

# Human-AI Collaboration to Support Mental Health and Well-Being

Ashish Sharma

A dissertation  
submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy

University of Washington  
2024

*Reading Committee:*  
Tim Althoff, Chair  
Noah A. Smith  
Ranjay Krishna

Program Authorized to Offer Degree:  
Computer Science and Engineering

© Copyright 2024

Ashish Sharma

University of Washington

**Abstract**

Human-AI Collaboration to Support  
Mental Health and Well-Being

Ashish Sharma

Chair of the Supervisory Committee:

Tim Althoff

School of Computer Science and Engineering

As mental health conditions surge worldwide, healthcare systems are struggling to provide accessible and high-quality mental health care for all. Although therapy can support people struggling with mental health challenges, barriers like clinician shortages and mental health stigma commonly limit people’s access to therapy. In this thesis, I study how human-AI collaboration can improve access to and quality of mental health support.

First, I study how human-AI collaboration can empower people who provide support to conduct effective and high-quality conversations. Specifically, I focus on *peer supporters* on online peer support platforms like Reddit and TalkLife. While peer supporters are motivated and well-intentioned to help support seekers, they are typically untrained and unaware of key psychotherapy skills, such as empathy, that foster effective support. Using a reinforcement learning-based method, evaluated through a randomized trial with 300 peer supporters from the largest peer support platform, I demonstrate that AI-based feedback helps peer supporters express empathy more effectively in their conversations.

Second, I investigate how human-AI collaboration can empower people who seek support by making self-guided mental health interventions more accessible and easier to engage with. Self-guided interventions, such as “do-it-yourself” tools to learn and practice coping skills, are often cognitively demanding and emotionally triggering, creating accessibility barriers that limit their

wide-scale implementation and adoption. Using cognitive restructuring of negative thoughts as a case study, evaluated through a randomized trial on a large mental health website with 15,531 participants, I show that human-AI collaboration supports people in overcoming negative thoughts and informs psychology theory about processes that lead to positive outcomes.

Third, I systematically evaluate human-AI collaboration systems used for mental health support. While there is great interest in utilizing AI for mental health support, there is a significant lack of methods to evaluate their effectiveness, quality, equity, and safety. I study how clinical trials can be conducted to effectively evaluate short-term and long-term outcomes, equity, and safety of AI-based mental health interventions comparing them to traditional approaches. Moreover, I develop a computational framework to automatically assess the behavior of large language models (LLM) when employed as therapists. By analyzing 13 different psychotherapy techniques, I compare the behavior of LLM therapists against that of high- and low-quality human therapy. My analysis reveals that LLMs often resemble behaviors more commonly exhibited in low-quality therapy rather than high-quality therapy, such as offering a higher degree of problem-solving advice when clients share emotions, which is against typical recommendations.

My thesis develops two human-AI collaboration systems to support mental health and well-being, along with an evaluation framework for such systems. My work opens opportunities to improve the learning and practice of mental health strategies and coping skills for both support seekers and support providers through human-AI collaboration interventions.

# Acknowledgements

First and foremost, I would like to thank my advisor, Tim Althoff, for taking a chance on me as one of his first PhD students. I still remember my first research project with him and the paper we wrote. Though I had prior research experience, his mentorship on that single project fundamentally transformed my approach to research, something that has stayed with me ever since. I am deeply grateful to him for pushing me towards pursuing high-impact, high-quality work. Learning how to achieve this has been one of the biggest lessons of my PhD journey. I also want to thank him for enabling so many partnerships and collaborations with organizations like TalkLife, Mental Health America, Talkspace, Headspace, and many others that have been foundational to my research and personal growth.

I want to thank my committee members, Noah Smith, Ranjay Krishna, Alexis Hiniker, and Adam Miner, for their unique perspectives that have significantly improved my thesis. Throughout my PhD, Noah's guidance on conducting NLP research has been instrumental, and I'm grateful for his mentorship. Adam's expertise in mental health became a key component of my PhD research, and I'm immensely appreciative of his support. My thesis has drawn a lot of inspiration from the human-AI collaboration research that Ranjay and Alexis do. I am thankful for their time and efforts in supporting my PhD thesis.

I'm deeply grateful to have Inna Lin as my close collaborator throughout my PhD. Being a PhD student can be isolating, but having Inna as my go-to person for both personal and research challenges was invaluable. I also want to thank Ethan Hsu and Kelly Chiu. Ethan and Kelly re-taught me what it means to be a young student with so much energy and passion. I am grateful for their significant contributions to my work.

I'm immensely grateful for the opportunity to collaborate with Mental Health America during

my PhD. Working alongside Kevin Rushton, Theresa Nguyen, and others at MHA was a true privilege. A significant highlight of my PhD was deploying a tool I developed on the Mental Health America platform, impacting hundreds of thousands of users. This would not have been possible without the trust and invaluable contributions of Kevin and Theresa.

Collaborating with psychotherapy and mental health experts was a key part of my research. I want to thank Adam Miner, Dave Atkins, Zac Imel, and Torrey Creed for their guidance in grounding our research in psychotherapy principles. Their expertise was critical in helping us develop AI systems with the potential for real-world impact.

I also want to thank my collaborators from TalkLife, particularly Jamie Druitt. The licensed dataset that Jamie provided became one of the foundations of my research and helped me learn about the key challenges and problems in this space.

A career in research wouldn't have been possible without the support of my mentors. I am grateful to Niloy Ganguly and Koustav Rudra for introducing me to research and hand-holding me through my undergraduate research endeavors. I also want to thank Hari Sundaram and Adit Krishnan for their mentorship. I was initially very hesitant about pursuing a PhD (as everyone is). It was only after they took me as an intern in the final year of my undergrad that I became excited to pursue this goal. I especially thank Adit for being a mentor throughout my career, guiding me through personal to professional struggles.

I'm grateful to Monojit Choudhury and Kalki Bali for facilitating my smooth transition into the PhD program through the excellent Research Fellow program at Microsoft. Pursuing a PhD and doing research can be stressful, but the year I spent with them fostered a sense of calm. Working with Monojit, in particular, was instrumental in helping me discover the problems I'm passionate about.

I also want to thank Sudha Rao, Chris Brockett, Nebojsa Jojic, and Bill Dolan who mentored me during my internship at Microsoft which was such a refreshing experience and helped me in exploring the other perspectives of human-AI collaboration research.

I also thank my collaborators, Dave Wadden, Isabell Griffith Fillipo, Jina Suh, Chris Rytting, Xinyi Zhou, Mike Pullmann, Meghan Romanelli, Brittany Mosser, Khendra Lucas, John Marion, Rou Rouvere, and Akanksha Malhotra.

Thanks to the bdata team – Galen Weld, Mike Merrill, Margaret Li, Ken Gu, Vinayak Gupta,

Bret Nestor – for helping pilot my data collection endeavors and for giving invaluable feedback on my research and paper writing. Thanks to Galen and Margaret for all the fun social events as well.

Thanks to my amazing friends – Prabhat Agarwal, Jatin Arora, Jeenu Grover, Sumeet Kumar, Chinmaya Pancholi, Sebastin Santy, Raghav Agarwal, Tapan Chugh, Jaehun Jung, Aditya Kusupati, Artidoro Pagnoni, Pratyush Patel, Raghav Somani, Sahil Verma – for always keeping the fun alive in my life. Thanks to Sebastin for also helping me with the design of various human-facing systems throughout my PhD.

Thanks to my grandparents – Gugan Ram Sharma and Vimla Sharma, my parents – Bijesh Kumar Sharma and Bharti Sharma, my sister and brother-in-law – Nidhi Sharma and Chanchal Sharma, my uncle and aunt – Pinkesh Sharma and Sunita Sharma, and my cousins – Aditya Sharma and Ridhi Sharma. During the first year of my PhD, COVID hit, and I decided to move back to India to be with my family. Having them by my side during those difficult times was a blessing, providing me with the support I needed to stay positive and relaxed.

This research would not have been possible without funding and support from NSF grant IIS-1901386, NSF CAREER IIS-2142794, NSF grant CNS-2025022, NIH grant R01MH125179, Bill & Melinda Gates Foundation (INV-004841), the Office of Naval Research (#N00014-21-1-2154), a Microsoft AI for Accessibility grant, a Garvey Institute Innovation grant, and UW Azure Cloud Computing Credits.

Last, but not least, I am deeply grateful to UW CSE for facilitating an exceptional research environment.



# Contents

<b>1</b>	<b>Introduction</b>	<b>51</b>
1.1	Thesis Statement	54
1.2	Contributions and Thesis Overview	54
1.2.1	Human-AI Collaboration to Empower People Providing Support	56
1.2.2	Human-AI Collaboration to Empower People Seeking Support	57
1.2.3	Evaluating Human-AI Collaboration Mental Health Interventions	58
<b>2</b>	<b>Human-AI Collaboration to Empower People Providing Support</b>	<b>61</b>
2.1	Understanding Empathy Expressed in Peer-to-Peer Mental Health Support	63
2.1.1	How to Measure Empathy?	64
2.1.2	Framework of Expressed Empathy	65
2.1.3	Data Collection	67
2.1.4	Model	69
2.1.5	Results	73
2.1.6	Model-based Insights into Mental Health Platforms	76
2.1.7	Implications for Empathy-Based Feedback	78
2.2	Empathic Rewriting: Computationally Transforming Low-Empathy Conversations into Higher Empathy	78
2.2.1	Problem Definition and Goals	80
2.2.2	Data Description	83
2.2.3	PARTNER: Empathic Rewriting Using Reinforcement Learning	85
2.2.4	PARTNER : Experiments	91

2.2.5	Discussion . . . . .	99
2.3	Helping Untrained Peer Supporters Express Empathy More Effectively . . . . .	99
2.3.1	HAILEY: Human-AI Collaboration Approach for Empathy . . . . .	99
2.3.2	Study Design . . . . .	104
2.3.3	Results . . . . .	107
2.3.4	Discussion . . . . .	110
2.4	Related Work . . . . .	118
2.5	Summary of Contribution to Thesis . . . . .	120
<b>3</b>	<b>Human-AI Collaboration to Empower People Seeking Support</b>	<b>121</b>
3.1	Cognitive Restructuring of Negative Thoughts Through Human-AI Collaboration . . . . .	123
3.1.1	Problem Definition and Goals . . . . .	126
3.1.2	Framework of Linguistic Attributes of Reframed Thoughts . . . . .	126
3.1.3	Data Collection . . . . .	128
3.1.4	Method . . . . .	129
3.1.5	Experiments and Results . . . . .	133
3.2	Randomized Field Study on a Large Mental Health Platform . . . . .	135
3.2.1	Study Overview . . . . .	137
3.2.2	RQ1: How can we design a self-guided cognitive restructuring intervention that is supported through human-language model interaction? . . . . .	138
3.2.3	RQ2a – To what extent does our intervention help individuals in alleviating negative emotions and overcoming negative thoughts? . . . . .	144
3.2.4	RQ2b – What is the impact of individual design hypotheses on the effec- tiveness of the intervention? . . . . .	148
3.2.5	RQ3a - What types of reframed thoughts do people prefer? . . . . .	153
3.2.6	RQ3b - How do the linguistic attributes of reframed thoughts relate to the desired outcomes of cognitive reframing? . . . . .	155
3.2.7	RQ4 – How equitable is the intervention and what strategies may improve its equity? . . . . .	156
3.2.8	Discussion . . . . .	162

3.3	Related Work	167
3.4	Summary of Contribution to Thesis	169
<b>4</b>	<b>Evaluating AI-based Mental Health Interventions</b>	<b>171</b>
4.1	A Clinical Trial of AI-Based Mental Health Intervention	173
4.1.1	Intervention Groups for Randomized Controlled Trials	176
4.1.2	Study Measures to Assess Short-term and Long-term Effects	176
4.1.3	Ethical and Safety Considerations	179
4.1.4	Study Procedure	180
4.1.5	Results	181
4.2	A Computational Framework for Behavioral Assessment of LLM Therapists	185
4.2.1	BOLT: Framework for Assessing Conversational Behavior of LLM Therapists	189
4.2.2	Results	197
4.2.3	Discussion	208
4.3	Related Work	211
4.4	Summary of Contribution to Thesis	212
<b>5</b>	<b>Conclusion</b>	<b>213</b>
5.1	Summary of Contributions	213
5.2	Future Work	215
5.2.1	Long-term Assistance	215
5.2.2	Robust Safety Evaluation	216
5.2.3	Adapting to Other Applications and Contexts	216
<b>A</b>	<b>Supplementary Materials for Human-AI Collaboration to Empower People Providing Support</b>	<b>251</b>
A.1	Data Collection Details	251
A.2	Reproducibility	253
A.3	Potential Confounding Factors in Analysis of Variation of Empathy over Time	254
A.4	Randomized Controlled Trial Details	255

<b>B</b>	<b>Supplementary Materials for Human-AI Collaboration to Empower People Seeking Support</b>	<b>295</b>
B.1	Method . . . . .	295
B.2	Reproducibility . . . . .	296
B.3	List of Thinking Traps . . . . .	297
B.4	Example Illustrating Our Rationality Measurement . . . . .	298
B.5	Randomized Field-Study Interface . . . . .	299
B.6	Data Collection Instructions . . . . .	300
B.7	Consent Form Used in the Randomized Field Study . . . . .	304
<b>C</b>	<b>Supplementary Material for Evaluating AI-based Mental Health Interventions</b>	<b>313</b>

# List of Figures

1.1	Overview of my dissertation research and contributions. I build and evaluate human-AI collaboration systems to support people who seek mental health support and people who provide mental health support by developing human-centered NLP methods and designing and deploying human-AI interaction. My dissertation research bridges the critical gap between need and access to mental health support. . . . .	55
2.1	We use two independently pre-trained RoBERTa-based encoders for encoding seeker post and response post respectively. We leverage attention between them for generating seeker-context aware representation of the response post, used to perform the two tasks of empathy identification and rationale extraction. . . . .	71
2.2	(a) Peer-supporters do not self-learn empathy over time. Only users who joined in 2015 were included but similar trends hold for other user groups; (b) Stronger communications of emotional reactions and interpretations are received positively by seekers. Stronger explorations get 47% more replies; (c) A lot more seekers follow peers after empathic interactions; (d) Females are more empathic towards females. . . . .	77

2.3	An overview of the empathic rewriting task. Given a post from support seeker and a low-empathy response, the task is to rewrite the response for making it more empathic, through text <b>insertions</b> and <b>deletions</b> . This task requires inferring <i>specific</i> feelings and experiences from seeker’s post and using them for making appropriate changes to the response through empathic mechanisms like emotional reactions, interpretations, and explorations. <i>Examples have been paraphrased for anonymization (Matthews et al., 2017).</i> . . . . .	79
2.4	PARTNER uses a deep reinforcement learning approach for Empathic Rewriting. It leverages a transformer language model for performing the two actions of (1) selecting positions for insertion or replacement and (2) generating candidate empathic sentences. It uses four reward functions that promote increase in empathy, text fluency, sentence coherence, context specificity, and diversity. . . . .	81
2.5	Expression of high levels of empathy is very low in online support platforms, especially for Interpretations (IP) and Explorations (EX). Emotional reactions (ER) are slightly more common. . . . .	84
2.6	Analysis of empathic rewritings. Error bars are 95% confidence intervals. . . . .	94
2.7	Human evaluation of empathy, fluency, and specificity in rewritings from PARTNER vs. (a) rewritings from baseline models, and (b) rewritings from ablations. PARTNER is preferred over baselines and ablations in empathy and specificity and is competitive in fluency. . . . .	96

2.8 We performed a randomized controlled trial with 300 TalkLife peer supporters as participants. We randomly divided participants into Human Only (control) and Human + AI (treatment) groups and asked them to write supportive, empathic responses to seeker posts without feedback and with feedback, respectively. To identify whether just-in-time Human-AI collaboration helped increase expressed empathy beyond potential (but rare) traditional training methods, participants in both groups received initial empathy training before starting the study (Methods; Supplementary Figure A.2). **(a)** Without AI, human peer supporters are presented with an empty chatbox to author their response (the current status quo). As peer supporters are typically untrained on best-practices in therapy – such as empathy – they rarely conduct highly empathic conversations. **(b)** Our feedback agent (HAILEY) prompts peer supporters for providing just-in-time AI feedback as they write their responses. **(c)** HAILEY then suggests changes that can be made to the response to make it more empathic. These suggestions include new sentences that can be *inserted* and options for *replacing* current sentences with their more empathic counterparts. Participants can accept these suggestions by clicking on the *Insert* and *Replace* buttons and continue editing the response or get more feedback, if needed. . . . . 100

2.9 **(a)** Human evaluation from an independent set of TalkLife users showed that the Human + AI responses (N=139) were strictly preferred 46.9% of the time relative to a 37.4% strict preference for the Human Only responses (N=161). **(b)** Through automatic evaluation using an AI-based expressed empathy score (Chapter 2.1), we found that the Human + AI responses (N=139) had 19.6% higher empathy than the Human Only responses (N=161; 1.77 vs. 1.48; Cohen’s  $d=0.24$ ;  $p=5.1 * e^{-8}$ ;  $t = 5.46$ ;  $D_f = 2998$ ; Two-sided Student’s t-test). **(c)** For the participants who reported challenges in writing responses after the study, we found a stronger preference for the Human + AI responses vs. Human Only responses (49.1% vs. 34.0%), compared to participants who did not report challenges (44.6% vs. 41.5%). **(d)** For participants who reported challenges in writing responses after the study, we found a higher improvement in expressed empathy scores of the Human + AI responses vs. Human Only responses (38.9%; 1.74 vs. 1.25; Cohen’s  $d=0.43$ ), compared to participants who did not report challenges (11.9%; 1.79 vs. 1.60; Cohen’s  $d=0.15$ ). In **c** and **d**, the sample size varied to ensure comparable conditions (Methods). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 108

2.10 We derived a hierarchical taxonomy of Human-AI collaboration categories. **(a)** We clustered the interaction patterns of Human + AI (treatment) participants based on how often the AI was consulted during the study and how the AI suggestions were used (N=116/139). We excluded participants who belonged to multiple clusters (N=23/139). Very few participants always consulted and used AI (2.6%), indicating that participants did not rely excessively on AI feedback. Participants could use AI feedback directly through suggested actions (64.6%) or indirectly by drawing ideas from the suggested feedback and rewriting it in their own words in the final response (18.5%). **(b)** Empathy increased when participants consulted and used AI more frequently, with those who did not consult AI (22.4%) or did not use AI (9.5%) having significantly lower preference over Human Only responses (N=37;  $p = 6.4 * e^{-6}$ ; Two-sided Student's t-test). **(c)** Participants who did not consult AI had the lowest empathy levels based on our automatic evaluation (1.13 on average out of 6). The area of the points is proportional to the number of participants in the respective human-AI collaboration categories. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . 111

3.1 **(a)** We consider the task of reframing negative thoughts with different, more hopeful thoughts using LMs; **(b)** Different perspectives on a situation may result in different reframes. Here, we propose a framework of seven reframing attributes (see gray box). Given a reframed thought, we control each attribute (e.g., *actionability*) to generate reframes that score higher or lower on that attribute (e.g., *more or less actionable*); **(c)** We deploy this model on Mental Health America, a large U.S. national mental health website ([bit.ly/changing-thoughts](https://bit.ly/changing-thoughts)) and conduct a randomized field study with 2,067 participants. We suggest LM-generated reframes to MHA visitors and assess which reframing attributes are desirable and what constitutes a relatable, helpful and memorable reframe. . . . . 125

3.2	We design a human-language model interaction based system for self-guided cognitive restructuring of negative thoughts. The system involves (a-b) describing the context by participants, (c-d) LM-assisted identification of thinking traps, and (e-g) LM-assisted writing of reframed thoughts. . . . .	139
3.3	(a) Reduction in emotion intensity of participants before and after using the system (emotion scale: 1 to 10). We found that 67% of the participants reported having a positive reduction in (negative) emotions (N=1,922). (b) Participants with higher emotion intensity before using the system reported a higher reduction in emotion intensity post the system usage (N=1,922). Error bars represent 95% bootstrapped confidence intervals. . . . .	145
3.4	Randomized controlled trial to estimate the effects of contextualizing thoughts through situation (N=1,636). (a) Contextualizing participant thoughts through their situations led to 2.80% more helpful reframes (3.31 vs. 3.22; $p = 0.0104$ ) but did not lead to more relatable reframes. (b) Asking for additional context did not lead to a lower completion rate. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at $\alpha = 0.05$ . . . .	149
3.5	Randomized controlled trial to estimate the effects of enabling participants to iteratively edit reframes through increased interaction with the language model (N=2,165). 38% of participants chose to use this intervention. (a) Having the option of interactive reframe edits available to participants led to a 23.73% greater reduction in emotion intensity (2.19 vs. 1.77). (b-e) However, it did not lead to significant differences in other outcomes (at $\alpha = 0.05$ ). Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at $\alpha = 0.05$ . . . . .	151

- 3.6 Participants were provided with an option to seek one or more of the following types of more specific suggestions from the language model – actionable, empathic, or personalized. Among those who used any of the options (N=992), (a) those who chose to make their reframes actionable experienced superior effectiveness across all five outcomes; (b) those who chose to make their reframes empathic reported 21.86% higher reduction in emotion intensity (2.23 vs. 1.83), 5.52% higher reframe helpfulness (3.44 vs. 3.26), and 5.14% higher skill learnability (3.48 vs. 3.31) and no significant differences based on reframe relatability and reframe memorability (at  $\alpha = 0.05$ ); (c) those who chose to make their reframes personalized reported no significant differences in outcomes (at  $\alpha = 0.05$ ). Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . . . . . 152
- 3.7 **Which linguistic attributes of reframed thoughts do people prefer?** For a given situation and thought from MHA visitors, we show them multiple LM-generated reframes with variance across a randomly selected attribute (e.g., low, medium, and high actionability). We find that highly empathic and highly specific reframings are more preferred. On the other hand, reframes with high positivity are less preferred. N: No; Y: Yes; L: Low; M: Medium; H: High. Error bars represent 95% bootstrapped confidence intervals. . . . . 153
- 3.8 **Which linguistic attributes are associated with desired cognitive reframing outcomes?** For a given situation and thought, we show one LM-generated reframe to MHA participants and ask them to rate it on relatability, helpfulness and memorability on a 1 to 5 scale. For each linguistic attribute, we compare the first (Q1) and the fourth quartile (Q4). We find that (a) reframes that have higher rationality are more relatable; (b) reframes that address thinking traps, have higher actionability or higher specificity are more helpful; (c) reframes that have higher actionability or higher specificity are more memorable. Error bars represent 95% bootstrapped confidence intervals. \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ . . . . . 154

3.9 Randomized controlled trial to estimate the effects of making reframes easier to understand and more casual on adolescents in age group 13 to 14 (N=148). Adolescents reported 8.60% higher reframe relatability (4.04 vs. 3.72) and 14.44% higher reframe helpfulness (3.17 vs. 2.77) if they were suggested easier to understand and more casual reframes compared to instances where such reframes were not suggested. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . . . . . 161

4.1 A conceptual model outlining our study. We conducted a between-subjects study where participants were divided into DIY w/ AI, DIY w/o AI, or Psychoeducation Only groups. All three groups had education components that included the education of restructuring skills. DIY w/o AI and DIY w/ AI groups have DIY components that include the practice of restructuring skills including identifying distortions and reframing thoughts. Finally, DIY w/ AI group has AI components that include just-in-time assistance on restructuring skills and interactive refinement. We study how these different interventions may affect outcomes both short-term and long-term in a study spanning multiple weeks. We also study how user attributes like demographics and depression/anxiety levels have an influence on outcomes. . . . . 175

4.2 We found that DIY w/ AI led to significantly higher reframing effectiveness compared to DIY w/o AI during the first usage (Cohen's  $d = 0.18$ ). We also found that DIY w/ AI led to similar emotional regulation as DIY w/o AI (Cohen's  $d=0.07$ ). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. \*:  $p < 0.05$  . . . . . 181

4.3 We assess the outcomes of DIY w/ AI, DIY w/o AI, and Psychoeducation Only groups over a period of 8 weeks. We found that participants in the DIY w/ AI and DIY w/o AI conditions demonstrated similar levels of Emotional Mechanisms, Skill Acquisition, Reduction in PHQ-9, and Reduction in GAD-7. However, participants in the Psychoeducation Only condition demonstrated significantly lower emotional mechanisms and skill acquisition levels than participants in DIY w/ AI and DIY w/o AI conditions. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 182

4.4 Overview of BOLT, a computational framework that enables systematic assessment of the behavior of LLM therapists and compares them to high- and low-quality human therapy. . . . . 188

4.5 Difference in the frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. **A**: average-quality, **Low**: low-quality, and **High**: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in **blue** indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in **orange** indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in **gray** are not statistically significantly different from average-quality at  $p = 0.05$  at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test, following Bonferroni correction (m: number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. A key insight we find is that LLMs respond with significantly higher **PROBLEM-SOLVING** (subfigure (a)), similar to low-quality human therapy. On the other hand, LLMs respond with significantly higher **REFLECTIONS ON STRENGTHS** (subfigure (l)), similar to high-quality therapy, but with a frequency that significantly exceeds high-quality therapy. . . . . 200

4.6 Difference in the *temporal order* of conversational behaviors, operationalized as the turn numbers in which behaviors are first exhibited in a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. **A**: average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the order is increasing (we flip the axis if low-quality is exhibited later than high-quality, such that low-quality is visualized below the average quality marker). Values colored in **blue** indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in **orange** indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in **gray** are not statistically significantly different from average-quality at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test, following Bonferroni correction (m: number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. Most LLM therapists start providing PLANNING (subfigure (b)) earlier in the conversations but provide NORMALIZING (subfigure (j)) later against common recommendations (Cochran and Cochran, 2015). . . . . 202

4.7 Difference between the frequency of conversational behaviors observed in LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) or low-quality human therapy *in response to specific client behaviors (Adaptability)*, relative to average-, low-, and high-quality human therapy. **A**: average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in **blue** indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in **orange** indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in **gray** are not statistically significantly different from average-quality at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test, following Bonferroni correction (m: number of (client, therapist) intents tested =  $13 \times 6 = 78$ ). Error bars indicate 95% bootstrapped confidence intervals. Here, a key finding is that LLMs respond with significantly lower QUESTIONS ON EMOTIONS when clients express SUSTAINING UNHEALTHY BEHAVIOR (subfigure (b)), similar to low-quality human therapy. . . . . 205

4.8 We incorporate simple prompts that aim to calibrate LLM therapists, specifically (a) increase QUESTIONS ON EXPERIENCES, (b) decrease PROBLEM-SOLVING, and (c) decrease NORMALIZING. Subfigures show changes in the frequency of conversational behaviors based on changes in prompts to different LLM therapists (GPT-4, GPT-3.5 turbo, Llama2-70b, Llama2-13b). Changes to individual LLMs are shown in pairs (left – Original prompt; right – modulated prompt). The corresponding high-quality human therapy behavior frequency is shown as green dashed lines. For instance, GPT-4 increases the frequency of QUESTIONS ON EXPERIENCES from 29.9% to 57.0% with the modulated prompt. In general, we find that only GPT-4 is able to modulate behavior frequency to a statistically and practically significant amount into the desired direction always, whereas the modulation is inconsistent for other models. Error bars indicate 95% bootstrapped confidence intervals. . . . 207

A.1 Empathy over time analysis of various user groups. We find similar trends across multiple groups. . . . . 255

A.2 Empathy training used in our study. Participants in both the Human + AI (treatment) and Human Only (control) groups received the same training. The training included the empathy definition, a framework of common ways of expressing empathy in responses, and examples of empathic responses. This ensures that participants were working under similar understandings of empathy. In practice, such training is very rare and the effect of training typically diminishes over time. The identified difference in empathy between treatment and control groups in our study therefore conservatively estimates the impact of our AI-in-the-loop feedback system, and not baseline differences in empathy definitions. The effect in practice may be larger than the intentionally conservative estimates produced here, as such training is uncommon on current mental health platforms. . . . . 257

A.3 Comparison of Human Only (control) and Human + AI (treatment) responses with AI Only responses (generated directly from PARTNER(Sharma et al., 2021), the deep reinforcement learning model for empathic rewriting, used as a foundation for HAILEY (Methods)). **(a)** Through human evaluation from an independent set of TalkLife users, we found that AI Only responses have a similar preference as the Human + AI responses (48.2% vs. 46.9%; N=161;  $p=0.23$ ; Two-sided Student's t-test) but a higher preference than the Human Only responses (48.2% vs. 37.4%; N=139;  $p=3.3 * 10^{-5}$ ; Two-sided Student's t-test). **(b)** Automatic estimation of empathy, on the contrary, suggested that AI Only responses have a higher expressed empathy score compared to Human + AI responses (2.10 vs. 1.77; N=139; Cohen's  $d=0.28$ ;  $p=5.3 * 10^{-13}$ ; Two-sided Student's t-test). Importantly however, note that the AI Only responses were optimized on the same scoring function that we use to automatically estimate empathy, which likely explains the high scores of the AI Only approach. **(c)** However, while the authenticity of Human Only and Human + AI responses was comparable (69.6% vs. 65.4%; N=139;  $p=0.01$ ; Two-sided Student's t-test), the authenticity of AI Only responses was significantly lower (36.5% vs. 65.4%; N=161;  $p=3.7 * 10^{-8}$ ; Two-sided Student's t-test). This highlights the key issue of authenticity with using AI Only, alongside safety, privacy, bias and other unintended consequences in the high-risk setting of mental health. To summarize, we find that Human + AI is the only approach that leads to both high empathy and high authenticity. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 258

A.4 Perceptions of Human + AI (treatment) group participants as reported in phase IV (post-intervention survey). We observed that more than 63.3% of participants found the current feedback helpful, 60.4% found it actionable and 69.8% of participants self-reported feeling more confident at providing support after our study. Also, 77.7% of participants wanted this type of feedback system to be deployed on TalkLife or other similar peer-to-peer support platforms, indicating potential opportunities for deployment in real-world. . . . . 259

A.5 Comparison of our rewriting-based AI treatment with a secondary classification-based AI treatment. A classification-based AI treatment provided participants with an option to request empathy classification scores for their responses, as opposed to the more granular feedback consisting of concrete suggestions to edit responses in our primary rewriting-based approach (Supplementary Figure A.6). Our hypothesis was that such a treatment should be less actionable and is likely to lead to less empathic responses than the rewriting-based treatment. In our study, we assigned a secondary classification-based treatment to 10% of the incoming participants at random (N=30). **(a)** Through human evaluation from an independent set of TalkLife users, we found that the Human + Classification responses have a significantly lower preference than the Human + Rewriting responses (37.9% vs. 47.8%; N=30;  $p=0.002$ ; Two-sided Student's t-test). **(b)** Automatic estimation of empathy, on the contrary, suggested that the Human + Classification responses have a higher expressed empathy score compared to Human + Rewriting responses (2.24 vs. 1.77; N=30; Cohen's  $d=0.37$ ;  $p=4.7 * 10^{-6}$ ; Two-sided Student's t-test). As the same score is also exposed to participants just-in-time in the classification-based treatment, it may have led participants to be put particular emphasis on a high expressed empathy score, which participants in the rewriting-based treatment feedback didn't have direct access to. **(c)** We found that less participants in the classification-based treatment group agree on deploying the system on TalkLife than the rewriting-based treatment (63.3% vs. 77.7%; N=30;  $p=0.0998$ ; Two-sided Student's t-test; Supplementary Figure A.4). Also, we observed that more participants in the classification-based treatment disagree on its actionability than participants in the rewriting-based treatment, but the difference may not be statistically significant due to the limited power (23.3% vs. 12.2%; N=30;  $p=0.1154$ ; Two-sided Student's t-test). The area of the points in the plots is proportional to the number of participants in the respective control/treatment conditions. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 260

A.6 Interface of our classification-based AI treatment (Supplementary Figure A.5). **(a)** Participant is asked to write a supportive, empathic response and given an option to receive feedback. **(b)** Participant starts writing the response. **(c)** Participant clicks on the “Get Feedback” button to request classification-based feedback. The feedback consists of classification scores on three empathy communication mechanisms – Emotional Reactions, Interpretations, and Explorations. **(d)** Participant edits the response based on the classification scores, often improving on the communication mechanisms with low scores and requests “More Feedback” if needed. 261

A.7 Both Human Only (control) and Human + AI (treatment) group participants showed a significant drop in empathy levels in the last 5 responses of our study. With Human + AI, however, we observed a significantly lower drop in empathy (5.3% vs. 26.0%; N=139; p=0.0062; Two-sided Student’s t-test). This indicates the effectiveness of just-in-time AI feedback in alleviating challenges like empathy fatigue, associated with providing mental health support. The empathy differences between Human Only (N=161) and Human + AI (N=139) responses are statistically significant for both first 5 and last 5 responses ( $p=1.1 \times 10^{-8}$ ; Two-sided Student’s t-test). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 262

A.8 Comparison of existing Human Only responses on TalkLife with Human Only and Human + AI responses in our study. Human Only responses on TalkLife had significantly lower preference for empathy (18.4% vs. 37.4% vs. 46.9%; N=139;  $p=1.4 * 10^{-6}$ ; Two-sided Student’s t-test) and significantly lower expressed empathy score (1.11 vs. 1.48 vs. 1.77;  $p=2.8 * 10^{-46}$ ; Two-sided Student’s t-test). This difference might be attributed to the additional initial empathy training provided to participants, as well as a potential selection effect in our study that may have attracted Talklife users who particularly care about expressing empathy in supporting others. As our study shows that Human-AI collaboration improves empathy expression even for those participants who already express empathy more often, practical gains for the average user of the Talklife platform could be even higher. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 263

A.9 Qualitative examples of just-in-time AI feedback provided to participants by HAI-LEY. In (b) and (c), the original peer supporter response was empty. Seeker posts in these examples have been paraphrased for anonymization. . . . . 264

A.10 The distribution of post and response lengths. The seeker posts in our dataset had a mean length of 25.9 words, a standard deviation of 25.3 words and a median of 18.0 words. The response posts collected in our study had a mean length of 25.9 words, a standard deviation of 34.6 words and a median of 19.0 words. Also, Human + AI responses (mean = 22.4 words; std = 34.6; median = 19.0) were 28.9% shorter in length compared to Human Only responses on average (mean = 32.1 words; std = 44.4; median = 21.0;  $p < 0.001$ ; Two-sided Student’s t-test). In addition, we found that Human + AI responses had 5.2% higher diversity than the Human Only responses based on the Distinct-1 metric(Li et al., 2016a), which computes the number of distinct unigrams divided by the total number of tokens (0.146 vs. 0.139;  $p = 0.019$ ; Two-sided Wilcoxon signed-rank test). . . . . 265

A.11 Background and demographics of participants in Human Only (control) and Human + AI (treatment) groups, as reported in phase I (pre-intervention survey). . . . 266

A.12 Differences between expressed empathy scores of participants in Human Only (control) and Human + AI (treatment) groups, stratified by demographics of participants and their prior experience with online peer support. The area of the points is proportional to the number of participants in the respective categories. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. . . . . 267

A.13 Perceptions of participants in Human Only (control) and Human + AI (treatment) groups, as reported in phase I (pre-intervention survey). . . . . 268

A.14 Distribution of participants in Human Only (control) and Human + AI (treatment) groups who report writing responses as challenging or stressful, as reported in phase IV (post-intervention survey). . . . . 269

A.15 Distribution of participants in the Human Only (control) group who indicate that feedback could have improved responses, as reported in phase IV (post-intervention survey). . . . . 270

A.16 Perceptions of participants in the Human + AI (treatment) group, as reported in phase IV (post-intervention survey). . . . . 271

A.17 Expressed empathy levels of responses with perceptions of Human + AI (treatment) group participants, as reported in phase IV (post-intervention survey). The area of the points is proportional to the number of participants with respective perceptions. Error bars indicate bootstrapped 95% confidence intervals. . . . . 272

A.18 Participant perceptions, as reported in phase IV (post-intervention survey), with different human-AI collaboration categories. . . . . 273

A.19 Consent form used in our study. . . . . 274

A.20 Form used for collecting demographics and background of participants [phase I: pre-intervention survey]. . . . . 275

A.21 Onboarding survey used for collecting perceptions of participants [phase I: pre-intervention survey]. . . . . 276

A.22 Instructions shown to the control group participants [phase II: empathy training and instructions]. Continued on the next page (1/2). . . . . 277

A.23 Instructions shown to the control group participants [phase II: empathy training and instructions] (2/2).	278
A.24 Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (1/6).	279
A.25 Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (2/6).	280
A.26 Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (3/6).	281
A.27 Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (4/6).	282
A.28 Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (5/6).	283
A.29 Instructions shown to the treatment group participants [phase II: empathy training and instructions] (6/6).	284
A.30 An example workflow for Human Only (control) participants [phase III: write supportive, empathic responses]. <b>(a)</b> Participant is asked to write a supportive, empathic response. <b>(b)</b> Participant starts writing the response.	285
A.31 An example workflow for Human + AI (treatment) participants [phase III: write supportive, empathic responses]. <b>(a)</b> Participant is asked to write a supportive, empathic response and given an option to receive feedback. <b>(b)</b> Participant starts writing the response. <b>(c)</b> Participant clicks on the prompt to request feedback from HAILEY. <b>(d)</b> Participant accepts the suggested changes and gets an option to request more feedback. <b>(e)</b> Participant continues editing the response and requests more feedback as needed. <b>(f)</b> When the response is already highly empathic, the participant simply receives a positive feedback.	286
A.32 Interface for flagging feedback [phase III: write supportive, empathic responses].	287
A.33 Exit survey used for collecting perceptions of control group participants [phase IV: post-intervention survey].	288
A.34 Exit survey used for collecting perceptions of treatment group participants [phase IV: post-intervention survey]. Continued on the next page (1/2).	289

A.35 Exit survey used for collecting perceptions of treatment group participants [phase IV: post-intervention survey] (2/2). . . . .	290
A.36 Consent form used for human evaluation of responses. . . . .	291
A.37 Instructions for human evaluation of responses. . . . .	292
A.38 Interface for human evaluation of responses. . . . .	293
B.1 To measure reasoning strength, we generate two explanations for each reframe – one for why it might be sound; another for why it may be flawed. To check if the explanations themselves are well-reasoned, we recursively generate explanations for the explanations. Here, we choose a recursive tree depth of 3. Also, at every step, we generate three explanations in favour of a reframe and three explanations against it. . . . .	298
B.2 Illustration of the interface used for our randomized field-study: <b>(a)</b> Participant starts by writing the negative thought they are struggling with in the moment; <b>(b)</b> We ask the participant to describe a recent situation that may have led to their thought; <b>(c)</b> An AI model identifies possible cognitive distortion(s) in the thought. Participant selects the cognitive distortions that they most relate to; <b>(d)</b> An AI model generates and suggests three different reframed thoughts that may help overcome negative thinking and the associated cognitive distortion. Participant selects the reframe they find the most relatable, helpful and memorable. Some of the instructions provided to the participants, including informed consent and evaluation, have been omitted from this illustration for brevity. . . . .	299
B.3 Instructions shown during data collection with mental health experts. Continued on the next page (1/3). . . . .	301
B.4 Instructions shown during data collection with mental health experts. Continued on the next page (2/3). . . . .	302
B.5 Instructions shown during data collection with mental health experts (3/3). . . . .	303
B.6 Consent form shown to the MHA visitors. Continued on the next page (1/2). . . . .	305
B.7 Consent form shown to the MHA visitors (2/2). . . . .	306

B.8 Randomized controlled trial to estimate the effects of contextualizing thoughts through emotions (N=4,016). (a) Contextualizing participant thoughts through their emotions led to 3.86% lower relatable reframes (3.87 vs. 3.72). Note that our language model does not necessarily incorporate emotions unless they are expressed in the thought or situation as well. (b) Asking for additional context did not lead to lower completion rate. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . . . . 307

B.9 Randomized controlled trial to estimate the effects of integrating psychoeducation (N=1,850). We did not find significant quantitative improvement in outcomes on integrating psychoeducation. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . . . . . 307

B.10 Participants who further interacted with the language model to seek additional reframing suggestions of specific types (actionable, empathic, or personalized) reported 5.57% higher reframe helpfulness (3.41 vs. 3.23) and 4.86% higher skill learnability (3.45 vs. 3.29) and no significant differences in reduction in emotion intensity, reframe relatability, and reframe memorability (at  $\alpha = 0.05$ ; N=992). Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . For a randomized trial assessing the effects of this intervention, see Figure 3.5. . . . . 307

B.11 Reading Complexity (Coleman–Liau Index) of the thoughts written by participants based on their age. We find that adolescents (those below the age of 18) write thoughts with the least reading complexity. Error bars represent 95% bootstrapped confidence intervals. . . . . 308

B.12 Randomized controlled trial to estimate the effects of making reframes easier to understand and more casual on age groups 15 to 17 and  $\geq 18$ . Participants in the age group 15 to 17 reported 15.58% higher reframe helpfulness (3.19 vs. 2.76) when they were suggested such reframes. We did not find significant differences for participants in age group  $\geq 18$  due to this intervention. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . . . . . 308

B.13 Detailed interface and process for iterative edits of reframes through further interaction with the language model. . . . .	311
C.1 Different prompts variations for classifying conversational behavior embedded in both therapist and client utterance in the counseling setting. We explored two variations of prompting – 1) Prompting with definitions of different conversational behaviors, with the sources of two studies (Lee et al., 2019; Cao et al., 2019). 2) Prompting with the definitions and also examples from annotated dataset. For 2), we explored two ways of asking questions, which are the multi-label setting and binary-label setting. We experimented with different variations of these prompts (e.g., using different language for the instructions) but observed similar results. See the detailed methodology description in Methods. . . . .	337
C.2 Instruction prompts given to LLMs for simulating counseling conversation using two types of methods (single response and full conversation): 1) using LLMs to write a single response within an ongoing conversation between human therapist and client. 2) creating a full and multi-turn conversation between LLMs and clients. See the detailed description in Methods. . . . .	338



# List of Tables

2.1	EPITOME incorporates both emotional and cognitive aspects of empathy that were previously only studied in face-to-face therapy and never computationally in text-based, asynchronous conversations. *Rashkin et al. (2019) implicitly enable empathic conversations through grounding in emotions instead of communication. . . .	66
2.2	Statistics of the collected empathy dataset. The crowdworkers were trained on EPITOME through a series of phone calls and manual/automated feedback on sample posts to ensure high quality annotations. . . . .	69
2.3	Empathy identification task results. We observe substantial gains over baselines with our seeker-context aware, multi-tasking approach. . . . .	72
2.4	Rationale extraction task results. We evaluate both at the level of tokens (T-f1) and spans (IOU-f1). . . . .	74
2.5	Ablation results. Most of our gains are due to context provided through attention and seeker post; higher gains for the rationale extraction task. *Note that rationales cannot be predicted after removing them from training. . . . .	74
2.6	Statistics of the TalkLife dataset. . . . .	84
2.7	Performance of PARTNER and comparisons with dialogue generation and other sequence-to-sequence generation baselines on the set of automatic metrics. PARTNER outperforms all baselines in empathy improvement and generates fluent, specific, and diverse outputs with lower edits. (↑) indicates higher is better, (↓) indicates lower is better. . . . .	91

2.8	Ablation results. Warm-start improves over DialoGPT but is still much worse than PARTNER in empathy improvement, highlighting the effectiveness of our RL-based training. . . . .	94
2.9	BLEU scores of all models against empathic rewritings from experts. The rewritings from PARTNER are closest to expert rewritings. . . . .	96
2.10	Qualitative examples of empathic rewriting using PARTNER and baseline methods. PARTNER generates rewritings that leverage both seeker post and original response post for empathic responses. It infers perspective from seeker posts (“ <i>I feel like nobody cares about my existence</i> ” → “ <i>It’s hard to find others who can relate</i> ”) and generates empathic sentences which can be coherently inserted to response posts (“ <i>What happened between you two?</i> ” → “ <i>What happened between you two? What caused you to break?</i> ”) . . . . .	97
3.1	Correlation of our proposed attribute measures by with human judgments from mental health experts. *: $p < 0.05$ ; **: $p < 0.001$ ; ***: $p < 10^{-5}$ . . . . .	132
3.2	Automatic and Human Evaluation Results. R-1: ROUGE-1; R-L: ROUGE-L; BScore: BertScore; Rel.: Relatability; Help.: Helpfulness. . . . .	133
3.3	Mean and standard deviation of the five quantitative measures as reported by participants. . . . .	146
3.4	Participants with higher emotion intensity before using the system reported a higher reduction in emotion intensity post the system usage. Participants with higher initial emotion intensities reported worse reframing outcomes, suggesting that writing effective reframes and learning the cognitive restructuring skill was harder when individuals were emotionally agitated (N=1,922). . . . .	146

3.5 Effectiveness of our system across different issues expressed by participants. Numbers highlighted in **green** indicate outcomes that are significantly better than the population mean ( $p < 0.05$ ). Numbers highlighted in **red** indicate outcomes that are significantly worse than the population mean ( $p < 0.05$ ). We found that participants who expressed *Parenting* and *Work* related issues reported better outcomes than the population means. Moreover, participants who expressed *Hopelessness*, *Loneliness*, and *Tasks & Achievement* related issues reported worse outcomes. . . . . 156

3.6 Effectiveness of our system across different demographic population. Numbers highlighted in **green** indicate outcomes that are significantly better than the population mean ( $p < 0.05$ ). Numbers highlighted in **red** indicate outcomes that are significantly worse than the population mean ( $p < 0.05$ ). AIAN: American Indian or Alaska Native; MENA: Middle Eastern or North African; NHPI: Native Hawaiian and Pacific Islander. We found that adolescents, males, and those with middle school education reported worse outcomes. Moreover, adults (age  $\geq 25$ ) and those with graduate and doctorate education reported better outcomes. . . . . 159

4.1 Effect sizes (Cohen’s  $d$ ) comparing outcomes between DIY w/ AI participants and DIY w/o AI participants, across various subpopulations. Numbers highlighted in **green** indicate outcomes where DIY w/ AI is significantly better than DIY w/o AI ( $p < 0.05$ ). Numbers highlighted in **red** indicate outcomes where DIY w/ AI is significantly worse than DIY w/o AI ( $p < 0.05$ ). We found that young adults (age 18 to 24) assigned to DIY w/ AI reported significantly higher levels of reframing effectiveness and emotional regulation compared to young adults assigned to the DIY w/o AI. . . . . 184

4.2 Percentage of analysis in which the models were found to be desirable (✓), undesirable (✗), or uncertain (?). On one hand, in some cases, LLMs demonstrate desirable behaviors resembling high-quality therapy, which could be promising. However, the prevalence of undesirable behaviors resembling low-quality therapy is concerning. These behaviors deviate from established therapeutic guidelines and could affect the overall quality of care. . . . . 209

A.1	Train/Dev/Test Splits. . . . .	253
A.2	Description of our randomized controlled trial (RCT) study population, setting and model, following reporting standards for artificial intelligence in health care from Hernandez-Boussard et al. (2020) . . . . .	256
B.1	Definitions and examples of the list of issues identified in our open coding process.	309
B.2	Examples illustrating the rewriting of reframes to make them simpler and casual for adolescents. Simpler and casual reframes use more straightforward words and phrases (e.g., “valuable” → “awesome”, “things I have accomplished” → “stuff I’ve already done”, “potential I have” → “how much more I can do”, “improve my physical health” → “make my body healthier”) and use informal language (e.g., “But you know what?”, “Plus...”). This could potentially be more appealing especially to adolescents.	310
C.1	Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at $p = \frac{0.05}{m}$ using Two-sided Student’s t-test after Bonferroni correction, where $m$ is the number of t-test per one model ( $m = 13$ ). $d$ is the cohen’s $d$ when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS by 40.0% while human therapists responds by 23.4% on average ( $P = 8.0 \times 10^{-73}$ , Cohen’s $d = 0.4$ , two-sided Student’s $t$ -test). . . . .	314

C.2 Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS by 31.1% and GPT-4 with full conversation simulation responds by 42.5% when compared with the human therapists responds by 23.4% on average from Table C.1. (Single:  $P = 2.3 \times 10^{-6}$ , Cohen’s  $d = 0.2$ ; two-sided student t-test) (Full:  $P = 6.4 \times 10^{-84}$ , Cohen’s  $d = 0.4$ ; two-sided student t-test) . . . . . 314

C.3 Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS at 3.1 turns earlier than average human therapists responds. (GPT-4: 3.1 turn, Average human therapists: 6.2 turn;  $P = 4.3 \times 10^{-15}$ , Cohen’s  $d = 0.7$ , two-sided Student’s  $t$ -test). . . . . 315

C.4 Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS at 4.3 turn which is 1.9 turns earlier than average human therapists. Similarly, GPT-4 with full conversation simulation responds at 2.0 turn which is 4.2 turns earlier than average human therapists. (Single: 4.3 turn, Average human therapists: 6.2 turn from Table C.3;  $P = 1.1 \times 10^{-3}$ , Cohen’s  $d = 0.4$ , two-sided Student’s  $t$ -test) (Full: 2.0 turn, Average human therapists: 6.2 turn from Table C.3;  $P = 9.2 \times 10^{-16}$ , Cohen’s  $d = 0.7$ ). . . . . 315

C.5 Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON CONSEQUENCES by 19.9% and human therapists responds by 27.3% on average when client expresses CHANGING UNHEALTHY BEHAVIOR. ( $P = 4.1 \times 10^{-6}$ , Cohen’s  $d = -0.2$ , two-sided Student’s  $t$ -test). . . . . 316

C.6 (Continue Table C.5) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 38.9% and human therapists responds by 24.5% on average when client expresses SHARING NEGATIVE EMOTIONS. ( $P = 4.0 \times 10^{-9}$ , Cohen’s  $d = 0.3$ , two-sided Student’s  $t$ -test). . . . . 317

C.7 Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON CONSEQUENCES by 38.9% which is 15.8% less than average human therapists. (Single: 11.5%, Average human therapists: 27.3% from Table C.5;  $P = 5.3 \times 10^{-17}$ , Cohen’s  $d = -0.4$ , two-sided Student’s  $t$ -test) . . . . . 318

C.8 (Continue Table C.7) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON NEEDS by 38.9% which is 23.6% more than average human therapists. (Single: 48.1%, Average human therapists: 24.5% from Table C.6;  $P = 2.3 \times 10^{-10}$ , Cohen’s  $d = 0.5$ , two-sided Student’s  $t$ -test) . . . . . 319

C.9 Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 14$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits POSITIVE EMOTIONS by 1.5% while human therapists responds by 1.0% on average ( $P = 8.4 \times 10^{-13}$ , Cohen’s  $d = 0.1$ , two-sided Student’s  $t$ -test). . . . . 320

C.10 Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen's d when compared with human therapist average. For instance, GPT-4 with single response simulation responds with POSITIVE EMOTIONS by 2.0% and GPT-4 with full conversation simulation responds by 1.3% when compared with the human therapists responds by 1.0% on average from Table C.9. (Single:  $P = 2.0 \times 10^{-14}$ , Cohen's  $d = 0.3$ ; two-sided student t-test) (Full:  $P = 3.7 \times 10^{-5}$ , Cohen's  $d = 0.1$ ; two-sided student t-test). . . . . 320

C.11 Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen's d when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS by 40.2% while human therapists responds by 23.4% on average ( $P = 6.9 \times 10^{-90}$ , Cohen's  $d = 0.4$ , two-sided Student's t-test). . . . . 321

C.12 Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS by 33.5% and GPT-4 with full conversation simulation responds by 41.9% when compared with the human therapists responds by 23.4% on average from Table C.11. (Single:  $P = 2.6 \times 10^{-14}$ , Cohen’s  $d = 0.2$ ; two-sided student t-test) (Full:  $P = 8.1 \times 10^{-99}$ , Cohen’s  $d = 0.4$ ; two-sided student t-test) . . . . . 321

C.13 Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS at 3.3 turns earlier than average human therapists responds. (GPT-4: 2.9 turn, Average human therapists: 6.2 turn;  $P = 1.0 \times 10^{-19}$ , Cohen’s  $d = 0.8$ , two-sided Student’s  $t$ -test). . . . . 322

C.14 Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS at 4.0 turn which is 2.2 turns earlier than average human therapists. Similarly, GPT-4 with full conversation simulation responds at 1.8 turn which is 4.4 turns earlier than average human therapists. (Single: 4.0 turn, Average human therapists: 6.2 turn from Table C.13;  $P = 6.4 \times 10^{-6}$ , Cohen’s  $d = 0.5$ , two-sided Student’s  $t$ -test) (Full: 1.8 turn, Average human therapists: 6.2 turn from Table C.13;  $P = 4.7 \times 10^{-21}$ , Cohen’s  $d = 1.0$ ). . . . . 322

C.15 Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 10.2% and human therapists responds by 27.7% on average when client expresses CHANGING UNHEALTHY BEHAVIOR. ( $P = 1.9 \times 10^{-46}$ , Cohen’s  $d = -0.5$ , two-sided Student’s  $t$ -test). . . . . 323

C.16 (Continue Table C.15) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 14.0% and human therapists responds by 24.5% on average when client expresses SHARING NEGATIVE EMOTIONS. ( $P = 3.2 \times 10^{-7}$ , Cohen’s  $d = -0.3$ , two-sided Student’s  $t$ -test). . . . . 324

C.17 Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON CONSEQUENCES by 11.6% which is 15.7% less than average human therapists. (Single: 11.6%, Average human therapists: 27.3% from Table C.15;  $P = 3.2 \times 10^{-21}$ , Cohen’s  $d = -0.4$ , two-sided Student’s  $t$ -test) . . . . . 325

C.18 (Continue Table C.17) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen's d when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON NEEDS by 46.5% which is 22.0% more than average human therapists. (Single: 46.5%, Average human therapists: 24.5% from Table C.16;  $P = 1.1 \times 10^{-11}$ , Cohen's  $d = 0.5$ , two-sided Student's  $t$ -test) . . . . . 326

C.19 Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 14$ ).  $d$  is the cohen's d when compared with human therapist average. For instance, GPT-4 exhibits POSITIVE EMOTIONS by 1.5% while human therapists responds by 1.0% on average ( $P = 2.8 \times 10^{-19}$ , Cohen's  $d = 0.2$ , two-sided Student's  $t$ -test). . . . . 327

C.20 Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with POSITIVE EMOTIONS by 2.1% and GPT-4 with full conversation simulation responds by 1.3% when compared with the human therapists responds by 1.0% on average from Table C.19. (Single:  $P = 2.8 \times 10^{-24}$ , Cohen’s  $d = 0.3$ ; two-sided student t-test) (Full:  $P = 4.7 \times 10^{-8}$ , Cohen’s  $d = 0.1$ ; two-sided student t-test). . . . . 327

C.21 Dataset statistics. . . . . 327

C.22 Performance of classification models on the tasks of (1) identifying therapist behavior across 13 different therapists codes and (2) identifying client behavior across 6 different client codes. We create five random train-test splits of our annotated dataset and report the average macro-precision (P), macro-recall (R), and macro-F1 (F1) scores with standard deviations as subscripts. def.: Definitions; ex.: Examples. Best-performing models are **bolded**. We find that GPT-4-based prompting methods that make use of psychotherapy-based definitions and examples achieve the highest macro-F1 performance. We adopt the prompting (multi-label) method for classifying therapist behavior while the prompting (binary-label) method for classifying client behavior (both highlighted in lightblue ). . . . . 328

C.23 Therapist Conversational Behavior Definition and Corresponding Examples based on techniques identified by researchers (Lee et al., 2019; Cao et al., 2019). We used these behavior categories in our annotation dataset. See more details in Methods. . 329

C.24 Client Conversational Behavior Definition and Corresponding Examples based on expressions from clients identified by researchers (Lee et al., 2019). We used these behavior categories in our annotation dataset. See more details in Methods. . . . . 330

C.25 Therapist Conversational Behavior Classification Model Performance Details. We examine the performance of our model across different classes. See the result description in Methods. . . . .	331
C.26 Client Conversational Behavior Classification Model Performance Details. We examine the performance of our model across different classes. See the result description in Methods. . . . .	331
C.27 Explanations generated from our behavior prediction model. We ask the model to explain its predictions and examine the differences between true and predicted behavior. . . . .	332
C.28 First Part of the Illustrative Example of Single-Response Simulation Method by LLMs. This method involves the actual responses of both therapist and client from a reference dataset, which are written in <i>italic</i> , and simulated responses of LLM Therapists. This full example is the simulated conversation from GPT-4 on one of the conversations using High-Low quality dataset as reference. . . . .	333
C.29 (Continue Table C.28) Second Part of the Illustrative Example of Single-Response Simulation Method by LLMs. This method involves the actual responses of both therapist and client from a reference dataset, which are written in <i>italic</i> , and simulated responses of LLM Therapists. This full example is the simulated conversation from GPT-4 on one of the conversations using High-Low quality dataset as reference.	334
C.30 Illustrative Example of Full Conversation Simulation Method by LLMs. The reference conversations have been fed into LLM to simulate both therapist and client. This full example is the simulated conversation from Llama-70b on one of the conversations using High-low quality dataset as reference. . . . .	335
C.31 Illustrative Example of Full Conversation Simulation Method by LLMs. The reference conversations have been fed into LLM to simulate both therapist and client. This full example is the simulated conversation from GPT-3.5 on one of the conversations that used HOPE dataset as a reference. . . . .	336



# Chapter 1

## Introduction

Mental health is a rapidly escalating global crisis, with approximately 20% of people worldwide experiencing mental health problems (WHO, 2022). Established therapeutic treatments, such as cognitive behavioral therapy (CBT), have proven efficacy in treating mental illness. Unfortunately, access to these treatments is limited. Many people do not or cannot receive care for various reasons, including a shortage of professionals, lack of insurance coverage, high costs, and stigma surrounding therapy (Olfson, 2016).

The **critical gap between the overwhelming need for and limited access to mental health care** has prompted clinicians, stakeholders, and advocates to develop innovative technological interventions aimed at complementing professional therapy and ensuring accessible care for all (Miner et al., 2019; Imel et al., 2015; Kazdin and Rabbitt, 2013; Schleider et al., 2020). However, developing effective mental health interventions that do not rely on professionals to deliver care is challenging.

Some interventions connect people seeking support (*support seekers*) with a large number of online *peers* who can provide support on text-based peer-to-peer support platforms at scale. Examples of such platforms include TalkLife ([talklife.com](http://talklife.com)), YourDost ([yourdost.com](http://yourdost.com)), and Mental Health Subreddits ([reddit.com](http://reddit.com)). However, *peer supporters* providing support on such platforms are typically untrained and unaware of effective therapeutic practices like expressing empathy to support seekers. Therefore, high-quality supportive conversations such as those that are highly empathic are rare on these platforms (Sharma et al., 2020b).

Other interventions facilitate self-guided practicing and learning of mental health skills and

coping strategies. These “do-it-yourself” tools, like journaling, mirror the proven effectiveness of therapeutic homeworks. Also, they can provide on-demand support for those unable to access a therapist or waiting for treatment. They can also support therapists who are seeking to augment ongoing care with engaging homework. However, despite their inherent promise, the wide-scale implementation of these interventions remains a challenge owing to the cognitive and emotional barriers that they pose (Garrido et al., 2019; Baumel et al., 2019; Fleming et al., 2018; Torous et al., 2020). Most interventions that digitally facilitate self-guided interventions simply transform traditional manual therapeutic worksheets into digital online formats (Shkel et al., 2023). Support seekers, already facing distress, often find themselves cognitively challenged and emotionally triggered when using these self-guided interventions without the assistance of a professional therapist.

**Augmenting mental health interventions with AI and NLP-based methods** has the potential to provide scaffolding that helps support providers and support seekers engage with them more effectively. Empowering peer supporters with automated, actionable, just-in-time feedback and training, such as through human-AI collaboration systems based on language models, can help them interact more effectively with support seekers (Imel et al., 2015; Miner et al., 2019). Traditionally, providing feedback and training processes can be slow, expensive, labor-intensive, and requires extensive expert instruction (Tanana et al., 2019; Moyers et al., 2005). AI systems can be trained to provide feedback on key therapy skills such as empathy and help assess the quality of care just-in-time and at scale. This ultimately could enhance the efficiency and effectiveness of the way in which mental health support is delivered. Simultaneously, AI-based techniques can empower support seekers in practicing and learning cognitively challenging and emotionally triggering mental health skills. AI-augmented “Do-It-Yourself” tools could provide assistance to improve mental health “in-the-moment” of crisis, thereby making self-guided interventions more engaging and accessible.

Despite these opportunities, the use of AI in a high-risk setting like mental health also presents potential risks. AI-based methods have been scrutinized for generating biased, harmful, or factually incorrect information (Blodgett et al., 2020; Lin et al., 2022; Ganguli et al., 2022), potentially undermining their effectiveness. As AI is integrated with new and existing mental health interventions, it is important to develop mechanisms for their continuous monitoring and

evaluation to ensure responsible usage and adherence to established interventions (Stade et al., 2024; De Choudhury et al., 2023; Li et al., 2020; Tate et al., 2023; Van Veen et al., 2024).

Previous research in AI and NLP for mental health has primarily focused on identifying effective psychotherapy techniques (Althoff et al., 2016; Pérez-Rosas et al., 2019; Zhang and Danescu-Niculescu-Mizil, 2020; Sharma and De Choudhury, 2018; Yang et al., 2019). However, there is a lack of emphasis on methods that can generate actionable, just-in-time feedback and training to those seeking support and those providing support. Psychotherapy techniques like empathy (Elliott et al., 2011) are conceptually complex, nuanced, and multi-dimensional, which makes generating such feedback and training challenging.

Also, prior efforts exploring AI-based mental health interventions have predominantly been limited to small-scale, Wizard-of-Oz-style research, that do not accurately measure the underlying machine's efficacy (Ly et al., 2017; Smith et al., 2021; Morris et al., 2015; Kornfield et al., 2023; Kumar et al., 2023). These studies were conducted in controlled lab settings and evaluated on online crowdworker platforms, such as MTurk, which may not accurately represent the people who actively seek mental health care or use such interventions (Mohr et al., 2017). Therefore, much less is known about intervention effectiveness in ecologically valid settings with individuals experiencing mental health challenges and seeking care. This limits our understanding of end-user preferences within these emerging forms of intervention (Mohr et al., 2017; Blandford et al., 2018; Borghouts et al., 2021; Poole, 2013).

Moreover, while experts have highlighted the critical need to evaluate AI-based mental health interventions, there is a significant lack of research and evidence demonstrating their feasibility, safety, and effectiveness comparing them against traditional treatment approaches (Stade et al., 2024; De Choudhury et al., 2023; Li et al., 2020; Tate et al., 2023; Van Veen et al., 2024). The rapid pace at which AI-based interventions are being developed also necessitates exploring *automated* evaluation methods to keep up with advancements. Moreover, the evaluation requires different types of open-ended texts that AI may generate, in response to a diverse range of clients, for different types of support, and at different stages of support, making the assessment complex and multi-faceted.

**In this thesis, I develop and evaluate human-AI collaboration interventions to bridge the critical gap between need and access to mental health support. Specifically, I:**

1. Develop **human-AI collaboration methods that can empower people providing mental health support** to effectively leverage effective therapeutic practices through automated, actionable, just-in-time feedback.
2. Develop **human-AI collaboration methods that can empower people seeking support** to learn and practice mental health skills through automated, actionable, just-in-time feedback.
3. Develop **systematic evaluation methods that can be used to assess human-AI collaboration methods for mental health** by comparing them against established treatment methods.

## 1.1 Thesis Statement

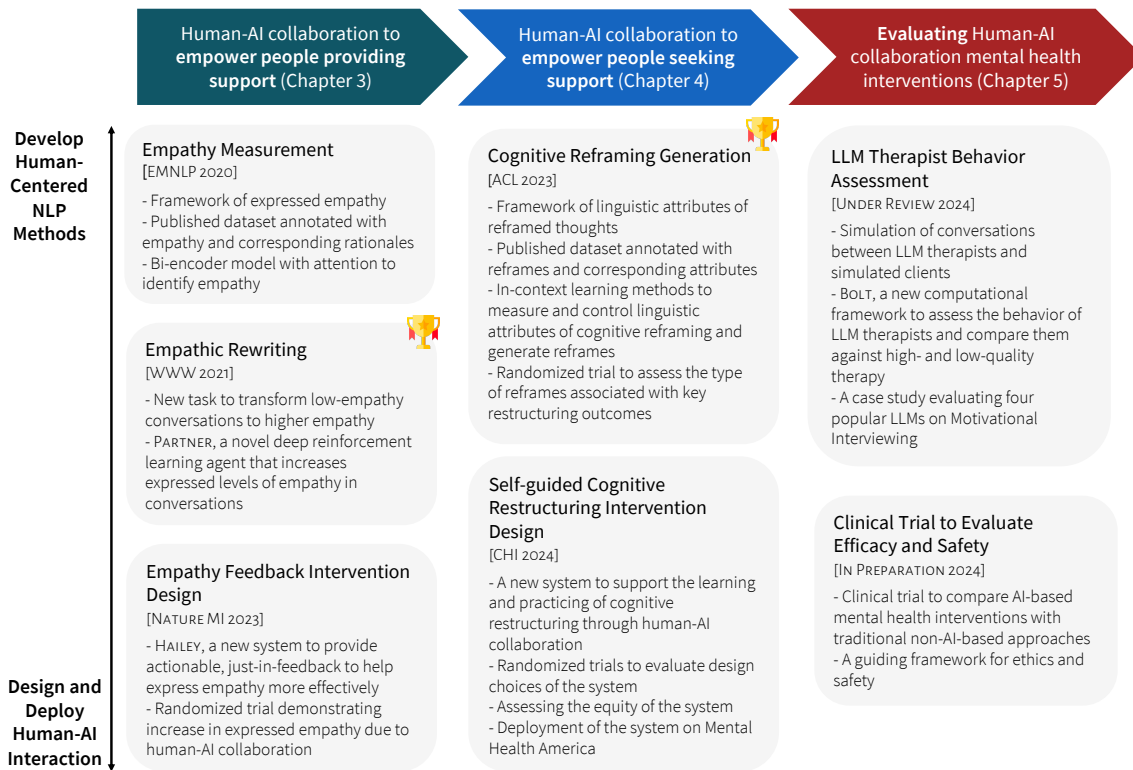
To summarize, my research demonstrates the following thesis statement:

*Human-AI collaboration methods can help develop more effective mental health interventions through automated, actionable, just-in-time feedback to individuals seeking and providing mental health support and can be evaluated by comparing their effectiveness against established treatment methods.*

## 1.2 Contributions and Thesis Overview

Figure 1.1 provides an overview of my thesis highlighting the contributions.

Making progress toward developing and evaluating human-AI collaboration interventions requires advancements in artificial intelligence, the integration of complex mental health constructs into current AI systems, and a deeper understanding of human behaviors and end-user needs and preferences. This calls for an interdisciplinary approach. In my thesis, I develop human-centered NLP methods and design and deploy human-AI interaction informed by well-established principles in psychology and mental health. I collaborate with experts in artificial intelligence, natural language processing, human-computer interaction, psychology, and mental health. My thesis makes significant contributions to the fields of AI, NLP, HCI, psychology, and mental health which I discuss next.



**Figure 1.1:** Overview of my dissertation research and contributions. I build and evaluate human-AI collaboration systems to support people who seek mental health support and people who provide mental health support by developing human-centered NLP methods and designing and deploying human-AI interaction. My dissertation research bridges the critical gap between need and access to mental health support.

### 1.2.1 Human-AI Collaboration to Empower People Providing Support

In Chapter 2, I develop a human-AI collaboration approach to empower people who provide mental health support. I focus on text-based, peer-to-peer mental health support and investigate how AI systems can collaborate with humans who provide support (*peer supporters*) to help facilitate the expression of empathy in textual supportive conversations.

First, I develop human-centered NLP methods to understand, measure, and improve how empathy is expressed on online peer-to-peer mental health support platforms. I create a novel unifying theoretically-grounded framework for characterizing the expression of empathy in text-based conversations. I collect and share a corpus of 10k (post, response) pairs annotated using this empathy framework with supporting evidence for annotations (rationales). I develop a multi-task RoBERTa-based bi-encoder model for identifying empathy in conversations and extracting rationales underlying its predictions. Experiments demonstrate that this approach can effectively identify empathic conversations. I further apply this model to analyze 235k mental health interactions and show that users do not self-learn empathy over time, revealing opportunities for empathy training and feedback. This work, described in Section 2.1, was originally published at the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020) (Sharma et al., 2020b). Moreover, I make our dataset and code publicly available at [github.com/behavioral-data/Empathy-Mental-Health](https://github.com/behavioral-data/Empathy-Mental-Health).

To improve the expressed levels of empathy in online mental health support conversations, I introduce a new task of *Empathic Rewriting* which aims to transform low-empathy conversational posts to higher empathy. Learning such transformations is challenging and requires a deep understanding of empathy while maintaining conversation quality through text fluency and specificity to the conversational context. I propose PARTNER, a deep reinforcement learning (RL) agent that learns to make sentence-level edits to posts in order to increase the expressed level of empathy while maintaining conversation quality. PARTNER leverages a policy network, based on a transformer language model adapted from GPT-2, which performs the dual task of generating candidate empathic sentences and adding those sentences at appropriate positions. During training, I reward transformations that increase empathy in posts while maintaining text fluency, context specificity, and diversity. Through a combination of automatic and human eval-

uation, I demonstrate that PARTNER successfully generates more empathic, specific, and diverse responses and outperforms NLP methods from related tasks such as style transfer and empathic dialogue generation. This work, described in Section 2.2, was originally *published at The Web Conference 2021 (WWW 2021)* and also won the **Best Paper Award** at the conference (Sharma et al., 2021). Moreover, I made our code publicly available at [github.com/behavioral-data/Partner](https://github.com/behavioral-data/Partner).

Finally, to translate these NLP models into real-world impact, I **design a human-AI interaction** system, HAILEY, an AI-in-the-loop agent that provides just-in-time feedback to help peer supporters respond more empathically to support seekers. I evaluate HAILEY in a non-clinical randomized controlled trial with real-world peer supporters on TalkLife (N=300), a large online peer-to-peer support platform. I show that our Human-AI collaboration approach leads to a 19.60% increase in conversational empathy between peers overall. Furthermore, I find a larger 38.88% increase in empathy within the subsample of peer supporters who self-identify as experiencing difficulty providing support. This work, described in Section 2.3, was originally *published at Nature Machine Intelligence 2023 (NATURE MI 2023)* (Sharma et al., 2023a).

Broadly, my findings demonstrate the potential of feedback-driven, AI-in-the-loop writing systems to empower humans in open-ended, social, creative tasks such as empathic conversations.

## 1.2.2 Human-AI Collaboration to Empower People Seeking Support

Next, in Chapter 3, I shift my focus towards individuals seeking mental health support and **develop a human-AI collaboration approach to empower people who seek mental health support**. I focus on self-guided mental health interventions and investigate how such interventions can be facilitated through human-AI collaboration. I take cognitive restructuring, an evidence-based therapeutic technique to overcome negative thinking (Beck, 1976), as a case study.

First, I **develop human-centered NLP methods** for cognitive restructuring. Based on psychology literature, I define a framework of seven linguistic attributes that can be used to reframe a thought. I develop automated metrics to measure these attributes and validate them with expert judgements from mental health practitioners. I collect a dataset of 600 situations, thoughts, and reframes from practitioners and use it to train a retrieval-enhanced in-context learning model

that effectively generates reframed thoughts and controls their linguistic attributes. This work, described in Section 3.1, was originally *published at the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)* and also won an **Outstanding Paper Award** at the conference (Sharma et al., 2023b). Moreover, I made our data and code publicly available at [github.com/behavioral-data/Cognitive-Reframing](https://github.com/behavioral-data/Cognitive-Reframing).

I use these NLP methods to **design and deploy a human-AI interaction** system to support people through various steps of cognitive restructuring. In an IRB-approved randomized field study on Mental Health America with 15,531 participants, I show that this system positively impacts emotional intensity for 67% of participants and helps 65% overcome negative thoughts. I also study what constitutes a “high-quality” reframe, and find that people prefer highly empathic or specific reframes, as opposed to reframes that are overly positive. Although adolescents report relatively worse outcomes, I find that tailored interventions that simplify language model generations improve overall effectiveness and equity. This work, described in Section 3.2, was originally *published at the ACM CHI conference on Human Factors in Computing Systems 2024 (CHI 2024)* (Sharma et al., 2024).

After appropriate ethical and safety considerations, I **deployed this system on Mental Health America**, a large mental health website that provides mental health resources and tools to millions of users website. The system is hosted at [bit.ly/changing-thoughts](https://bit.ly/changing-thoughts) and has been **used by over 157,000+ users**.

### 1.2.3 Evaluating Human-AI Collaboration Mental Health Interventions

While mental health interventions based on human-AI collaboration interventions bring a lot of promise, there is a lack of systematic approaches to evaluate them. In the last part of my thesis in Chapter 4, I establish how we can evaluate the efficacy and safety of such interventions on human participants through clinical trials. Moreover, I develop a framework to automatically evaluate conversations between people seeking mental health support and large language models.

First, I **study how clinical trials can be conducted to effectively evaluate short-term and long-term outcomes, equity, and safety** of AI-based mental health interventions. I use the AI-based cognitive restructuring intervention developed in Chapter 3 as a case study. I conduct

a clinical trial in an ecologically valid setting to compare our AI-based self-guided cognitive restructuring tool against traditional interventions including an equivalent non-AI tool and psychoeducation only. Also, I develop a guiding framework for ethical and safety considerations. Through this clinical trial, described in Section 4.1, I show that such AI-based self-guided cognitive restructuring intervention demonstrates superior short-term efficacy and long-term engagement compared to traditional approaches, while achieving similar levels of long-term efficacy.

Moreover, I study how large language models (LLMs) used in mental health support can be automatically evaluated. I propose BOLT, a **novel computational framework to study the conversational behavior of LLMs when employed as therapists**. I develop an in-context learning method to quantitatively measure the behavior of LLMs based on 13 different psychotherapy techniques including reflections, questions, solutions, normalizing, and psychoeducation. Subsequently, I compare the behavior of LLM therapists against that of high- and low-quality human therapy, and study how their behavior can be modulated to better reflect behaviors observed in high-quality therapy. The analysis of GPT and Llama-variants reveals that these LLMs often resemble behaviors exhibited in low-quality therapy rather than high-quality therapy, such as offering a higher degree of problem-solving advice when clients share emotions, which is against typical recommendations. At the same time, unlike low-quality therapy, LLMs reflect significantly more upon clients' needs and strengths. My analysis framework suggests that despite the ability of LLMs to generate anecdotal examples that appear similar to human therapists, LLM therapists are currently not fully consistent with high-quality care, and thus require additional research to ensure quality care. This work, described in Section 4.2, is *currently under review at Nature Medicine* (Chiu et al., 2024). Moreover, I make our dataset and code publicly available at [github.com/behavioral-data/BOLT](https://github.com/behavioral-data/BOLT).



## Chapter 2

# Human-AI Collaboration to Empower People Providing Support

This chapter focuses on how we can develop interventions to support people who provide mental health support. We focus on text-based, peer-to-peer mental health support and investigate how AI systems can collaborate with humans to help facilitate the expression of *empathy* in textual supportive conversations.

Research suggests that for people in distress, connecting and interacting with *peers* can help in coping with mental illness, enhancing mental well-being, and developing social integration (Davidson et al., 1999). This form of social support (Kaplan et al., 1977) through peers can be provided online which has led to the development of online mental health support platforms such as [TalkLife](#), [YourDost](#), or [Mental Health Subreddits](#) (Naslund et al., 2016). With an estimated 400 million people suffering from mental health disorders worldwide (WHO), combined with a pervasive lack of qualified mental health professionals (Kazdin and Blase, 2011; Olfson, 2016), these platforms have pioneered avenues for seeking social support and discussing mental health issues for millions of people who use them for expressing emotions, sharing stigmatized experiences, and providing and receiving peer support (Naslund et al., 2016).

A key component of providing successful support is *empathy*, the ability to understand and relate to the emotions and experiences of others and to effectively communicate that understanding (Robert et al., 2011). Empathic support is one of the critical factors that contributes to

successful conversations in mental health support, showing strong correlations with symptom improvement (Elliott et al., 2018) and the formation of alliance and rapport (Robert et al., 2011; Watson et al., 2002; Sharma et al., 2020b). While online peer-to-peer platforms enable supportive conversations between *support seekers* (people who seek support) and *peer supporters* (people who provide support), peer supporters are typically untrained in expressing complex and open-ended skills like empathy (Davis et al., 1980; Blease et al., 2020; Doraiswamy et al., 2020; Riess, 2017) and may lack the required expertise.

As shown in prior work (Kemp and Henderson, 2012; Mahlke et al., 2014), untrained peer supporters report difficulties in writing supportive, empathic responses to support seekers. Without deliberate training or specific feedback, the difficulty persists over time (Schwalbe et al., 2014; Goldberg et al., 2016; Sharma et al., 2020b) and may even lead to a gradual decrease in supporters’ effectiveness due to factors such as empathy fatigue (Nunes et al., 2011; Hojat et al., 2009; Stebnicki, 2007). Furthermore, current efforts to improve empathy (e.g., in-person empathy training) do not scale to the millions of peer supporters providing support online. Thus, empowering peer supporters with automated, actionable, just-in-time feedback and training, such as through human-AI collaboration systems, can help them express higher levels of empathy and, as a result, improve the overall effectiveness of these platforms (Imel et al., 2015; Miner et al., 2019; Sharma et al., 2020b, 2021).

Here, we investigate how AI systems can collaborate with humans to help facilitate the expression of *empathy* in textual supportive conversations. First, we design a novel framework and dataset of empathic conversations and develop a RoBERTa-based bi-encoder model for identifying empathy in conversations. Using this model, we demonstrate that highly empathic conversations are rare on online mental health platforms (Section 2.1). Therefore, to improve empathy in conversations, we introduce Empathic Rewriting, a new task that aims to transform low-empathy conversational posts to higher empathy. We propose PARTNER, a deep reinforcement learning agent that learns to make sentence-level edits to conversations in order to increase the expressed level of empathy while maintaining conversation quality through fluency, specificity, and diversity (Section 2.2). We use PARTNER to design and evaluate a new AI-in-the-loop agent that provides just-in-time feedback to help peer supporters respond more empathically to support seekers (Section 2.3).

## 2.1 Understanding Empathy Expressed in Peer-to-Peer Mental Health Support

AI systems that understand conversational empathy could empower peer supporters with feedback and training. However, the current understanding of empathy is limited to traditional face-to-face, speech-based therapy (Gibson et al., 2016; Pérez-Rosas et al., 2017) due to lack of resources and methods for new asynchronous, text-based interactions (Patel et al., 2019). Also, while previous NLP research has focused predominantly on empathy as reacting with emotions of warmth and compassion (Buechel et al., 2018), a separate but key aspect of empathy is to communicate a cognitive understanding of others (Selman, 1980).

Here, we present a novel computational approach to understanding how empathy is expressed in text-based, asynchronous mental health conversations. We introduce EPITOME,<sup>1</sup> a conceptual framework for characterizing communication of empathy in conversations that synthesizes and adapts the most prominent empathy scales from speech-based, face-to-face contexts to text-based, asynchronous contexts (Section 2.1.2). EPITOME consists of three communication mechanisms of empathy: *Emotional Reactions*, *Interpretations*, and *Explorations* (Figure 3.1).

To facilitate computational modeling of empathy in text, we create a new corpus based on EPITOME. We collect annotations on a dataset of 10k (post, response) pairs from extensively-trained crowdworkers with high inter-rater reliability (Section 2.1.3).<sup>2</sup> We develop a RoBERTa-based bi-encoder model for identifying empathy communication mechanisms in conversations (Section 2.1.4). Our multi-task model simultaneously extracts the underlying supportive evidences, *rationales* (DeYoung et al., 2020), for its predictions (spans of input post; e.g., highlighted portions in Figure 3.1) which serve the dual role of (1) explaining the model’s decisions, thus minimizing the risk of deploying harmful technologies in sensitive contexts, and (2) enabling rationale-augmented feedback for peer supporters.

We show that our computational approach can effectively identify empathic conversations with underlying rationales ( $\sim 80\%$  acc.,  $\sim 70\%$  macro-f1) and outperforms popular NLP baselines with a 4-point gain in macro-f1 (Section 2.1.5). We apply our model to a dataset of 235k

---

<sup>1</sup>EmPathy In Text-based, asynchrOnous MEntal health conversations

<sup>2</sup>Our dataset can be accessed from <https://bit.ly/2Rwy2gx>.

supportive conversations on TalkLife and demonstrate that empathy is associated with positive feedback from seekers and the forming of relationships. Importantly, our results suggest that most peer supporters do not self-learn empathy with time. This points to critical opportunities for training and feedback for peer supporters to increase the effectiveness of mental health support (Miner et al., 2019; Imel et al., 2015). Specifically, NLP-based tools could give actionable, real-time feedback to improve expressed empathy, and we demonstrate this idea in a small-scale proof-of-concept (2.1.6).

### 2.1.1 How to Measure Empathy?

Empathy is a complex multi-dimensional construct with two broad aspects related to emotion and cognition (Davis et al., 1980). The “emotion” aspect relates to the emotional stimulation in reaction to the experiences and feelings expressed by a user. The “cognition” aspect is a more deliberate process of understanding and interpreting the experiences and feelings of the user and communicating that understanding to them (Elliott et al., 2018).

Here, we study *expressed empathy* in text-based mental health support – empathy *expressed* or *communicated* by peer supporters in their textual interactions with seekers (cf. Barrett-Lennard (1981)).<sup>3</sup> Table 2.1 lists existing empathy scales in psychology and psychotherapy research. Truax and Carkhuff (1967) focus only on communicating cognitive understanding of others while Davis et al. (1980); Watson et al. (2002) also make use of expressing stimulated emotions.

These scales, however, have been designed for in-person interactions and face-to-face therapy, often leveraging audio-visual signals like expressive voice. In contrast, in text-based support, empathy must be expressed using textual response alone. Also, they are designed to operate on long, synchronous conversations and are unsuited for the shorter, asynchronous conversations of our context.

In this work, we adapt these scales to text-based, asynchronous support. We develop a new comprehensive framework for text-based, asynchronous conversations (Table 2.1; Section 2.1.2), use it to create a new dataset of empathic conversations (Section 2.1.3), a computational approach

---

<sup>3</sup>Note that *expressed* empathy may differ from the empathy *perceived* by seekers. However, obtaining perceived empathy ratings from seekers is challenging in sensitive contexts and involves ethical risks. Psychotherapy research indicates a strong correlation of expressed empathy with positive outcomes and frequently uses it as a credible alternative (Robert et al., 2011).

for identifying empathy (Section 2.1.4; Section 2.1.5), & gaining insights into mental health platforms (2.1.6).

## Computational Approaches for Empathy

Computational research on empathy is based on speech-based settings, exploiting audio signals like pitch which are unavailable in text-based platforms (Gibson et al., 2016; Pérez-Rosas et al., 2017). Moreover, previous NLP research has predominantly focused on empathy as reacting with emotions of warmth and compassion (Buechel et al., 2018). For mental health support, however, communicating cognitive understanding of feelings and experiences of others is more valued (Selman, 1980). Recent work also suggests that grounding conversations in emotions implicitly makes them empathic (Rashkin et al., 2019). Research in therapy, however, highlights the importance of expressing empathy in interactions (Truax and Carkhuff, 1967). In this work, we present a computational approach to (1) understanding empathy expressed in textual, asynchronous conversations; (2) address both emotional and cognitive aspects of empathy.

### 2.1.2 Framework of Expressed Empathy

To understand empathy in text-based, asynchronous, peer-to-peer support conversations, we develop EPITOME, a new conceptual framework of expressed empathy (Figure 3.1). In close collaboration with clinical psychologists, we adapt and synthesize existing empathy definitions and scales to text-based, asynchronous context. EPITOME consists of three communication mechanisms providing a comprehensive outlook of empathy – *Emotional Reactions*, *Interpretations*, and *Explorations*. For each of these mechanisms, we differentiate between – (0) peers not expressing them at all (*no communication*), (1) peers expressing them to some weak degree (*weak communication*), (2) peers expressing them strongly (*strong communication*).

Here, we describe our framework in detail using the following seeker post as context for all example responses: *I am about to have an anxiety attack.*

**Emotional Reactions.** Expressing emotions such as warmth, compassion, and concern, experienced by peer supporter after reading seeker’s post. Expressing these emotions plays an important role in establishing empathic rapport and support (Robert et al., 2011). A **weak commu-**

		Context	Applicable to text-based peer-support	Communication Mechanisms		
				Emotional Reactions	Interpretations ( <i>Cognitive</i> )	Explorations ( <i>Cognitive</i> )
Scales	Truax and Carkhuff (1967)	Face-to-face therapy	✗	✗	✓	✓
	Davis et al. (1980)	Daily human interactions	✗	✓	✓	✗
	Watson et al. (2002)	Face-to-face therapy	✗	✓	✓	✓
Methods	Buechel et al. (2018)	Reaction to news stories	✗	✓	✗	✗
	Rashkin et al. (2019)	Emotionally grounded convs.	✗	✗*	✗*	✗*
	Pérez-Rosas et al. (2017)	Motivational interviewing	✗	✗	✓	✓
EPITOME		Text-based, asynchronous support	✓	✓	✓	✓

**Table 2.1:** EPITOME incorporates both emotional and cognitive aspects of empathy that were previously only studied in face-to-face therapy and never computationally in text-based, asynchronous conversations. \*Rashkin et al. (2019) implicitly enable empathic conversations through grounding in emotions instead of communication.

**nication** of emotional reactions alludes to these emotions without the emotions being explicitly labeled (e.g., *Everything will be fine*). On the other hand, **strong communication** specifies the experienced emotions (e.g., *I feel really sad for you*).

**Interpretations.** Communicating an understanding of feelings and experiences inferred from the seeker's post. Such a cognitive understanding in responses is helpful in increasing awareness of hidden feelings and experiences, and essential for developing alliance between the seeker and peer supporter (Watson, 2007). A **weak communication** of interpretations contains a mention of the understanding (e.g., *I understand how you feel*) while a **strong communication** specifies the inferred feeling or experience (e.g., *This must be terrifying*) or communicates understanding through descriptions of similar experiences (e.g., *I also have anxiety attacks at times which makes me really terrified*).

**Explorations.** Improving understanding of the seeker by exploring the feelings and experiences not stated in the post. Showing an active interest in what the seeker is experiencing and feeling and probing gently is another important aspect of empathy (Miller et al., 2003; Robert et al., 2011). A **weak exploration** is generic (e.g., *What happened?*) while a **strong exploration** is specific and labels the seeker's experiences and feelings which the peer supporter wants to explore (e.g., *Are you feeling alone right now?*).

Consistent with existing scales, responses that only give advice (*Try talking to friends*), only provide factual information (*mindful meditation overcomes anxiety*), or are offensive or abusive (*shut the f\*\*k up*)<sup>4</sup> are not empathic and are characterized as no communication of empathy in our framework.

### 2.1.3 Data Collection

To facilitate computational methods for empathy, we collect data based on EPITOME.

#### Data Source

We use conversations on the following two online support platforms as our data source:

---

<sup>4</sup>Our approach is focused on supporting peers who are trying to help seekers. This is different from toxic language identification tasks. Such content can be independently flagged using existing techniques (e.g., [perspectiveapi.com](https://perspectiveapi.com))

**(1) TalkLife.** TalkLife ([talklife.co](http://talklife.co)) is the largest global peer-to-peer mental health support network ([talklife.co/about](http://talklife.co/about)). It enables seekers to have textual interactions with peer supporters through conversational threads. The dataset contains 6.4M threads and 18M interactions (seeker post, response post pairs) on TalkLife between May 2012 to Jan 2019.

**(2) Mental Health Subreddits.** Reddit ([reddit.com](http://reddit.com)) hosts a number of sub-communities aka *subreddits* (e.g., *r/depression*). We use threads posted on 55 mental health focused subreddits (Sharma and De Choudhury, 2018). This publicly accessible dataset contains 1.6M threads and 8M interactions on Reddit between Jan 2015 to Jan 2019.

We use the entire dataset for in-domain pre-training (Section 2.1.4) and annotate a subset of 10k interactions on empathy. We further analyze empathy on a carefully filtered dataset of 235k mental health interactions on TalkLife (2.1.6).

## Annotation Task and Process

Empathy is conceptually nuanced and linguistically diverse so annotating it accurately is difficult in short-term crowdwork approaches. This is also reflected in prior work that found it challenging to annotate therapeutic constructs (Lee et al., 2019). To ensure high inter-rater reliability, we designed a novel training-based annotation process.

**Crowdworkers Recruiting and Training.** We recruited and trained eight crowdworkers on identifying empathy mechanisms in EPITOME. We leveraged Upwork ([upwork.com](http://upwork.com)), a freelancing platform that allowed us to hire and work interactively with crowdworkers. Each crowdworker was trained through a series of phone calls (30 minutes to 1 hour in total) and manual/automated feedback on 50-100 posts. Refer Appendix A.1 for more details.

**Annotating Empathy.** In our annotation task, crowdworkers were shown a pair of (seeker post, response post) and were asked to identify the presence of the three communication mechanisms in EPITOME (Emotional Reactions, Interpretations, and Explorations), one at a time. For each mechanism, crowdworkers annotated whether the response post contained no communication, weak communication, or strong communication of empathy in the context of the seeker post.

**Highlighting Rationales.** Along with the categorical annotations, crowdworkers were also asked to highlight portions of the response post that formed the rationale behind their annotation. E.g,

	Data Source	No	Weak	Strong	Total
Emotional Reactions	TalkLife	3656	2945	461	7062
	Reddit	2034	899	148	3081
Interpretations	TalkLife	5533	178	1351	7062
	Reddit	1645	115	1321	3081
Explorations	TalkLife	5137	767	1158	7062
	Reddit	2600	104	377	3081

**Table 2.2:** Statistics of the collected empathy dataset. The crowdworkers were trained on EPITOME through a series of phone calls and manual/automated feedback on sample posts to ensure high quality annotations.

in the post *“That must be terrible! I’m here for you”*, the portion *“That must be terrible”* is the rationale for it being a strong communication of interpretations.

**Data Quality.** Overall, our corpus has an average inter-annotator agreement of 0.6865 (average over pairwise Cohen’s  $\kappa$  of all pairs of crowdworkers; each pair annotated  $>50$  posts in common) which is higher than previously reported values for the annotation of empathy in face-to-face therapy ( $\sim 0.60$  in Pérez-Rosas et al., 2017; Lord et al., 2015).<sup>5</sup> Our ground-truth corpus contains 10,143 (seeker post, response post) pairs with annotated empathy labels from trained crowdworkers (Table 2.2).

**Privacy and Ethics.** The TalkLife dataset was sourced with license and consent from the TalkLife platform. All personally identifiable information (user and platform identifiers) in both the datasets were removed. This study was approved by University of Washington’s Institutional Review Board. In addition, we tried to minimize the risks of annotating mental health related content by providing crisis management resources to our annotators, following Sap et al. (2020). This work does not make any treatment recommendations or diagnostic claims.

### 2.1.4 Model

With our collected dataset, we develop a computational approach for understanding empathy.

<sup>5</sup>We achieve an inter-annotator agreement of 0.69 for emotional reactions, 0.61 for interpretations, and 0.76 for explorations.

## Problem Definition

Let  $\mathbf{S}_i = s_{i1}, \dots, s_{im}$  be a seeker post and  $\mathbf{R}_i = r_{i1}, \dots, r_{in}$  be a corresponding response post. For the pair  $(\mathbf{S}_i, \mathbf{R}_i)$ , we want to perform two tasks:

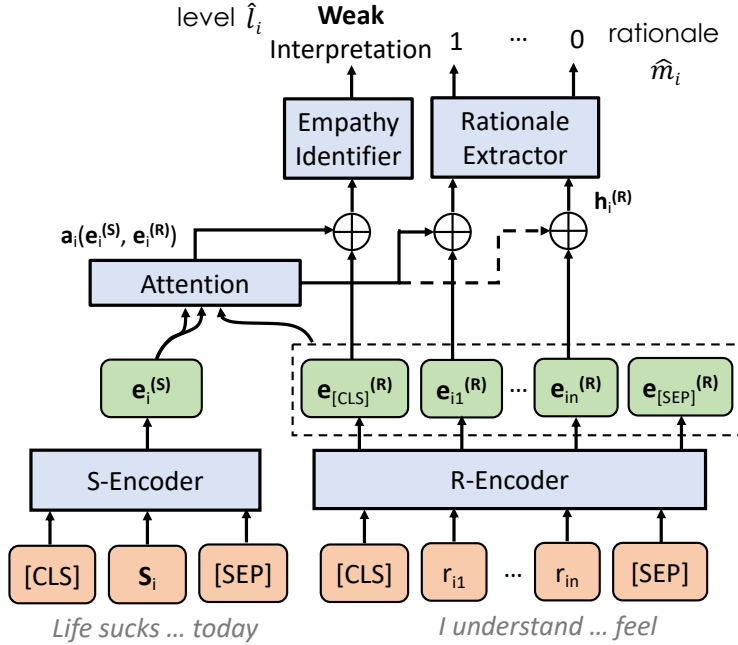
**Task 1: Empathy Identification.** Identify how empathic  $\mathbf{R}_i$  is in the context of  $\mathbf{S}_i$ . For each of the three communication mechanisms in EPITOME (Emotional Reactions, Interpretations, Explorations), we want to identify their level of communication ( $l_i$ ) in  $\mathbf{R}_i$  – no communication (0), weak communication (1), or strong communication (2).

**Task 2: Rationale Extraction.** Extract rationales underlying the identified level  $l_i \in \{\text{no, weak, strong}\}$  of each of the three communication mechanism in EPITOME. The extracted rationale is a subsequence of words  $x_i$  in  $\mathbf{R}_i$ . We represent this subsequence as a mask  $m_i = (m_{i1}, \dots, m_{in})$  over the words in  $\mathbf{R}_i$ , where  $m_{ij} \in \{0, 1\}$  is a boolean variable: 1 – rationale word; 0 – non-rationale word. Correspondingly,  $x_i = m_i \odot \mathbf{R}_i$ .

## Bi-Encoder Model with Attention

We propose a multi-task bi-encoder model based on RoBERTa (Liu et al., 2019) for identifying empathy and extracting rationales (Figure 2.1). We multi-task over the two tasks of empathy identification and rationale extraction and train three independent but identical architectures for the three empathy communication mechanisms in EPITOME (Section 2.1.2). The bi-encoder architecture (Humeau et al., 2019) facilitates a joint modeling of  $(\mathbf{S}_i, \mathbf{R}_i)$  pairs. Moreover, the use of attention helps in providing context from the seeker post,  $\mathbf{S}_i$ . We find that such an approach is more effective than methods that concatenate  $\mathbf{S}_i$  with  $\mathbf{R}_i$  with a [SEP] token to form a single input sequence (Section 2.1.5).

**Two Encoders.** Our model uses two independently pre-trained transformer encoders from RoBERTa<sub>BASE</sub> – S-Encoder & R-Encoder – for encoding seeker post and response post respectively. S-Encoder encodes context from the seeker post whereas R-Encoder is responsible for understanding empathy in the response post.



**Figure 2.1:** We use two independently pre-trained RoBERTa-based encoders for encoding seeker post and response post respectively. We leverage attention between them for generating seeker-context aware representation of the response post, used to perform the two tasks of empathy identification and rationale extraction.

$$\mathbf{e}_i^{(S)} = \text{S-Encoder}([\text{CLS}], \mathbf{S}_i, [\text{SEP}]) \quad (2.1)$$

$$\mathbf{e}_i^{(R)} = \text{R-Encoder}([\text{CLS}], \mathbf{R}_i, [\text{SEP}]) \quad (2.2)$$

where [CLS] and [SEP] are special start and end tokens adapted from BERT (Devlin et al., 2019).

**Domain-Adaptive Pre-training.** Both the S-Encoder and R-Encoder are initialized using the weights learned by RoBERTa<sub>BASE</sub>. We further perform a domain-adaptive pre-training (Gururangan et al., 2020) of the two encoders to adapt to conversational and mental health context. For this additional pre-training of the two encoders, we use the datasets of 6.4M seeker posts (182M tokens) and 18M response (279M tokens) posts respectively sourced from TalkLife (Section 2.1.3). We use the masked language modeling task for pre-training (3 epochs, batch size = 8).

**Attention Layer.** We use a single-head attention over the two encodings for generating seeker-context aware representation of the response post. Using the terminology of transformers (Vaswani et al., 2017), our query is the response post encoding  $\mathbf{e}_i^{(R)}$ , and the keys and the values are the

	Model	Emotional Reactions		Interpretations		Explorations	
		acc.	f1	acc.	f1	acc.	f1
TalkLife	Log. Reg.	58.02	51.58	55.53	41.19	63.23	51.97
	RNN	69.09	54.02	82.25	47.94	73.40	28.22
	HRED	78.91	48.70	79.26	29.48	73.40	28.22
	BERT	76.98	70.31	85.06	62.24	85.87	71.56
	GPT-2	76.89	70.76	80.00	58.43	83.25	65.65
	DialoGPT	76.71	70.42	85.67	66.60	83.95	66.34
	RoBERTa	78.28	71.06	86.25	62.69	85.79	71.83
	<b>Our Model</b>	<b>79.93</b>	<b>74.29</b>	<b>87.50</b>	<b>67.46</b>	<b>86.92</b>	<b>73.47</b>
Reddit	Log. Reg.	41.69	42.69	70.58	49.77	67.08	46.63
	RNN	71.63	42.85	76.21	51.76	85.58	30.74
	HRED	71.11	44.10	79.65	54.16	85.58	30.74
	BERT	72.13	50.41	82.16	61.20	89.35	56.54
	GPT-2	76.69	71.65	82.32	62.27	88.25	58.28
	DialoGPT	66.07	51.16	81.85	<b>68.95</b>	89.65	70.65
	RoBERTa	76.99	70.35	82.16	61.38	90.58	63.41
	<b>Our Model</b>	<b>79.43</b>	<b>74.46</b>	<b>84.04</b>	62.60	<b>92.61</b>	<b>72.58</b>

**Table 2.3:** Empathy identification task results. We observe substantial gains over baselines with our seeker-context aware, multi-tasking approach.

seeker post encoding  $\mathbf{e}_i^{(S)}$ . Our attention is computed as:

$$\mathbf{a}_i(\mathbf{e}_i^{(R)}, \mathbf{e}_i^{(S)}) = \text{softmax} \left( \frac{\mathbf{e}_i^{(R)} \mathbf{e}_i^{(S)}}{\sqrt{d}} \right) \mathbf{e}_i^{(S)} \quad (2.3)$$

where  $d = 768$  (hidden size in RoBERTa<sub>BASE</sub>). We sum the encoded response  $\mathbf{e}_i^{(R)}$  with its representation transformed through attention  $\mathbf{a}_i(\mathbf{e}_i^{(R)}, \mathbf{e}_i^{(S)})$  to obtain a residual mapping (He et al., 2016) –  $\mathbf{h}_i^{(R)}$ , which forms the final seeker-context aware representation of the response post.

**Empathy Identification.** For the task of identifying empathy, we use the final representation of the [CLS] token in the response post ( $\mathbf{h}_i^{(R)}[\text{[CLS]}]$ ) and pass it through a linear layer to get the predictions of the empathy level  $\hat{l}_i$  (0, 1, or 2) of each empathy communication mechanism. Note that we train three independent models for the three communication mechanisms in EPITOME (Section 2.1.2).

**Extracting Rationales.** For extracting rationales  $y_i$  underlying the predictions, we use final representations of the individual tokens in  $\mathbf{R}_i$  ( $\mathbf{h}_i^{(R)}[r_{i1}, \dots, r_{in}]$ ) and pass them through a linear layer

for making boolean predictions,  $\hat{m}_i$ .

**Loss Function.** We use cross-entropy between the true and predicted labels as the loss functions of our two tasks. The overall loss of our multi-task architecture is:  $\mathcal{L} = \lambda_{\text{EI}} * \mathcal{L}_{\text{EI}} + \lambda_{\text{RE}} * \mathcal{L}_{\text{RE}}$ .

**Experimental Setup.** We split both datasets into train, dev, and test sets (75:5:20). We train our model for 4 epochs using a learning rate of  $2e-5$ , batch size of 32,  $\lambda_{\text{EI}} = 1$ , and  $\lambda_{\text{RE}} = 0.5$  (Refer Appendix A.2 for fine-tuning details).

### 2.1.5 Results

Next, we analyze how effectively we can identify empathy with underlying rationales using our computational approach.

#### Overall Results

We compare the performance of our approach with a range of models popularly used in related tasks (e.g., sentiment classification, conversation analysis). We quantify how challenging identifying empathy with underlying rationales is, how well do existing models perform, and what performance is achieved by our proposed approach.

**Baselines.** Our baselines are: **1.** Log. reg. (logistic regression over tf.idf vectors); **2.** RNN (two-layer recurrent neural network); **3.** HRED (hierarchical recurrent encoder-decoder, often used for modeling conversations (Sordoni et al., 2015)); **4.** BERT<sub>BASE</sub> (Devlin et al., 2019); **5.** GPT-2 (typically used for language generation (Radford et al., 2019)); **6.** DialoGPT (GPT-2 adapted to asynchronous conversations (Zhang et al., 2020)); and **7.** RoBERTa<sub>BASE</sub> (Liu et al., 2019).

**Empathy Identification Task.** Table 2.3 reports the accuracy and macro-f1 scores of the three communication mechanisms (random baseline for each is 33% accurate; three levels). Log. reg., RNN, and HRED struggle to identify empathy with noticeably low macro-f1 scores indicative of failures to distinguish between the three levels of communication. Among the baseline transformer architectures, we obtain best performance using RoBERTa but observe substantial gains over them with our approach (+1.73 acc., +4.02 macro-f1 over RoBERTa). We analyze the sources of these gains in the Ablation Study section.

Model	Emotional Reactions		Interpretations		Explorations	
	T-f1	IOU	T-f1	IOU	T-f1	IOU
Log. Reg.	47.44	63.27	46.92	32.97	47.18	62.25
RNN	62.80	58.22	67.26	57.31	63.29	64.65
HRED	60.56	55.01	64.26	70.92	61.54	70.85
BERT	61.29	51.20	61.06	67.33	62.50	64.80
GPT-2	47.39	51.27	64.06	81.12	<b>66.71</b>	78.21
DialoGPT	66.24	61.24	64.05	79.64	57.95	76.95
RoBERTa	59.12	63.82	60.08	84.85	60.05	78.21
<b>Our Model</b>	<b>68.49</b>	<b>66.82</b>	<b>67.81</b>	<b>85.76</b>	64.56	<b>83.19</b>
Log. Reg.	43.26	61.27	49.85	31.31	48.21	70.36
RNN	45.54	43.94	48.22	51.35	65.11	78.27
HRED	46.34	45.65	48.88	52.12	66.66	80.33
BERT	51.06	54.81	48.38	50.75	67.91	71.00
GPT-2	51.44	57.10	54.53	52.38	73.39	82.89
DialoGPT	51.83	49.37	54.43	55.85	<b>73.43</b>	<b>85.20</b>
RoBERTa	51.89	58.31	55.62	54.60	69.76	83.33
<b>Our Model</b>	<b>53.57</b>	<b>64.83</b>	<b>57.40</b>	<b>55.90</b>	71.56	84.48

**Table 2.4:** Rationale extraction task results. We evaluate both at the level of tokens (T-f1) and spans (IOU-f1).

Model	Emotional Reactions				Interpretations				Explorations			
	identification acc.	rationale f1	rationale T-f1	rationale IOU	identification acc.	rationale f1	rationale T-f1	rationale IOU	identification acc.	rationale f1	rationale T-f1	rationale IOU
<b>Our Model</b>	<b>79.93</b>	<b>74.29</b>	<b>68.49</b>	<b>66.82</b>	<b>87.50</b>	<b>67.46</b>	67.81	<b>85.76</b>	<b>86.92</b>	<b>73.47</b>	<b>64.56</b>	<b>83.19</b>
TalkLife -attention	79.00	73.02	59.59	63.49	87.41	66.97	67.12	79.20	84.86	63.45	59.42	73.82
-seeker post	79.37	73.52	61.08	62.58	86.04	63.23	65.56	77.23	86.16	70.80	60.05	81.87
-rationales	79.12	71.21	—*	—*	87.01	66.71	—*	—*	86.38	72.14	—*	—*
-pre-training	78.95	73.41	60.34	62.91	87.31	65.86	<b>69.03</b>	84.95	86.21	70.54	64.53	80.19
<b>Our Model</b>	<b>79.43</b>	<b>74.46</b>	<b>53.57</b>	<b>64.83</b>	<b>84.04</b>	<b>62.60</b>	57.40	<b>55.90</b>	<b>92.61</b>	<b>72.58</b>	<b>71.56</b>	<b>84.48</b>
-attention	75.51	52.66	51.79	59.83	83.26	62.25	54.90	52.79	91.98	64.75	68.81	81.91
-seeker post	79.15	71.47	45.87	58.56	83.57	62.41	55.59	55.51	91.67	64.59	68.73	81.56
-rationales	78.50	73.21	—*	—*	83.26	62.13	—*	—*	91.51	64.44	—*	—*
-pre-training	76.97	69.03	51.58	57.35	82.32	61.38	<b>57.61</b>	55.34	91.99	65.26	70.44	81.71

**Table 2.5:** Ablation results. Most of our gains are due to context provided through attention and seeker post; higher gains for the rationale extraction task. \*Note that rationales cannot be predicted after removing them from training.

**Rationale Extraction Task.** We perform both token level and span level evaluation for this task. We use two metrics, commonly used in discrete rationale extraction tasks (DeYoung et al., 2020): 1. T-f1 (token level f1); 2. IOU-f1 (intersection over union overlap of predicted spans with ground truth spans; threshold of 0.5 on the overlap for finding true positives and the corresponding f1). We find that GPT-2 and DialoGPT perform better than BERT and RoBERTa likely due to appropriateness to the related task of generating free-text rationales (Table 2.4). Our approach obtains gains of +2.58 T-f1 and +6.45 IOU-f1 over DialoGPT, potentially due to the use of attention and seeker post (see the Ablation Study section).

## Ablation Study

We next analyze the components and training strategies in our approach through an ablation study.

**No Attention.** Instead of using attention, we concatenate the seeker post encoding ( $e_i^{(S)}$ ) with the response post encoding ( $e_i^{(R)}$ ) and use the concatenated representation as input to the linear layer.

**No Seeker Post.** We train without the S-Encoder, i.e., by only encoding from the R-Encoder.

**No Rationales.** We set  $\lambda_{RE}$  to 0 and only train on the empathy identification task.

**No Domain-Adaptive Pre-training.** We initialize by only using model weights from RoBERTa<sub>BASE</sub>.

**Results.** Our most significant gains come from using attention and the seeker post (Table 2.5) which greatly benefits the rationale extraction task (+4.88 T-f1, +5.74 IOU-f1). Also, using rationales and pre-training only leads to small performance improvements.

## Error Analysis

We qualitatively analyze the sources of our errors. For the empathy identification task, we found that the model sometimes failed to identify short expressions of emotions in responses that otherwise contained a lot of instructions (e.g., *Sorry to hear that! Try doing ...*). Also, certain responses trying to universalize the situation (e.g., *You are not alone*) got incorrectly identified as strong interpretations. Furthermore, a source of error for explorations was confusions due to

questions that were not an exploration of seeker’s feelings or experiences (e.g., offers to talk - *Do you want to talk?*).

For the rationale extraction task, the model accurately extracts phrases that are commonly used for expressing empathy (e.g., *this might be tough*), but also sometimes incorrectly extracts single words from sentences indicating errors in disambiguating word usage (e.g., ‘what’ in *Tell them what you need* gets extracted as an exploration, likely because *what* is really common in explorations as in *what happened?*).

### 2.1.6 Model-based Insights into Mental Health Platforms

We apply our model to study how empathy impacts online peer-to-peer support dynamics. To only focus on conversations related to significant mental health challenges and filter out common social media interactions (e.g., *Merry Christmas*), we carefully select 235k mental health related interactions on TalkLife using a seeker-reported indicator.<sup>6</sup>

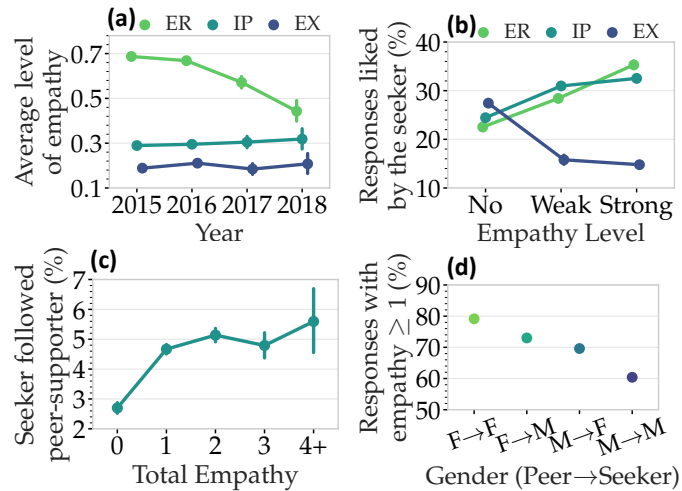
We investigate (1) the levels of empathy on the platform, its variation over time, and examine the relationship of empathy with (2) conversation outcomes, (3) relationship forming, and (4) gender.

**(1) Peer supporters do not self-learn empathy over time.** Overall, we observe that empathy expressed by peer supporters on the platform is low (avg. total score<sup>7</sup> of 1.09 out of 6). In addition, we find that the emotional reactivity of users decreases over time (36% decrease over three years) and their levels of interpretations and explorations remain practically constant (Figure 2.2a). Further analyzing whether a user individually improves, worsens, or remains constant in their expression of empathy, we find that 69% users either worsen or stay constant in their empathy expressions and only 10% improve by at least one point (i.e. one level in our framework). This is also reflected in prior work on therapy that shows that without deliberate practice and specific feedback, even trained therapists often diminish in skills over time (Goldberg et al., 2016). We find this trend robust to potential confounding factors (new users, user dropout) and users of different groups (low vs. high activity users, moderators; Appendix A.3). This indicates that most users do not self-learn empathy and highlights the need of providing them feedback.

---

<sup>6</sup>We focus analyzing TalkLife alone as Reddit lacks rich publicly available signals such as seeker liking the response.

<sup>7</sup>Total empathy score is obtained by adding the level of communication across the three mechanisms.



**Figure 2.2:** (a) Peer-supporters do not self-learn empathy over time. Only users who joined in 2015 were included but similar trends hold for other user groups; (b) Stronger communications of emotional reactions and interpretations are received positively by seekers. Stronger explorations get 47% more replies; (c) A lot more seekers follow peers after empathic interactions; (d) Females are more empathic towards females.

**(2) High empathy interactions are received positively by seekers.** We analyze the correlation of empathic conversations with positive feedback, concretely with seeker "liking" the post. We find that strong communications of empathy are received with 45% more likes by seekers than no communication (Figure 2.2b). Strong explorations get 44% less likes but receive 47% more replies than no explorations, leading to higher engagement.

**(3) Relationship forming more likely after empathic conversations.** Psychology research emphasizes the importance of empathy in forming alliance and relationship with seekers (Watson, 2007). Here, we operationalize relationship forming as seeker "following" the peer supporter after a conversation (within 24hrs). We find that seekers are 79% more likely to follow peer supporters after an empathic conversation (total score of 1+ vs. 0) than after a non-empathic one (Figure 2.2c).

**(4) Females are more empathic with females than males are with males.** Previous work has shown that seekers identifying as females receive more support in online communities (Wang and Jurgens, 2018). Here, we ask if empathic interactions are affected by the self-reported gender of seekers and peer supporters. We find that female peer supporters are 32% more empathic towards female seekers than males are towards male seekers (Figure 2.2d). Also, females are 6% more empathic towards males than males are towards females.

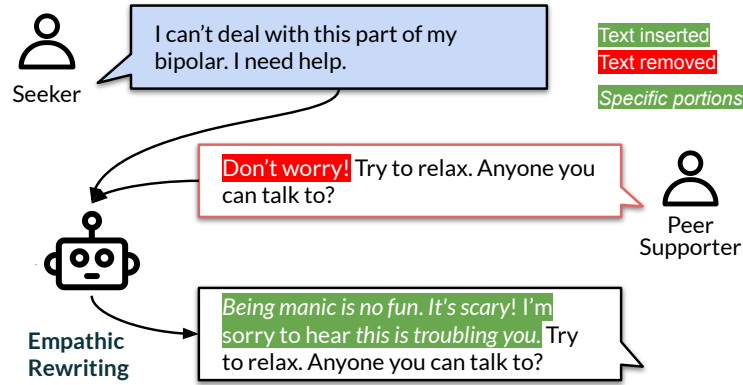
### 2.1.7 Implications for Empathy-Based Feedback

These results suggest that our approach not only successfully measures empathy according to a principled framework (Section 2.1.2), but that the measured empathy components are important to online supportive conversations as indicated by the positive reactions from seekers and meaningful reflections of social theories. However, peer supporters on the platform express empathy rarely and this does not improve over time. This points to critical opportunities for empathy-based feedback to peer supporters for making their interactions with seekers more effective.

## 2.2 Empathic Rewriting: Computationally Transforming Low-Empathy Conversations into Higher Empathy

To make conversations more empathic, we introduce *empathic rewriting*, a new task that aims to transform low-empathy conversational posts to higher empathy (Figure 2.3). For example, given a post from a support seeker *"I can't deal with this part of my bipolar. I need help."* and a low-empathy response *"Don't worry! Try to relax. Anyone you can talk to?,"* we want to increase the empathy quotient in the response by transforming it to *"Being manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?"*; the rewritten response should communicate more empathy through an understanding of feelings and experiences (*"Being manic is no fun. It's scary"*) and display of felt emotions (*"I'm sorry to hear this is troubling you"*).

Performing such transformations is a challenging task: First, empathy is a complex, conceptually nuanced construct and requires understanding the feelings and experiences shared by the support seeker. In the example above, one needs to understand that being *"bipolar"* can be *"scary"*, involves *"manic"* phases, and communicate this in the response. Second, for empathic rewriting to be purposeful, it should not undermine other conversation goals like language fluency, context specificity, and diversity. Making changes that lead to ungrammatical posts with empathic portions (e.g., *"Scary it is manic being"*) may not be helpful and obstruct useful feedback. Further, making the same transformation to every response (e.g., rewrite every response to *"I understand how you feel"*) would lead to non-specific and generic responses reducing the overall conversational quality (See et al., 2019; Li et al., 2016a). Third, the task of empathic rewriting re-



**Figure 2.3:** An overview of the empathic rewriting task. Given a post from support seeker and a low-empathy response, the task is to rewrite the response for making it more empathic, through text **insertions** and **deletions**. This task requires inferring *specific* feelings and experiences from seeker’s post and using them for making appropriate changes to the response through empathic mechanisms like emotional reactions, interpretations, and explorations. *Examples have been paraphrased for anonymization (Matthews et al., 2017).*

quires changes that go beyond simple word-level transformations, often requiring multiple new sentences to be added or replaced (e.g., three sentence insertions and one sentence removal in the example in Figure 2.3). This is different from related style transfer tasks (Shen et al., 2017; Li et al., 2018) where even changing a single word may suffice for transferring from negative to positive sentiment (e.g., replace “bad” with “good” in the sentence “the movie was bad”). Finally, supervised methods commonly used for similar tasks such as style transfer (Shen et al., 2017; Li et al., 2018) and content debiasing (Pryzant et al., 2020; Ma et al., 2020) usually require a large parallel dataset. Such a dataset is not yet available for empathic rewriting and hard to collect as it would require a large number of clinical psychologists and counselors well-versed in the complex construct of empathy.

To address the challenges described above, we propose PARTNER,<sup>8</sup> a deep reinforcement learning (RL) model for the task of empathic rewriting (Section 2.2.3). We design an RL agent which learns to add new empathic sentences to posts or replace existing sentences in posts with more empathic ones. The agent operates on a pair of seeker post and the original response post (which rarely is highly empathic (Section 2.1.6)) and makes edits to the response at the level of a sentence by simultaneously (a) identifying positions in the original response post where changes are required, and (b) generating empathic sentences for insertion or replacement at the identified

<sup>8</sup>emPAthic RewriTing in meNtal hEalth suppoRt

positions (Section 2.2.3). We model this agent using a policy network based on a transformer decoder model adapted from GPT-2 (Radford et al., 2019). We build upon existing large-scale pre-training of GPT-2 on conversations, as done in DialoGPT (Zhang et al., 2020), and modify it to perform the two simultaneous actions of identifying positions and generating empathic sentences for empathic rewriting (Section 2.2.3). Through carefully constructed scoring functions, we reward transformations that increase empathy in posts while maintaining text fluency, context specificity, and diversity (Section 2.2.3).

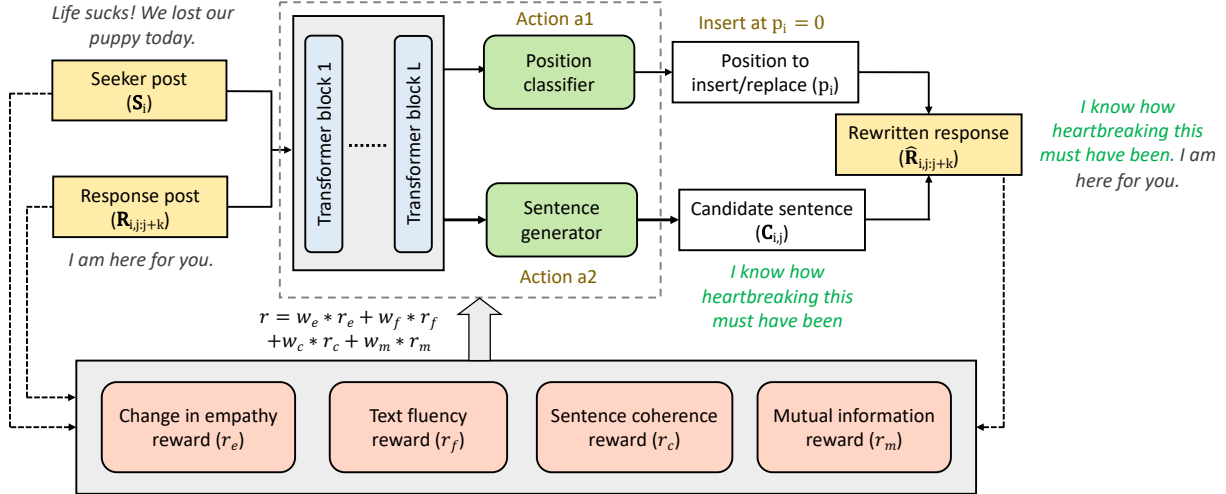
Evaluating complex conversational constructs such as empathy is fundamentally challenging. Therefore, we combine comprehensive automatic evaluation with expert-based human evaluation. Our experiments demonstrate that PARTNER can effectively increase empathy in posts in fluent, specific, and diverse ways and outperforms baselines used in related text generation tasks by  $> 35\%$  in empathy improvement (Section 2.2.4). Also, PARTNER is the only approach that consistently improves empathy and does not lead to a *loss* of empathy when rewriting an already highly empathic post, while all baselines tend to propose a large number of edits that only make the situation worse (Section 2.2.4). Lastly, through comprehensive human evaluation, we show that experts in clinical psychology prefer rewritings of PARTNER compared to baselines, based on empathy, specificity, and fluency (Section 2.2.4). We view our approach and findings as a key step towards building AI systems for facilitating empathic conversations on online mental health support platforms, but these insights may generalize beyond mental health to other conversational settings on web-based platforms. We share our code publicly at <https://github.com/behavioral-data/PARTNER>.

## 2.2.1 Problem Definition and Goals

In this section, we formulate the task of empathic rewriting and state the associated goals.

### Empathic Rewriting

We introduce *empathic rewriting*, a new task that aims to transform low-empathy conversational posts to higher empathy. In contrast with empathic dialogue generation (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020), where the objective is to generate empathic posts from scratch, this task requires making changes to existing posts in order to make them empathic. This



**Figure 2.4:** PARTNER uses a deep reinforcement learning approach for Empathic Rewriting. It leverages a transformer language model for performing the two actions of (1) selecting positions for insertion or replacement and (2) generating candidate empathic sentences. It uses four reward functions that promote increase in empathy, text fluency, sentence coherence, context specificity, and diversity.

is more consistent with realistic use-cases in difficult, high-stakes settings such as online support systems, which are likely to augment, rather than replace humans (Miner et al., 2019).

Formally, let  $S_i$  be a seeker post and  $R_i$  be a corresponding response post. We aim to transform  $R_i$  into its more empathic counterpart  $\hat{R}_i$ .

## Goals

For empathic rewriting to be useful in improving mental health support conversations, the rewriting process should achieve specific goals related to empathy, conversation and natural language generation quality, and purposeful and precise feedback:

**Theoretically-grounded empathy.** Empathy is complex and conceptually nuanced; over time psychology research has emphasized multiple aspects of empathy (Bohart and Greenberg, 1997; Duan and Hill, 1996; Batson, 2009; Davis et al., 1980). For example, computational research typically defines empathy as reacting with emotions of warmth and compassion (Buechel et al., 2018). However, psychotherapy research emphasizes aspects of empathy related to communicating cognitive understanding of feelings and experiences of others (Selman, 1980). For empathic rewriting to be useful and potentially adopted in online mental health support, we need to design methods grounded in psychology and psychotherapy research. Here, we adopt the

theoretically-grounded framework of empathy designed in Section 2.1.2. We leverage empathy measurements based on this framework as (1) reward signals in our model for empathic rewriting (Section 2.2.3), and (2) an automatic evaluation metric for judging improvements in empathy from various rewriting models (Section 2.2.4).

**Context specificity and response diversity.** Consider a rewriting approach that transforms every response to a generic but empathic response (e.g., *“That must have been really hard for you”*). While this approach may seem to “solve” empathic rewriting, it suffers from two key issues. First, the responses generated by this approach would lack specificity to the emotions and experiences shared in the seeker post, which is important for empathy and effective mental health support (Robert et al., 2011; Majumder et al., 2020). Second, performing this same transformation to millions of responses on online platforms would dramatically reduce response diversity which has been shown to be important for mental health support (Althoff et al., 2016) as well as in general dialogue research (See et al., 2019; Li et al., 2016a).

Thus, the task of empathic rewriting interplays with other issues related to conversation and natural language generation quality and effective mental health support. Ensuring that the rewritten response is specific and diverse, along with empathic is challenging but critical for obtaining purposeful transformations. In this work, we learn rewriting actions that simultaneously achieve the goals of context specificity and response diversity using a reinforcement learning approach (Section 2.2.3) and we evaluate these goals using a combination of automatic and human evaluation (Section 2.2.4,2.2.4).

**Text fluency and sentence coherence.** In addition, only generating empathic words or phrases may not be sufficient. Without appropriate measures, the rewriting process may lead to an ungrammatical, non-fluent final response (e.g., *“Scary being is it manic”*). Also, making changes that are incoherent with the original response may not be appropriate (e.g., changing *“Sorry to hear that you lost your job. I hope you get a new job soon.”* to *“Sorry to hear that you lost your job. Congrats on your job promotion. I hope you get a new job soon.”*). Here, we avoid such responses with non-fluent and incoherent portions through carefully constructed reward functions (Section 2.2.3) and conduct both automatic and human evaluations of models on text fluency and sentence coherence (Section 2.2.4,2.2.4).

**Rewriting for feedback and training.** An important way in which the task of empathic rewriting can be used is for providing feedback and training to people through machine-in-the-loop writing systems (Clark et al., 2018; Tanana et al., 2019). For humans to adopt such feedback, however, the rewriting process should make changes that are precise and specific to the original response. This means that the number of changes should be kept minimal and that the changes themselves should be suitable to the original response. For example, adding 10 sentences to a one-sentence response may not be useful. Here, we train a reinforcement learning agent which learns when to stop making changes through a special “stopping” action (Section 2.2.3). We evaluate the number of transformations different models need for empathic rewriting through a standard edit-distance based scoring metric (Section 2.2.4).

## 2.2.2 Data Description

In this section, we describe the dataset used for the task of empathic rewriting.

### The TalkLife platform

Table 2.6 describes the statistics of conversational threads on the TalkLife platform.

**Curating mental health-related conversations.** The TalkLife platform hosts a significant number of common social media interactions (e.g., *Happy mother’s day*). Here, we focus our analyses on mental health-related conversations and filter out such posts. We manually annotate  $\sim 3\text{k}$  posts with answers to the question “*Is the seeker talking about a mental health related issue or situation in his/her post?*”. Using this annotated dataset, we train a standard text classifier based on BERT (Devlin et al., 2019) (achieving an accuracy of  $\sim 85\%$ ). We apply this classifier to the entire TalkLife dataset and create a filtered dataset of mental health-related conversations. This dataset contains 3.33M interactions from 1.48M seeker posts.

### Creating a Dataset of Empathic Posts

Training supervised methods would require a large parallel dataset of corresponding pairs of posts with low and high empathy, respectively. As empathy is a complex phenomenon, collecting such a dataset is challenging and would likely require psychology experts. Here, we create

Dataset Statistics	TalkLife
# of Seeker posts	10.9M
# of Response posts	26.9M
# of Users	642K
Observation Period	May 2012 to June 2020

**Table 2.6:** Statistics of the TalkLife dataset.



**Figure 2.5:** Expression of high levels of empathy is very low in online support platforms, especially for Interpretations (IP) and Explorations (EX). Emotional reactions (ER) are slightly more common.

a large non-parallel dataset with empathy measurements for training unsupervised and self-supervised computational models and a small parallel dataset with expert empathic rewritings for conducting evaluations.

**Computational labeling with empathy measurements.** We computationally label our dataset of 3.33M interactions with empathy measurements using the classifier developed in Section 2.1. This framework consists of three empathy communication mechanisms – (1) *Emotional Reactions* (expressing emotions such as warmth, compassion), (2) *Interpretations* (communicating an understanding of feelings and experiences), and (3) *Explorations* (improving understanding of the seeker by exploring feelings and experiences). For each communication mechanism, the authors design a three-point scale (0 to 2). We computationally label all pairs of (seeker post, response post) in our dataset based on this empathy scale. Figure 2.5 shows the statistics which indicate that high levels of empathy expressions are uncommon in online support platforms, highlighting the need for building systems for improving empathy (e.g., through feedback using empathic rewriting (Section 2.2.1)). We use this dataset for a supervised warm-start training in our reinforcement learning model (Section 2.2.3) and for training unsupervised baselines (Section 2.2.4).

**Expert empathic rewritings.** Additionally, we create a small parallel dataset of 180 pairs of

corresponding low and rewritten high empathy response posts with rewritings from people having substantial expertise in empathy, mental health, and therapy (six graduate students in clinical psychology; none are co-authors). We showed them pairs of seeker and response posts and asked them to modify the response post for improving it in empathy. This expert-based dataset is designed to represent the best possible responses and we use it as ground truth for evaluation (Section 2.2.4).

## Privacy, Ethics, and Disclosure

The dataset was sourced with license and consent from the TalkLife platform. All personally identifiable information (user and platform identifiers) in our dataset was removed. This work was approved by University of Washington’s Institutional Review Board. We do not make any treatment recommendations or diagnostic claims.

**Towards preventing unsafe rewritings.** We acknowledge that building computational models for intervention in high-stakes settings such as mental health necessitates ethical considerations. There is a risk that in attempting to help, responses could have the opposite effect, which could be deadly in cases of self-harm. No current computational approach will identify and respond to harm-related utterances perfectly (Miner et al., 2020). Thus, risk mitigation steps are appropriate in this context. Here, we remove all posts that contain a pre-defined unsafe regular expression (e.g., *\*commit suicide\**) from our analyses and training in collaboration with mental health professionals. Future work testing or deploying AI systems should assess safety-related risk, and also potential sources of bias (e.g., race, ethnicity, age, or gender bias in training data or models).

### 2.2.3 Partner: Empathic Rewriting Using Reinforcement Learning

Here, we present PARTNER, a reinforcement learning model for the task of empathic rewriting. We first explain the general reinforcement learning framework and its applicability to our setting. We then describe the various components of our model (states, actions, policy, and rewards) and our training strategy.

## Reinforcement Learning Framework

We adopt the standard reinforcement learning framework consisting of a collection of states  $\mathcal{S}$ , a set of actions  $\mathcal{A}$ , a policy  $\pi$ , and rewards  $\mathcal{R}$  (Sutton and Barto, 2018). In this framework, given a state  $s \in \mathcal{S}$ , an agent takes an action  $a \in \mathcal{A}$  according to the policy  $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ . The policy defines whether the agent should take action  $a$  in a state  $s$ . The goal of the reinforcement learning agent is to learn a policy which maximizes the reward  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$ .

Here, we design a reinforcement learning model for the task of empathic rewriting. Conceptually, our agent leverages context from the seeker post which it uses for making specific empathic changes. Alongside, it operates on the response post, looks for areas where empathy could be improved, and works on those improvements in fluent, coherent, specific, and diverse ways. Moreover, it ensures that the changes are minimal and precise by learning when to stop through a special “stopping” action.

In our reinforcement learning model, we construct states based on seeker posts and fixed-length contiguous spans in the associated response posts (Section 2.2.3). Insertion, replacement, and deletion of sentences in response posts are defined as actions (Section 2.2.3). We learn a policy that uses transformer language models at its core (Section 2.2.3). We design a reward function that favors empathic, fluent, coherent, specific, and diverse transformations (Section 2.2.3).

### State: Seeker Post & Fixed-Length Contiguous Spans of Response Post

PARTNER simultaneously operates on the seeker post and fixed-length contiguous spans of the response post. The use of seeker post helps us in leveraging conversational context, thereby enabling transformations that are specific to the feelings and experiences shared in the seeker post. The response post is used for making transformations. The use of fixed-length contiguous spans enables a static action set.

Formally, let  $\mathbf{R}_i$  contain  $n$  sentences  $\mathbf{R}_{i,1}, \dots, \mathbf{R}_{i,n}$ . At each step, we focus on a contiguous window of  $k$  sentences starting from the  $j$ th sentence  $\mathbf{R}_{i,j:j+k} = \mathbf{R}_{i,j}, \dots, \mathbf{R}_{i,j+k-1}$ . Then, our state  $s \in \mathcal{S}$  is denoted by the pair  $(\mathbf{S}_i, \mathbf{R}_{i,j:j+k})$ . Our policy uses a string containing  $\mathbf{S}_i$  concatenated with  $\mathbf{R}_{i,j:j+k}$  separated by a special <SPLIT> token (as commonly used in BERT-like models (Devlin et al., 2019)).

## Actions: Sentence-level Edits

Our agent takes actions at the level of a sentence, i.e. it either inserts new sentences or replaces existing sentences with newer ones. A deletion operation is equivalent to replacing a sentence with an empty string. Our agent can make word-level changes by replacing the original sentence with a slightly different sentence containing only word-level edits. We focus on sentence-level edits because the task of empathic rewriting requires changes that go beyond simple word-level edits. Empathic responses typically contain multiple sentences with different goals such as emotional reactions, interpretations, and explorations (Section 2.1.2); generating these sentences and using them for making changes to the response is important for empathic rewriting.

In a state  $(\mathbf{S}_i, \mathbf{R}_{i,j:j+k})$ , our agent simultaneously takes two actions –  $(a_1)$  select a position in  $\mathbf{R}_{i,j:j+k}$  for insertion or replacement,  $(a_2)$  generate a candidate empathic sentence. The action space  $\mathcal{A}_1$  of  $a_1$  consists of  $2k+2$  actions –  $k+1$  positions for insertions,  $k$  positions for replacements, and one *special* action for no insertion or replacement, which stops the agent from making any further changes. The action space  $\mathcal{A}_2$  of  $a_2$  consists of all arbitrary-length sentences. We denote the action taken by our agent as  $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ .

## Policy

At its core, our policy has a transformer language model consisting of a stack of masked multi-head self-attention layers, based on GPT-2 (for a detailed description, see Vaswani et al. (2017), Radford et al. (2019)). It takes as input an encoded representation of our state  $(\mathbf{S}_i, \mathbf{R}_{i,j:j+k})$  and generates the action  $a = (a_1, a_2)$ .

**$(a_1)$  Selecting a position for insertion or replacement.** Given  $(\mathbf{S}_i, \mathbf{R}_{i,j:j+k})$  as input, we want to identify a position  $\mathbf{p}_i$  in  $\mathbf{R}_{i,j:j+k}$  where changes need to be made for improving empathy through insertion or replacement operations. A  $k$  sentence window  $\mathbf{R}_{i,j:j+k}$  has  $k + 1$  positions for insertions and  $k$  positions for replacement. Then, our task is to select one of these  $2k + 1$  positions. We formulate this as a classification problem with  $2k + 2$  classes. The first  $2k + 1$  classes represent one of the  $2k + 1$  potential positions and the last class represents the "stopping" action of not selecting any position, thereby stopping the agent from making any changes and keeping the response span unchanged.

For selecting this position, we first encode the input string " $\mathbf{S}_i$  <SPLIT>  $\mathbf{R}_{i,j+k}$ " using the transformer block of GPT-2. We then pass this encoded representation through a linear layer to get the prediction  $\hat{\mathbf{p}}_i$  of the position for insertion or replacement. We denote our position classifier as  $p_{\text{pos}}$ .

**(a<sub>2</sub>) Generating a candidate sentence.** Given  $(\mathbf{S}_i, \mathbf{R}_{i,j+k})$  as input, we want to generate a candidate sentence  $\mathbf{C}_{i,j}$  to be used for making changes to  $\mathbf{R}_{i,j+k}$ . We frame this task as a language modeling problem where the objective is to generate  $\mathbf{C}_{i,j}$  that maximizes the conditional probability  $p_{\text{sent}}(\mathbf{C}_{i,j}|\mathbf{S}_i, \mathbf{R}_{i,j+k})$ .

Similar to the position selection action, we first encode our input string " $\mathbf{S}_i$  <SPLIT>  $\mathbf{R}_{i,j+k}$ " using the transformer block of GPT-2. We then compute a probability distribution over vocabulary tokens by transforming the encoded representation into a vocabulary-sized vector through a softmax layer. Finally, we use top-p sampling (Holtzman et al., 2020)<sup>9</sup> over this probability distribution to generate the desired  $\mathbf{C}_{i,j}$ . The generation is terminated when the sampling process encounters a special end-of-sequence token.

## Rewards

Our reward functions aim to increase empathy in posts and maintain text fluency, sentence coherence, context specificity, and diversity:

**Change in empathy.** The task of empathic rewriting requires transformations that can increase empathy of posts. Thus, we want to reward actions that increase empathy of  $\mathbf{R}_i$  and penalize actions that decrease empathy of  $\mathbf{R}_i$ . Let  $f_e(\cdot)$  be a function that measures empathy of posts. Then, the change in empathy reward,  $r_e$ , is defined as:

$$r_e = f_e(\hat{\mathbf{R}}_i) - f_e(\mathbf{R}_i) \quad (2.4)$$

Here, we estimate  $f_e(\cdot)$  using the empathy classification model developed in Section 2.1 for predicting empathy levels of responses. We use the model as  $f_e(\cdot)$  which gives us empathy scores of  $\hat{\mathbf{R}}_i$ s in the range of 0 to 6.

---

<sup>9</sup>For generating every word in a sequence, top-p sampling (or nucleus sampling) chooses from the smallest set of words whose total probability is more than p.

**Text fluency.** We want to prevent actions that lead to outputs that are highly empathic but not fluent or grammatically correct. Therefore, we want to reward actions that lead to fluent outputs and penalize actions resulting in non-fluent outputs. Here, we operationalize *text fluency* as the inverse of perplexity of the generated  $\hat{\mathbf{R}}_i$ s. We define the text fluency reward,  $r_f$  as:

$$r_f = p_{\text{LM}}(\hat{\mathbf{R}}_i)^{(1/N)} \quad (2.5)$$

where  $p_{\text{LM}}$  is a general language model for English and  $N$  is the number of words in  $\hat{\mathbf{R}}_i$ . Here, we use GPT-2 (Radford et al., 2019) as our  $p_{\text{LM}}$ , following previous work (Ma et al., 2020; Dai et al., 2019).

**Sentence coherence.** A key component of our action space is the addition of the candidate sentence to the original response. While the candidate sentence might be highly empathic and fluent, it may not be well-suited for the response  $\mathbf{R}_i$  to which it would be added, leading to incoherent sentences in the transformed response  $\hat{\mathbf{R}}_i$ . This may not be handled by perplexity which tends to give high scores to posts where individual sentences are all fluent but are not coherent at the macro response level. Here, we design a reward function,  $r_c$  that measures coherence of the candidate sentence  $\mathbf{C}_{i,j}$  with the response span  $\mathbf{R}_{i,j:j+k}$ .  $r_c$  measures the average sentence coherence probability between a candidate sentence and existing sentences in the response.

First, we create a dataset of likely coherent and incoherent sentence pairs. Given two sentences  $\mathbf{R}_{i,j_1}$  and  $\mathbf{R}_{i,j_2}$  in a response  $\mathbf{R}_i$ , we call  $(\mathbf{R}_{i,j_1}, \mathbf{R}_{i,j_2})$  a *potential coherent sentence pair*. We randomly sample a sentence  $\mathbf{R}'$  which is not a part of responses posted to the current seeker post  $\mathbf{S}_i$  and call  $(\mathbf{R}', \mathbf{R}_{i,j})$  a *potential incoherent sentence pair* ( $\forall \mathbf{R}_{i,j} \in \mathbf{R}_i$ ). Next, we train a text classification model, based on BERT (Devlin et al., 2019), on this dataset. We take softmax at the last layer which gives us probabilities of a sentence pair being coherent ( $p_{\text{coherent}}$ ) or incoherent ( $p_{\text{incoherent}}$ ). Then, our sentence coherence reward is defined as:

$$r_c = \frac{\sum_{l=j}^{l=j+k} p_{\text{coherent}}(\mathbf{C}_{i,j}, \mathbf{R}_{i,l})}{k} \quad (2.6)$$

**Mutual information for specificity and diversity.** In the process of empathic rewriting, the final

rewritten response may become generic (e.g., “I understand how you feel”) thereby affecting the overall conversation quality (See et al., 2019; Li et al., 2016a). In order to ensure specificity to the seeker post and diversity of responses, we exploit the idea of maximizing mutual information between seeker post and the rewritten response post (Li et al., 2016a,b). Our mutual information reward is:

$$r_m = \lambda_{\text{MI}} * \log \vec{p}(\hat{\mathbf{R}}_i | \mathbf{S}_i) + (1 - \lambda_{\text{MI}}) * \log \overleftarrow{p}(\mathbf{S}_i | \hat{\mathbf{R}}_i) \quad (2.7)$$

where  $\vec{p}$  is the transformer language model used in our policy and  $\overleftarrow{p}$  is an identical language model for performing the reverse task of generating seeker post from the rewritten response.

**Total reward.** Our total reward is  $r = w_e * r_e + w_f * r_f + w_c * r_c + w_m * r_m$ .

## Optimization and Training

**Warm-start using supervised learning.** We use the pre-trained weights of DialoGPT (Zhang et al., 2020) for initializing our transformer language model. Next, we use a warm-start strategy using supervised learning on a parallel dataset of (low empathy, high empathy) pairs, following previous work in reinforcement learning for dialogue generation (Li et al., 2016b). For creating this dataset, we follow the reverse process of making highly empathic responses less empathic by removing sentences that are high in empathy. Similar "reverse-engineering" strategy has also been shown to work well in other complex linguistic phenomenon like humor (West and Horvitz, 2019). We first identify highly empathic sentences (with scores  $\geq 2$ ) in our dataset of empathic interactions (Section 2.2.2). For a seeker post  $\mathbf{S}_i$  and response post  $\mathbf{R}_i$  having a highly empathy sentence  $\mathbf{R}_{i,j}$ , we create a dataset with  $(\mathbf{S}_i \langle \text{SPLIT} \rangle \mathbf{R}_i, \mathbf{R}_i - \mathbf{R}_{i,j})$  pairs.<sup>10</sup> We use this dataset to finetune our DialoGPT-initialized transformer language model.

**REINFORCE with a baseline value for training.** We use the standard REINFORCE algo-

---

<sup>10</sup> $\mathbf{R}_i - \mathbf{R}_{i,j}$  refers to the full response post  $\mathbf{R}_i$  with the sentence  $\mathbf{R}_{i,j}$  removed.

	Model	Change in empathy (↑)	Perplexity (↓)	Specificity (↑)	Diversity (↑)		Sentence coherence (↑)	Edit rate (↓)
					distinct-1	distinct-2		
Dialogue Generation	DialoGPT	0.4698	8.6500	0.8921	0.0382	0.1334	0.6683	1.3520
	MIME	1.2069	9.0171	0.8837	0.0031	0.0198	0.3687	1.8193
Seq-to-Seq Generation	Latent Seq.	0.9745	8.7143	0.8512	0.0001	0.0002	<b>0.9252</b>	7.8853
	BART	-0.0611	<b>7.2040</b>	0.8878	<b>0.0722</b>	<b>0.3945</b>	0.4560	<b>0.7496</b>
<b>Partner</b>		<b>1.6410</b>	7.3641	<b>0.9052</b>	0.0659	0.3807	0.3030	0.9654

**Table 2.7:** Performance of PARTNER and comparisons with dialogue generation and other sequence-to-sequence generation baselines on the set of automatic metrics. PARTNER outperforms all baselines in empathy improvement and generates fluent, specific, and diverse outputs with lower edits. (↑) indicates higher is better, (↓) indicates lower is better.

rithm (Williams, 1992) for training our agent. Our loss function is defined as:

$$\begin{aligned}
 J(\theta) = & -(r - b) * (\log p_{\text{pos}}(a_1 | \mathbf{S}_i, \mathbf{R}_{i,j:j+k}) \\
 & + \log p_{\text{sent}}(a_2 | \mathbf{S}_i, \mathbf{R}_{i,j:j+k}))
 \end{aligned}
 \tag{2.8}$$

where  $\theta$  is our set of parameters and  $b$  is a baseline estimate of the reward (running average of previous 100 reward values) used for stabilizing training.

**Experimental setup.** We use a batch size of 16 and train our model for 20000 steps using a learning rate of 1e-5. We use  $w_e = 1.0$ ,  $w_f = 10.0$ ,  $w_c = 0.1$ , and  $w_m = 0.1$  (selected using a grid-search approach with three values (0.1, 1.0, 10.0) for each hyperparameter). Moreover, we choose  $k = 2$ ,  $p = 0.92$ , and  $\lambda_{\text{MI}} = 0.5$ . We truncate both seeker and response post to 64 tokens each.

## 2.2.4 Partner : Experiments

Next, we present experiments for analyzing the performance of PARTNER on the task of empathic rewriting. We first describe automatic evaluation metrics (Section 2.2.4) based on the desired goals for empathic rewriting (Section 2.2.1), baseline approaches and ablations (Section 2.2.4), and demonstrate results on the automatic evaluation metrics (Section 2.2.4). Since evaluation using automated metrics in language generation tasks are often not robust (Liu et al., 2016), we additionally present human evaluation results from people having expertise in therapy and mental health (Section 2.2.4). We end with a qualitative discussion on the model’s performance

(Section 2.2.4).

## Automatic Evaluation Metrics

We use a number of automatic metrics that are based on the goals associated with empathic rewriting (Section 2.2.1):

- **Change in empathy:** A key metric for successful empathic rewriting is how much the empathy has changed from the original response to the rewritten response. Similar to our reward function (Section 2.2.3), we measure this change using the empathy classification model developed in Section 2.1. The model computes empathy scores in the range 0 to 6 (leading to change of empathy ranging from -6 to 6).
- **Perplexity:** Similar to our text fluency reward (Section 2.2.3), we measure perplexity for quantifying fluency of the rewritten responses. For this, we use a pre-trained GPT-2 language model that has not been fine-tuned on our dataset, following previous work (Ma et al., 2020; Dai et al., 2019).
- **Sentence coherence:** Since empathic rewriting requires changes at the sentence level, ensuring coherent sentences in the final rewritten response is crucial. Here, we measure sentence coherence using the scoring mechanism developed in Section 2.2.3.
- **Specificity:** The rewritten response should be specific to the seeker post. Following Xu et al. (2018), we measure specificity using word embedding similarity between seeker post and rewritten response post (using embeddings from BERT (Devlin et al., 2019)).
- **Diversity:** Since empathic rewriting has implications on millions of conversations on online mental health platforms, ensuring diversity of responses is important. Here, we measure diversity using the *distinct-1* and *distinct-2* metrics, following Li et al. (2016a). The two metrics compute the number of distinct unigrams and bigrams respectively divided by the total number of tokens.
- **Edit rate:** The changes in empathic rewriting should be minimal and precise. Here, we use edit rate (Snover et al., 2006) to measure the number of changes between the original re-

sponse and the rewritten response. Edit rate is defined by the Levenshtein distance between the two responses divided by the length of the original response.

## Baselines and Ablations

As the task of empathic rewriting has not been explored before, we compare against baseline approaches from the related tasks of dialogue generation and style transfer. Our baselines are:

- **DialoGPT (Zhang et al., 2020):** A large dialogue generation model, based on GPT-2 (Radford et al., 2019) and pre-trained on Reddit conversations.
- **MIME (Majumder et al., 2020):** An empathic dialogue generation model which exploits emotion mimicking while accounting for emotion polarity (positive or negative).
- **Deep latent sequence model (He et al., 2019):** A deep generative model designed for unsupervised style transfer.
- **BART (Lewis et al., 2019):** An encoder-decoder model for sequence-to-sequence language generation.

DialoGPT and MIME baselines completely disregard the original response; the rewritten response is the response generated given a seeker post by the respective dialogue generation models. Deep latent sequence model and BART perform a sequence-to-sequence generation from a (seeker post, original response post) pair to a response with higher empathy. We use publicly available implementations of all our baselines. We further fine-tune deep latent sequence model on the dataset of empathy-labeled interactions (Section 2.2.2) and BART on the heuristic-based dataset created for warm-start (Section 2.2.3).

Additionally, we investigate the importance of different components of our model using the following ablated baselines:

- **Warm-start only, no RL training:** We analyze the performance of the model at the end of our warm-start stage, i.e. without any RL training.
- **No coherence reward:** We train the model without using the sentence coherence reward.



(a) PARTNER and MIME are effective at increasing empathy in zero-empathy responses. However, PARTNER is more effective in increasing empathy in low, non-zero empathic responses and doesn't make an already empathic post worse.

(b) PARTNER makes fewer number of changes compared to baselines. The changes are relatively more for less empathic responses which also tend to be shorter.

**Figure 2.6:** Analysis of empathic rewritings. Error bars are 95% confidence intervals.

Model	Change in empathy ( $\uparrow$ )	Perplexity ( $\downarrow$ )	Specificity ( $\uparrow$ )	Diversity ( $\uparrow$ )		Sentence coherence ( $\uparrow$ )	Edit rate ( $\downarrow$ )
				distinct-1	distinct-2		
<b>Partner</b>	<b>1.6410</b>	7.3641	0.9052	0.0659	0.3807	0.3030	<b>0.9654</b>
- no coherence	1.6127	7.2806	<b>0.9055</b>	0.0663	0.3844	0.3005	1.0108
- no mutual info.	1.6132	7.3274	0.9045	0.0674	0.3859	<b>0.3078</b>	1.0071
- warm-start only	0.7481	<b>7.1858</b>	0.9027	<b>0.0816</b>	<b>0.4238</b>	0.2935	1.0327

**Table 2.8:** Ablation results. Warm-start improves over DialoGPT but is still much worse than PARTNER in empathy improvement, highlighting the effectiveness of our RL-based training.

- **No mutual information:** We train the model without using the mutual information component.

## Automatic Metrics Results

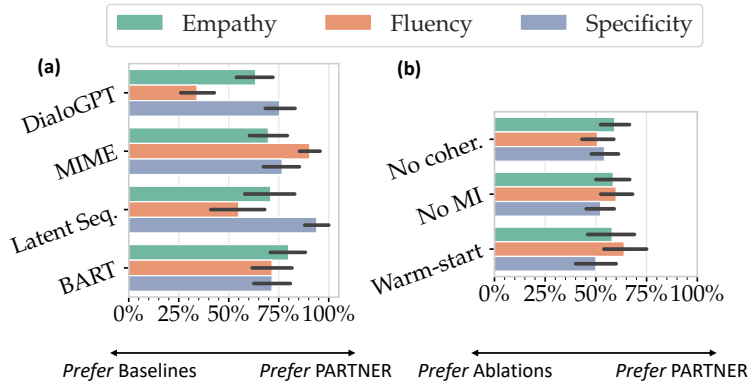
**Baseline Results.** Table 2.7 reports the results of PARTNER on the automatic evaluation metrics and comparisons with baselines. We find that empathic rewriting through PARTNER achieves the largest change in empathy (35% more than the next best approach, MIME) and is more specific than all baselines. MIME generates empathic outputs (+1.21 change in empathy) but the generations have low diversity (86% less than PARTNER) indicating similar responses for most seeker posts. BART generates outputs with lowest perplexity, highest diversity, and lowest edit rate, which is consistent with substantial improvements to language models in recent years (Brown et al., 2020). However, to our surprise, the rewritten responses through BART receive an overall

drop of 0.06 in empathy, indicating that the model is unable to perform the task of empathic rewriting well and only generates non-empathic, fluent, diverse text.

Our specificity metric can be hard to interpret with values having a really small range (0.85 to 0.9). However, with human-based evaluation (Section 2.2.4), we find that a difference of 0.05 on this metric (between PARTNER and latent seq.) translates to a 90% preference towards PARTNER. Moreover, while PARTNER has the lowest sentence coherence score, we find that this is likely due to higher number of sentences generated by it compared to baselines. The baselines generate 1-2 sentence responses on an average, where achieving high coherence between sentences is expected (e.g., a one-sentence response by design has a coherence of 1.0). PARTNER, on the contrary, generates responses with  $\sim 70\%$  more sentences than baselines, affecting the overall coherence score.

**Adaptability of rewritings to original post.** Adapting to different types of original responses and making appropriate changes is an important aspect of empathic rewriting. A low empathic response needs a lot more improvements and edits than a highly empathic response. Figure 2.6a shows the change in empathy of responses given their original empathy levels. We find that PARTNER performs better than baselines in improving responses with low empathy. Importantly, only PARTNER succeeds at not deteriorating responses that are already highly empathic, indicating the effectiveness of PARTNER at adapting to responses with different empathy levels. We also analyze the number of edits by each model on responses with different original empathy levels (Figure 2.6b). PARTNER not only effects a greater change in empathy than baselines, it achieves so with the least number of edits for both low and high empathy responses.

**Ablation results.** Table 2.8 reports results on ablated versions of PARTNER. Only using warm-start and no RL training is +0.2783 points better than the related off-the-shelf DialoGPT baseline on empathy improvement. However, the RL training in PARTNER further improves over this warm-start model by +0.8929 points. Using the coherence and mutual information rewards leads to small performance improvements, particularly in empathy (+0.03).



**Figure 2.7:** Human evaluation of empathy, fluency, and specificity in rewritings from PARTNER vs. (a) rewritings from baseline models, and (b) rewritings from ablations. PARTNER is preferred over baselines and ablations in empathy and specificity and is competitive in fluency.

	Model	BLEU score
	<b>Partner</b>	<b>0.1391</b>
Baselines	DialoGPT	0.0722
	MIME	0.0808
	Latent Seq	0.0254
	BART	0.0956
Ablations	- no coherence	0.1335
	- no mutual info.	0.1297
	- warm-start only	0.1281

**Table 2.9:** BLEU scores of all models against empathic rewritings from experts. The rewritings from PARTNER are closest to expert rewritings.

## Human Evaluation Results

Since automatic evaluation in language generation is often not robust (Liu et al., 2016), we perform a human evaluation on our key metrics (empathy, fluency, and specificity) through A/B testing. We recruit six graduate students in clinical psychology with expertise in empathy and mental health support<sup>11</sup> and ask them to compare outputs from PARTNER against other baseline models, ablations, and expert empathic rewritings (Section 2.2.2) given the same input. Presenting a seeker post, a rewritten response post from PARTNER, and a rewritten response post from

<sup>11</sup>Most participants were PhD students in second or subsequent years of their degree program. Research in Psychology has shown that clinical psychology graduate students are, in general, representative of mental health professionals (Öst et al., 2012). Although there are likely some differences between students and licensed psychologists, clinical outcomes in empathy-related measures such as therapeutic alliance have been shown to be comparable while students receive supervision (Goldstein et al., 2020).

Seeker post	Response post	Model	Rewritten response post
I feel like nobody cares about my existence.	I feel the same.	PARTNER	It's hard to find others who can relate. I feel the same.
		DialoGPT	I care.
		MIME	That is awful, i am sorry to hear that. I hope you feel better.
		Latent Seq.	I'm sorry to hear that.
		BART	I think i feel the same way. I do.
I overthink everything my friends say. I disregard them when they praise me but if they say something negative I take it so personally. I'm sick of being so paranoid.	I understand completely. I feel the same way.	PARTNER	It's hard when you're with such people and they are being like that to you. I understand completely. I feel the same way.
		DialoGPT	I'm the same way. I'm not sure why, but I just don't like being told I'm being paranoid.
		MIME	That is a shame, i am sure it will not be a bit of a bit!
		Latent Seq.	I know that, i feel the same way i'm sorry to.
		BART	I'm sorry, I understand completely. I feel the same way. I'm here for you.
I hope I didn't lose my only two best friends. They are the only ones I could relate to.	What happened between you two?	PARTNER	That's awful. What happened between you two? What caused you to break?
		DialoGPT	I hope you didn't lose them. I don't want to lose them.
		MIME	That is a good thing to do.
		Latent Seq.	I'm sorry to hear that.
		BART	What happened? What happened between you two?

**Table 2.10:** Qualitative examples of empathic rewriting using PARTNER and baseline methods. PARTNER generates rewritings that leverage both seeker post and original response post for empathic responses. It infers perspective from seeker posts (“*I feel like nobody cares about my existence*” → “*It’s hard to find others who can relate*”) and generates empathic sentences which can be coherently inserted to response posts (“*What happened between you two?*” → “*What happened between you two? What caused you to break?*”)

a baseline/ablation/expert-rewrite, we ask them to choose (a) response post which is more empathic, (b) response post which is more fluent, and (c) response post which is more specific. For each model, we collect evaluations on 50-100 examples.

**Results: Baselines and ablations.** Figure 2.7 shows the percentage of instances in which PARTNER was preferred over other baselines and ablations (values  $> 50\%$  indicate preference towards PARTNER). We find that rewritten responses from PARTNER are preferred for empathic and specific responses over all baselines. DialoGPT is judged more fluent (Figure 2.7a) but generates responses following similar templates (e.g., “*I’m sorry you.... I hope you....*”). Moreover, PARTNER has  $\sim 55\%$  preference for empathy over ablations where coherence and mutual information rewards are not used ( $p < 0.01$ ).

**Results: Expert rewritings.** The most appropriate way of performing empathic rewriting is through human experts. However, experts with training in therapy and mental health support are limited (Olfson, 2016) which makes it infeasible to employ them for millions of conversations on online support platforms. We use the small dataset of 180 empathic rewritings from experts to establish what the gold-standard performance for empathic rewritings in mental health support looks like. Unsurprisingly, experts are preferred  $\sim 80\text{-}90\%$  times over PARTNER in empathy, fluency, and specificity ( $p < 0.001$ ). However, in 10-20% cases PARTNER rewritings are preferred; these are typically instances where PARTNER is able to make empathic changes to responses while the experts leave it unchanged.

**Results: BLEU scores.** We also use the dataset of expert empathic rewritings (Section 2.2.2) as a ground truth and compare outputs of PARTNER, baselines, and ablations based on this ground truth using the BLEU metric (Papineni et al., 2002) (Table 2.9). We find that the outputs from PARTNER are closest to expert rewritings (86% better than the next best baseline, BART).

## Qualitative Examples

We present example rewritings from PARTNER and baselines in Table 2.10. PARTNER generates rewritings that leverage both seeker post and original response post for empathic responses. For example, from the seeker post “*I feel like nobody cares about my existence*”, PARTNER is able to infer “*It’s hard to find others who can relate*”. Also, PARTNER can coherently transform the response post

*“What happened between you two?” to “What happened between you two? What caused you to break?”.*

## 2.2.5 Discussion

Rewriting human-generated responses may be an effective approach to balancing the benefits and risks of using artificial intelligence in mental health settings. By combining human knowledge of context and experience, our approach can both provide feedback to online peer-supporters with actionable, real-time examples, and provide support seekers with more empathic responses. Importantly, this machine-in-the-loop approach can help mitigate some of the risks related to toxicity and safety of AI systems in settings of suicidal ideation, self-harm, or insensitive comments related to race/ethnicity/gender (Li et al., 2020; Luxton et al., 2012; Collings and Niederkrotenthaler, 2012a).

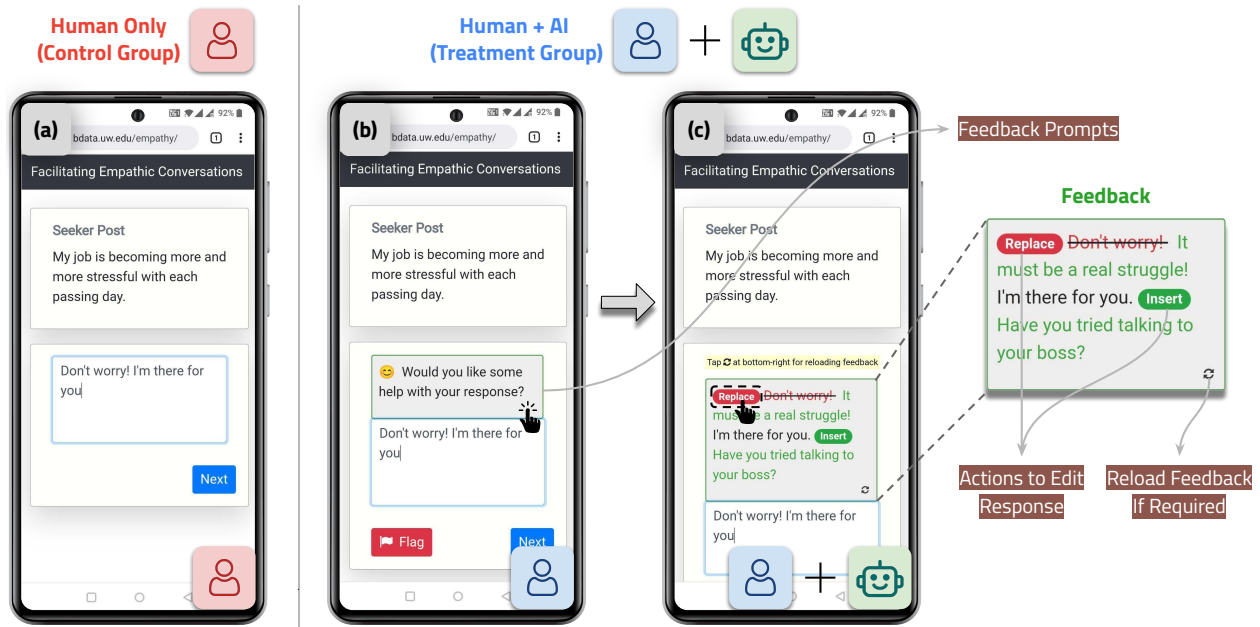
## 2.3 Helping Untrained Peer Supporters Express Empathy More Effectively

We next use PARTNER to provide just-in-feedback AI-based feedback to peer supporters.

### 2.3.1 Hailey: Human-AI Collaboration Approach for Empathy

We develop and evaluate a Human-AI collaboration approach for helping untrained peer supporters write more empathic responses in online, text-based peer-to-peer support. We propose HAILEY (Human-AI coLLaboration approach for EmpathY), an AI-in-the-loop agent that offers just-in-time suggestions to express empathy more effectively in conversations (Figure 2.8b, 2.8c). We design HAILEY to be collaborative, actionable, and mobile friendly.

Unlike the AI-only task of empathic dialogue generation (generating empathic responses from scratch) (Lin et al., 2019; Majumder et al., 2020; Rashkin et al., 2019), HAILEY adopts a collaborative design that edits existing human responses to make them more empathic. This design reflects the high-stakes setting of mental health, where AI is likely best used to augment, rather than replace, human skills (Miner et al., 2019; Chen and Asch, 2017). Furthermore, while current AI-in-the-loop systems are often restricted in the extent to which they can guide humans (e.g.,



**Figure 2.8:** We performed a randomized controlled trial with 300 TalkLife peer supporters as participants. We randomly divided participants into Human Only (control) and Human + AI (treatment) groups and asked them to write supportive, empathic responses to seeker posts without feedback and with feedback, respectively. To identify whether just-in-time Human-AI collaboration helped increase expressed empathy beyond potential (but rare) traditional training methods, participants in both groups received initial empathy training before starting the study (Methods; Supplementary Figure A.2). **(a)** Without AI, human peer supporters are presented with an empty chatbox to author their response (the current status quo). As peer supporters are typically untrained on best-practices in therapy – such as empathy – they rarely conduct highly empathic conversations. **(b)** Our feedback agent (HAILEY) prompts peer supporters for providing just-in-time AI feedback as they write their responses. **(c)** HAILEY then suggests changes that can be made to the response to make it more empathic. These suggestions include new sentences that can be *inserted* and options for *replacing* current sentences with their more empathic counterparts. Participants can accept these suggestions by clicking on the *Insert* and *Replace* buttons and continue editing the response or get more feedback, if needed.

simple classification methods that tell users to be empathic when they are not) (Tanana et al., 2019; Peng et al., 2020; Hui et al., 2018; Kelly et al., 2018), we ensure actionability by guiding peer supporters with concrete steps they may take to respond with more empathy. HAILEY is designed to suggest the *insertion* of new empathic sentences or *replacement* of existing low-empathy sentences with their more empathic counterparts (Figure 2.8c). For complex, hard-to-learn skills like empathy, this enables just-in-time suggestions on not just “what” to improve but on “how” to improve it.

We consider the general setting of text-based, asynchronous conversations between a support seeker and a peer supporter (Figure 2.8). In these conversations, the support seeker authors a post seeking mental health support (e.g., “*My job is becoming more and more stressful with each passing day.*”) to which the peer supporter writes a supportive response (e.g., “*Don’t worry! I’m there for you.*”). In this context, we support the peer supporters by providing just-in-time AI feedback to improve the empathy of their responses. To do so, HAILEY prompts the peer supporter through a pop-up (“*Would you like some help with your response?*”) placed above the response text box. On clicking this prompt, HAILEY shows just-in-time AI feedback consisting of *Insert* (e.g., Insert “*Have you tried talking to your boss?*” at the end of the response) and *Replace* (e.g., Replace “*Don’t worry!*” with “*It must be a real struggle!*”) suggestions based on the original seeker post and the current peer supporter response. The peer supporter can incorporate these suggestions by directly clicking on the appropriate Insert or Replace buttons, by further editing them, and/or by deriving ideas from the suggestions to indirectly use in their response. These suggestions are generated using PARTNER (Chapter 2.2).

## Design Goals

HAILEY is designed (1) with a collaborative “AI-in-the-loop” approach, (2) to provide actionable feedback, and (3) to be mobile friendly.

**Collaborative AI-in-the-loop Design.** In the high-stakes setting of mental health support, AI is best used to augment, rather than replace, human skill and knowledge (Miner et al., 2019; Chen and Asch, 2017). Current natural language processing technology – including language models, conversational AI methods, and chatbots – continue to pose risks related to toxicity, safety, and

bias, which can be life-threatening in contexts of suicidal ideation and self-harm (Wolf et al., 2017; Bolukbasi et al., 2016; Daws, 2020; Richardson et al., 2021). To mitigate these risks, researchers have called for Human-AI collaboration methods, where primary communication remains between two humans with an AI system “in-the-loop” to assist humans in improving their conversation (Miner et al., 2019; Chen and Asch, 2017). In HAILEY, humans remain at the center of the interaction, receive suggestions from our AI “in-the-loop,” and retain full control over which suggestions to use in their responses (e.g., by selectively choosing the most appropriate Insert or Replace suggestions and editing them as needed).

**Actionable Feedback.** Current AI-in-the-loop systems are often limited to addressing “what” (rather than “how”) participants should improve (Tanana et al., 2019; Peng et al., 2020; Hui et al., 2018; Kelly et al., 2018). For such a goal, it is generally acceptable to design simple interfaces that prompt participants to leverage strategies for successful supportive conversations (e.g., prompting “*you may want to empathize with the user*”) without any instructions on how to concretely apply those strategies. However, for complex, hard-to-learn constructs such as empathy (Davis et al., 1980; Elliott et al., 2011), there is a need to address the more actionable goal of steps to take for participants to improve. HAILEY, designed to be actionable, suggests concrete actions (e.g., sentences to insert or replace) that participants may take to make their current response more empathic.

**Mobile Friendly Design.** Online conversations and communication are increasingly mobile based. This is also true for peer-to-peer support platforms, which generally provide their services through a smartphone application. Therefore, a mobile friendliness design is critical for the adoption of conversational assistive agents like ours. However, the challenge here relates to the complex nature of the feedback and the smaller, lower-resolution screen on a mobile device as compared to a desktop. We therefore designed a compact, minimal interface that works equally well on desktop and mobile platforms. We created a conversational experience based on the mobile interface of peer-to-peer support platforms that was design minimal, used responsive prompts that adjusted in form based on screen sizes, placed AI feedback compactly above the response text box for easy access, and provided action buttons that were easy for mobile users to click on.

## Feedback Workflow

Through HAILEY, we showed prompts to participants that they could click on to receive feedback. Our feedback, driven by a previously validated Empathic Rewriting model, consists of actions that users can take to improve the empathy of their responses (Supplementary Figure A.31).

**Prompts to Trigger Feedback.** We showed the prompt “*Would you like some help with your response?*” to participants, which was placed above the response text box (Figure 2.8b). Participants could at any point click on the prompt to receive feedback on their current response (including when it is still empty). When this prompt is clicked, HAILEY acts on the seeker post and the current response to suggest changes that will make the response more empathic. Our suggestions consisted of Insert and Replace operations generated through empathic rewriting of the response.

**Generating Feedback through Empathic Rewriting.** The goal of empathic rewriting (Section 2.2) is to transform low empathy text to higher empathy. We proposed PARTNER, a deep reinforcement learning model that learns to take sentence-level edits as actions in order to increase the expressed level of empathy while maintaining conversational quality. PARTNER’s learning policy is based on a transformer language model (adapted from GPT-2 (Radford et al.)), which performs the dual task of generating candidate empathic sentences and adding those sentences at appropriate positions. PARTNER-generated rewritings increase empathy by 1.6 (on the 6-point empathy scale), which is >35% more than all state-of-the-art baseline methods and are judged more empathic over 65% of the time than baselines by human annotators. Here, we build on PARTNER by further improving training data quality through additional filtering, supporting multiple generations for the real-world use-case of multiple types of feedback for the same post, and evaluating a broader range of hyperparameter choices.

**Showing Feedback as Actions.** We map the rewritings generated by our optimized version of PARTNER to suggestions to *Insert* and *Replace* sentences. These suggestions are then shown as actions to edit the response. To incorporate the suggested changes, the participant clicks on the respective Insert or Replace buttons. Continuing our example from Figure 2.8, given the seeker post “*My job is becoming more and more stressful with each passing day.*” and the original response “*Don’t worry! I’m there for you.*”, PARTNER takes two insert actions – Replace “*Don’t worry!*” with

*“It must be a real struggle!”* and Insert *“Have you tried talking to your boss?”* at the end of the response. These actions are shown as feedback to the participant. See Supplementary Figure A.9 for more qualitative examples.

**Reload Feedback If Required.** For the same combination of seeker post and original response, multiple feedback suggestions are possible. In the Figure 1 example, instead of suggesting the insert *“Have you tried talking to your boss?”*, we could also propose inserting *“I know how difficult things can be at work”*. These feedback variations can be sampled from our model and, if the initial sampled feedback does not meet participant needs, iterated upon to help participants find better-suited feedback. HAILEY provides an option to *reload* feedback, allowing participants to navigate through different feedback and suggestions if necessary.

### 2.3.2 Study Design

We employed a between-subjects study design in which each participant was randomly assigned to one of Human + AI (treatment; N=139) or Human Only (control; N=161) conditions. Participants in both groups were asked to write supportive, empathic responses to a unique set of 10 existing seeker posts (one at a time), sourced at random from a subset of TalkLife posts. The Human + AI (treatment) group participants were given the option of receiving feedback through prompts as they typed their responses. Participants in the Human Only (control) group, in contrast, wrote responses with no option for feedback.

**Power Analysis.** We used a power analysis to estimate the number of participants required for our study. For an effect size of 0.1 difference in empathy, a power analysis with a significance level of 0.05, powered at 80%, indicated that we required 1,500 samples of (seeker post, response post) pairs each for treatment and control groups. To meet the required sample size, we collected 10 samples per participant and therefore recruited from 300 participants in total (with the goal of 150 participants per condition), for a total of 1,500 samples each.

**Dataset of Seeker Posts.** We obtained a unique set of 1500 seeker posts, sampled at random with consent from the TalkLife platform, in the observation period from May 2012 to June 2020. Prior to sampling, we filtered posts related to (1) critical settings of suicidal ideation and self-harm, using pre-defined unsafe regular expressions (e.g., *“.\*(commit suicide).\*”*, *“.\*(cut).\*”*), to

ensure participant safety (Discussion), and (2) common social media interactions not related to mental health (e.g., “Happy mother’s day”) using a standard BERT-based text classifier (Devlin et al., 2019), trained on a manually annotated dataset of ~3k posts with answers to the question “Is the seeker talking about a mental health related issue or situation in his/her post?” (~ 85% accuracy). We randomly divided these 1500 posts into 150 subsets of 10 posts each. We used the same 150 subsets for both treatment and control conditions for consistent context for both groups of participants.

**Participant Demographics.** In our study, 54.3% of the participants identified as female, 36.7% as male, 7.3% as non-binary, and the remaining 1.7% preferred not to report their gender. The average age of participants was 26.3 years (std = 9.5). 45.7% of the participants identified as White, 20.3% as Asians, 10.7% as Hispanic or Latino, 10.3% as Black or African American, 0.7% as Pacific Islander or Hawaiian, 0.3% as American Indian or Alaska Native, and the remaining 12.0% preferred not to report their race/ethnicity. 62.3% of the participants were from the United States, 13.7% were from India, 2.3% were from United Kingdom, 2.3% were from Germany, and the remaining 19.3% were from 36 different countries (spanning six of seven continents excluding Antarctica). Moreover, 31.7% of the participants reported having no experience with peer-to-peer support despite having been recruited from the TalkLife platform, 26.3% as having less than one year of experience, and 42.0% reported having greater than or equal to one year of experience with peer-to-peer support.

**RCT Group Assignment.** On clicking the advertised pop-up used for recruitment, a TalkLife user was randomly assigned to one of the Human + AI (treatment) or Human Only (control) conditions for the study duration.

**Study Workflow.** We divided our study into four phases:

- **Phase I: Pre-Intervention Survey.** First, both control and treatment group participants were asked the same set of survey questions describing their demographics, background and experience with peer-to-peer support (Supplementary Figure A.20, A.21).
- **Phase II: Empathy Training and Instructions.** Next, to address whether participants held similar understandings of empathy, both groups received the same initial empathy train-

ing, which included empathy definitions, frameworks, and examples based on psychology theory, before starting the main study procedure of writing empathic responses (Supplementary Figure A.2). Participants were also shown instructions on using our study interface in this phase (Supplementary Figure A.22, A.23, A.24, A.25, A.26, A.27, A.28, A.29).

- **Phase III: Write Supportive, Empathic Responses.** Participants then started the main study procedure and wrote responses to one of the 150 subsets of 10 existing seeker posts (one post at a time). For each post, participants in both the groups were prompted “*Write a supportive, empathic response here*”. The Human + AI (treatment) group participants were given the option of receiving feedback through prompts as they typed their responses (Supplementary Figure A.31). Participants in the Human Only (control) group wrote responses without any option for feedback (Supplementary Figure A.30).
- **Phase IV: Post-Intervention Survey.** After completing the 10 posts, participants in both groups were asked to assess the study by answering questions about the difficulty they faced while writing responses, the helpfulness and actionability of the feedback, their self-efficacy after the study, and the intent to adopt the system (Supplementary Figure A.33, A.34, A.35).

If participants dropped out of the study before completing it, their data was removed from our analyses. Participants took 20.6 minutes on average to complete the study. US citizens and permanent US residents were compensated with a 5 USD Amazon gift card. Furthermore, the top-2 participants in the human evaluation (Evaluation) received an additional 25 USD Amazon gift card. Based on local regulations, we were unable to pay non-US participants. This was explicitly highlighted in the participant consent form on the first landing page of our study (Supplementary Figure A.19, A.36).

**Participant Recruitment.** We worked with TalkLife to recruit participants directly from their platform. Because users on such platforms are typically untrained in best-practices of providing mental health support, their work offers a natural place to deploy feedback systems like ours. To recruit participants, we advertised our study on TalkLife. Recruitment started in April 2021 and continued until September 2021. The study was approved by the University of Washington’s Institutional Review Board (determined to be exempt; IRB ID STUDY00012706).

### 2.3.3 Results

#### Increase In Expressed Empathy Due To Human-AI Collaboration

Our primary finding is that providing just-in-time AI feedback to participants leads to more empathic responses (Figure 2.9). Specifically, through human evaluation from an independent set of TalkLife users (Methods), we found that the Human + AI responses were rated as being more empathic than the Human Only responses 46.8% of the time and were rated equivalent in empathy to Human Only responses 15.7% of the time. On the other hand, Human Only responses were preferred only 37.4% of the time ( $p < 0.01$ ;  $t = 5.88$ ;  $D_f = 2998$ ; Two-sided Student's t-test; Figure 2.9a). In addition, by automatically estimating empathy levels of responses using a previously validated empathy classification model on a scale from 0 to 6 (Methods), we found that the Human + AI approach led to 19.6% higher empathic responses compared to the Human Only approach (1.77 vs. 1.48; Cohen's  $d = 0.24$ ;  $p < 10^{-5}$ ;  $t = 5.46$ ;  $D_f = 2998$ ; Two-sided Student's t-test; Figure 2.9b).

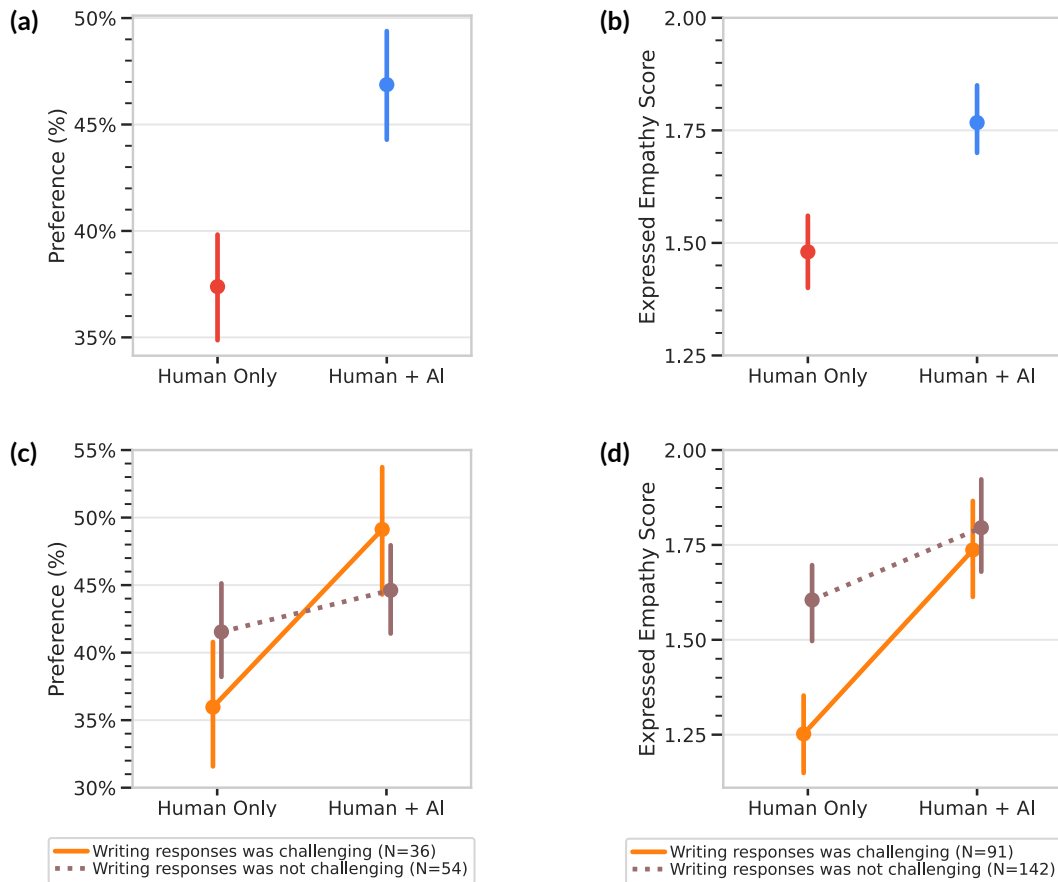
#### Higher Gains For Those Who Report Peer-support Challenges

Prior work has shown that online peer supporters find it extremely challenging to write supportive and empathic responses (Kemp and Henderson, 2012; Mahlke et al., 2014). Some participants have little to no prior experience with peer support (e.g., if they are new to the platform;  $N=95/300$ ; Methods). Even as the participants gain more experience, in the absence of explicit training or feedback, the challenge of writing supportive responses persists over time and may even lead to a gradual decrease in empathy levels due to factors such as empathy fatigue (Schwalbe et al., 2014; Goldberg et al., 2016; Stebnicki, 2007; Nunes et al., 2011; Hojat et al., 2009), as also observed during the course of our 30-minute study (Supplementary Figure A.7). Therefore, it is particularly important to better assist the many participants who struggle with writing responses.

For the subsample of participants who self-reported challenges in writing responses at the end of our study ( $N=91/300$ ; Methods), a post hoc analysis revealed significantly higher empathy gains using the Human-AI collaboration approach. For such participants, we found an absolute 4.5% stronger preference for the Human + AI responses (49.1% vs. 44.6%;  $p < 0.01$ ;  $t = 4.05$ ;  $D_f =$

**Human Evaluation:**  
Which response is more empathetic?

**Automatic/AI-based Evaluation:**  
Expressed empathy score



**Figure 2.9:** (a) Human evaluation from an independent set of TalkLife users showed that the Human + AI responses (N=139) were strictly preferred 46.9% of the time relative to a 37.4% strict preference for the Human Only responses (N=161). (b) Through automatic evaluation using an AI-based expressed empathy score (Chapter 2.1), we found that the Human + AI responses (N=139) had 19.6% higher empathy than the Human Only responses (N=161; 1.77 vs. 1.48; Cohen’s  $d=0.24$ ;  $p=5.1 \times 10^{-8}$ ;  $t = 5.46$ ;  $D_f = 2998$ ; Two-sided Student’s t-test). (c) For the participants who reported challenges in writing responses after the study, we found a stronger preference for the Human + AI responses vs. Human Only responses (49.1% vs. 34.0%), compared to participants who did not report challenges (44.6% vs. 41.5%). (d) For participants who reported challenges in writing responses after the study, we found a higher improvement in expressed empathy scores of the Human + AI responses vs. Human Only responses (38.9%; 1.74 vs. 1.25; Cohen’s  $d=0.43$ ), compared to participants who did not report challenges (11.9%; 1.79 vs. 1.60; Cohen’s  $d=0.15$ ). In c and d, the sample size varied to ensure comparable conditions (Methods). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

718; Two-sided Student's t-test; Figure 2.9c) and a 27.0% higher increase in expressed empathy using the Human + AI approach (38.8% vs. 11.8%;  $p < 10^{-5}$ ;  $t = 5.90$ ;  $D_f = 1818$ ; Two-sided Student's t-test; Figure 2.9d) compared to participants who did not report any challenges. For the subsample of participants who self-reported no previous experience with online peer support at the start of our study ( $N=95/300$ ; 37 of these participants also self-reported challenges), we found a 8.1% stronger preference for the Human + AI responses (51.8% vs. 43.7%;  $p < 0.01$ ;  $t = 4.42$ ;  $D_f = 758$ ; Two-sided Student's t-test;) and a 21.2% higher increase in expressed empathy using the Human + AI approach (33.7% vs. 12.5%;  $p < 10^{-5}$ ;  $t = 4.58$ ;  $D_f = 1898$ ; Two-sided Student's t-test; Supplementary Figure A.12d) compared to participants who reported experience with online peer support.

## Key Human-AI Collaboration Patterns

The collaboration between humans and AI can take many forms but specific formulations of human-AI collaboration remain poorly defined and challenging to measure (Bansal et al., 2021). Investigating how humans collaborate with our AI can help us better understand the system's use-cases and inform better design decisions. Here, we analyzed collaboration patterns of participants both over the course of the study as well as during a single response instance. We leveraged this analysis to derive a hierarchical taxonomy of Human-AI collaboration patterns based on how often the AI was consulted during the study and how AI suggestions were used (Figure 2.10a; Methods).

Our analysis revealed several categories of collaboration. For example, some participants chose to always rely on the AI feedback, whereas others only utilized it as a source of inspiration and rewrote it in their own style. Based on the number of posts in the study for which AI was consulted (out of the 10 posts for each participant), we found that participants consulted AI either always (15.5%), often (56.0%), once (6.0%), or never (22.4%). Very few participants always consulted and used the AI (2.6%), indicating that they did not rely excessively on AI feedback. A substantial number of participants also chose to never consult the AI (22.4%). Such participants, however, also expressed the least empathy in their responses (1.13 on average out of 6; Figure 2.10b), suggesting that consulting the AI could have been beneficial.

Furthermore, based on how AI suggestions were used, we found that participants used the suggestions either directly (64.6%), indirectly (18.5%), or not at all (16.9%). As expected given our system's design, the most common way of usage was direct, which entailed clicking on the suggested actions to incorporate them in the response. In contrast, participants who indirectly used AI (Methods) drew ideas from the suggested feedback and rewrote it in their own words in the final response. Some participants, however, chose not to use suggestions at all (16.9%); a review of these instances by the researchers, as well as the subjective feedback from participants, suggested that reasons included the feedback not being helpful, the feedback not being personalized, or their response already being empathic and leaving little room for improvement. Finally, multiple types of feedback are possible for the same combination of seeker post and original response, and some participants (16.9%) used our reload functionality (Methods) to read through these multiple suggestions before they selected a final response.

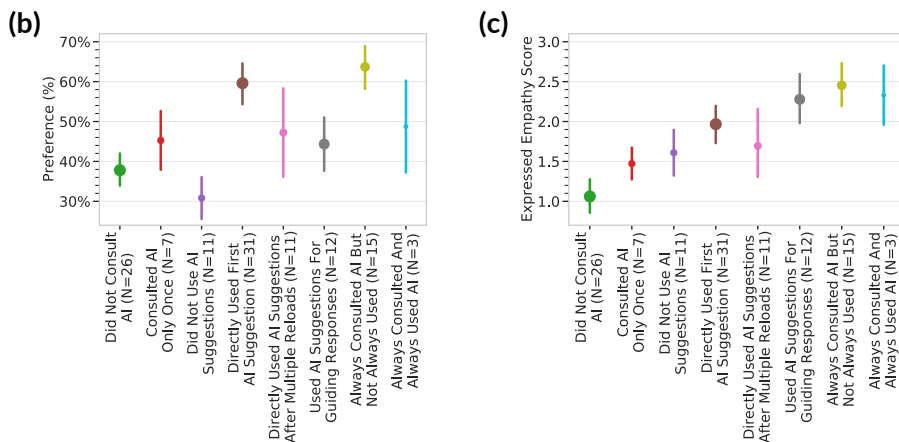
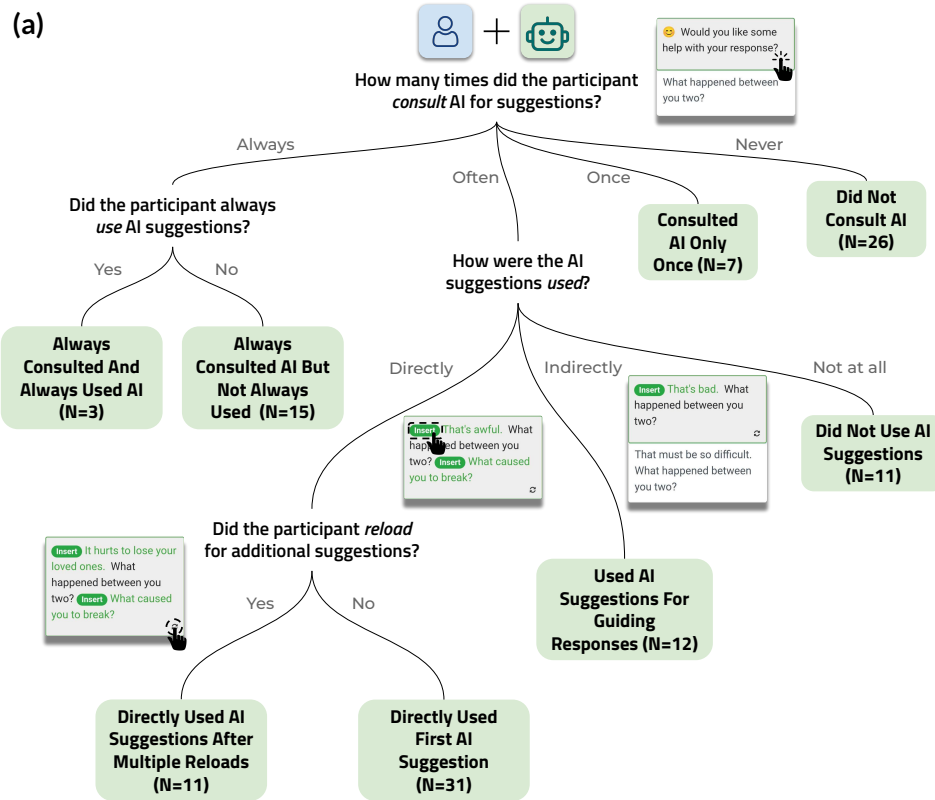
In general, participants who consulted and used AI more often expressed higher empathy, though this pattern was more pronounced when evaluated through our automatic expressed empathy score (Figure 2.10c) than through human evaluation (Figure 2.10b).

## **Positive Perceptions Of Participants**

At the end of our study, we collected study participants' perceptions about the usefulness and actionability of the feedback and their intention to adopt the system. We observed that 63.3% of participants found the feedback they received helpful, 60.4% found it actionable, and 77.7% of participants wanted this type of feedback system to be deployed on TalkLife or other similar peer-to-peer support platforms (Supplementary Figure A.4), indicating the overall effectiveness of our approach. We also found that 69.8% of participants self-reported feeling more confident at providing support after our study; this indicates the potential value of our system for training and increased self-efficacy (Supplementary Figure A.4).

### **2.3.4 Discussion**

Our work demonstrates how humans and AI might collaborate on open-ended, social, and high-stakes tasks such as conducting empathic conversations. Empathy is complex and nuanced



**Figure 2.10:** We derived a hierarchical taxonomy of Human-AI collaboration categories. **(a)** We clustered the interaction patterns of Human + AI (treatment) participants based on how often the AI was consulted during the study and how the AI suggestions were used (N=116/139). We excluded participants who belonged to multiple clusters (N=23/139). Very few participants always consulted and used AI (2.6%), indicating that participants did not rely excessively on AI feedback. Participants could use AI feedback directly through suggested actions (64.6%) or indirectly by drawing ideas from the suggested feedback and rewriting it in their own words in the final response (18.5%). **(b)** Empathy increased when participants consulted and used AI more frequently, with those who did not consult AI (22.4%) or did not use AI (9.5%) having significantly lower preference over Human Only responses (N=37;  $p = 6.4 \times e^{-6}$ ; Two-sided Student's t-test). **(c)** Participants who did not consult AI had the lowest empathy levels based on our automatic evaluation (1.13 on average out of 6). The area of the points is proportional to the number of participants in the respective human-AI collaboration categories. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

(Davis et al., 1980; Riess, 2017; Blease et al., 2020; Doraiswamy et al., 2020) and is thus more challenging for AI than many other Human-AI collaboration tasks, such as scheduling meetings, therapy appointments and checking grammar in text. We show how the joint effects of humans and AI can be leveraged to help peer supporters, especially those who have difficulty providing support, converse more empathically with those seeking mental health support.

Our study has implications for addressing barriers to mental health care, where existing resources and interventions are insufficient to meet the current and emerging need. According to a WHO report, over 400 million people globally suffer from a mental health disorder, with approximately 300 million suffering from depression (WHO). Overall, mental illness and related behavioral health problems contribute 13% to the global burden of disease, more than both cardiovascular diseases and cancer (Collins et al., 2011). Although psychotherapy and social support (Kaplan et al., 1977) can be effective treatments, many vulnerable individuals have limited access to therapy and counseling (Kazdin and Blase, 2011; Olfson, 2016). For example, most countries have less than one psychiatrist per 100,000 individuals, indicating widespread shortages of workforce and inadequate in-person treatment options (Rathod et al., 2017).

A scalable approach to improving access to mental health support globally is by connecting support seekers and peer supporters using online platforms like TalkLife ([talklife.com](https://talklife.com)), YourDost ([yourdost.com](https://yourdost.com)) or Mental Health Subreddits ([reddit.com](https://reddit.com)) (Naslund et al., 2016) to those with mental health issues. However, a key challenge in doing so lies in enabling effective and high-quality conversations between untrained peer supporters and those in need at scale. We show that Human-AI collaboration can considerably increase empathy in peer supporter responses, a core component of effective and quality support that ensures improved feelings of understanding and acceptance (Bohart et al., 2002; Elliott et al., 2011; Watson et al., 2002). While fully replacing humans with AI for empathic care has previously drawn skepticism from psychotherapists (Blease et al., 2020; Doraiswamy et al., 2020), our results suggest that it is feasible to empower untrained peer supporters with appropriate AI-assisted technologies in relatively lower-risk settings, such as peer-to-peer support (Lee et al., 2021; Vaidyam et al., 2021; Miner et al., 2019; Imel et al., 2015; Kazdin and Blase, 2011).

Our findings also point to potential secondary gain for peer supporters in terms of (1) increased self-efficacy, as indicated by 69.8% of participants feeling more confident in providing

support after the study, and (2) gained experience and expertise by multiple example learning when using reload functionality to scroll through multiple types of responses for the same seeker post. This has implications for helping untrained peer supporters beyond providing them just-in-time feedback. One criticism of AI is that it may steal or dampen opportunities for training more clinicians and workforce (Blease et al., 2020; Doraiswamy et al., 2020). We show that Human-AI collaboration can actually enhance, rather than diminish, these training opportunities. This is also reflected in the subjective feedback from participants (Methods), with several participants reporting different types of learning after interacting with the AI (e.g., one participant wrote, *“I realized that sometimes I directly jump on to suggestions rather than being empathic first. I will have to work on it.”*, while another wrote, *“Feedback in general is helpful. It promotes improvement and growth.”*).

Further, we find that participants not only directly accept suggestions but also draw higher-level inspiration from the suggested feedback (e.g., a participant wrote *“Sometimes it just gave me direction on what [should] be said and I was able to word it my way. Other times it helped add things that made it sound more empathic...”*), akin to having access to a therapist’s internal brainstorming, which participants can use to rewrite responses in their own style.

In our study, many participants (N=91) reported challenges in writing responses (e.g., several participants reported not knowing what to say: *“I sometimes have a hard time knowing what to say.”*), which is characteristic of the average user on online peer-to-peer support platforms (Kemp and Henderson, 2012; Mahlke et al., 2014). We demonstrate a significantly larger improvement in empathy for these users, suggesting that we can provide significant assistance in writing more empathic responses, thereby improving empathy awareness and expression of the typical platform user. Through qualitative analysis of such participants’ subjective feedback on our study, we find that HAILEY can guide someone who is unsure about what to say (e.g., a participant wrote, *“Feedback gave me a point to start my own response when I didn’t know how to start.”*) and can help them frame better responses (e.g., one participant wrote, *“Sometimes I didn’t really knew [sic] how to form the sentence but the feedback helped me out with how I should incorporate the words.”*, while another wrote, *“Sometimes I do not sound how I want to and so this feedback has helped me on sounding more friendly...”*; Methods). One concern is that AI, if used in practice, may cause harm to multiple stakeholders from support seekers to care providers and peer supporters, through inappropriate

interventions, role confusion, and data sharing concerns (Richardson et al., 2021; Blease et al., 2020; Doraiswamy et al., 2020). According to our findings, however, the individuals struggling to do a good job are the ones who benefit the most, which forms an important use case of AI in healthcare.

The reported differences in empathy between treatment and control groups conservatively estimates the impact of our AI-in-the-loop feedback system due to (1) additional initial empathy training to both Human + AI and Human Only groups (Supplementary Figure A.2), and (2) a potential selection effect that may have attracted TalkLife users who care more about supporting others (Supplementary Figure A.8). In practice, training of peer supporters is very rare, and the effect of training typically diminishes over time (Nunes et al., 2011; Hojat et al., 2009). We included this training to understand whether just-in-time AI feedback is helpful beyond traditional training methods. Moreover, the Human Only responses in our study had 34.5% higher expressed empathy than existing Human Only responses to the corresponding seeker posts on the TalkLife platform (1.11 vs. 1.48;  $p \ll 10^{-5}$ ; Two-sided Student's t-test; Supplementary Figure A.8), reflecting the effects of additional training as well as a potential selection effect. We show here that Human-AI collaboration improves empathy expression even for participants who already express empathy more often; practical gains for the average user of the TalkLife platform could be even higher than the intentionally conservative estimates presented here.

## **Safety, Privacy, And Ethics**

Developing computational methods for intervention in high-stakes settings such as mental health care involves ethical considerations related to safety, privacy, and bias (Collings and Niederkroenthaler, 2012b; Li et al., 2020; Luxton et al., 2012; Martinez-Martin et al., 2018). There is a risk that in attempting to help, AI may have the opposite effect on the potentially vulnerable support seeker or peer supporter (Richardson et al., 2021). The present study included several measures to reduce such risks and unintended consequences. First, our collaborative, AI-in-the-loop writing approach ensured that the primary conversation remains between two humans, with AI offering feedback only when it appears useful, and allowing the human supporter to accept or reject it. Providing such human agency is safer than relying solely on AI, especially in a high-stakes men-

tal health context (Miner et al., 2019). Moreover, using only AI results in loss of authenticity in responses; hence, our Human-AI collaboration approach leads to responses with high empathy as well as high authenticity (Supplementary Figure A.3).

Second, our approach intentionally assists only peer supporters, not support seekers in crisis, since they are likely to be at a lower risk and more receptive to the feedback. Third, we filtered posts related to suicidal ideation and self-harm by using pre-defined unsafe regular expressions (e.g., “.\*(commit suicide).\*”, “.\*(cut).\*”). Such posts did not enter our feedback pipeline, but instead we recommended escalating them to therapists. We applied the same filtering to every generated feedback, as well, to try and ensure that HAILEY did not suggest unsafe text as responses. Fourth, such automated filtering may not be perfect; therefore, we included a mechanism to flag inappropriate/unsafe posts and feedback by providing our participants with an explicit “Flag Button” (Supplementary Figure A.32). In our study, 1.6% posts (out of 1390 in the treatment group) and 2.9% feedback instances (out of 1939 requests) were flagged as inappropriate or unsafe. While the majority of them were concerned with unclear seeker posts or irrelevant feedback, we found six cases (0.2%) that warranted further attention. One of these cases involved the post containing intentionally misspelled self-harm content (e.g., “c u t” with spaces between letters in order to circumvent safety filters); another related to feedback containing a self-harm related term; three addressed the post or feedback containing a swear word that may not directly be a safety concern (e.g., “You are so f\*\*king adorable”); and one contained toxic/offensive feedback (“It’s a bad face”).

Future iterations of our system could address these issues by leveraging more robust filtering methods and toxicity/hate speech classifiers (e.g., Perspective API ([perspectiveapi.com](https://perspectiveapi.com))). Several platforms, including TalkLife, already have systems in place to prevent triggering content from being shown, which can be integrated into our system on deployment. Finally, we removed all personally identifiable information (user and platform identifiers) from the TalkLife dataset prior to training the AI model.

## Limitations

While our study results reveal the promise of Human-AI collaboration in open-ended and even high-stakes settings, the study is not without limitations. Some of our participants indicated that empathy may not always be the most helpful way to respond (e.g., when support seekers are looking for concrete actions). However, as demonstrated repeatedly in the clinical psychology literature (Elliott et al., 2011; Bohart et al., 2002; Watson et al., 2002), empathy is a critical, foundational approach to all evidence-based mental health support, plays an important role in building alliance and relationship between people, and is highly correlated with symptom improvement. It has consistently proven to be an important aspect of responding, but support seekers may sometimes benefit from additional responses involving different interventions (e.g., concrete problem solving, motivational interviewing (Tanana et al., 2016)). Future work should investigate when such additional responses are helpful or necessary.

Some participants may have been apprehensive about using our system, as indicated by the fact that many participants did not consult or use it (N=37). Qualitatively analyzing the subjective feedback from these participants suggested that this might be due to feedback communicating incorrect assumptions about the preferences, experience, and background of participants (e.g., assuming that a participant is dealing with the same issues as the support seeker: *“Not sure this can be avoided, but the feedback would consistently assume I’ve been through the same thing.”*). Future work should personalize prompts and feedback to individual participants. This could include personalizing the content and the frequency of the prompt as well as personalizing the type of feedback that is shown from multiple possible feedback options.

Our assessment includes validated yet automated and imperfect measures. Specifically, our evaluation of empathy is based only on empathy that was *expressed* in responses, not empathy that might have been *perceived* by the support seeker (Barrett-Lennard, 1981). In sensitive contexts like ours, however, obtaining perceived empathy ratings from support seekers is challenging and involves ethical risks (Safety, Privacy, and Ethics). We attempted to reduce the gap between expressed and perceived empathy in our human evaluation by recruiting participants from TalkLife who may be seeking support on the platform (Methods). Nevertheless, studying the effects of Human-AI collaboration on perceived empathy in conversations is a vital future

research direction. However, note that psychotherapy research indicates a strong correlation between expressed empathy and positive therapeutic outcomes and commonly uses it as a credible alternative (Elliott et al., 2011; Bohart et al., 2002; Watson et al., 2002).

Furthermore, we acknowledge that a variety of social and cultural factors might affect the dynamics of the support and the expression of empathy (De Choudhury et al., 2017; Cauce et al., 2002; Satcher, 2001). As such, our Human-AI collaboration approach must be adapted and evaluated in various socio-cultural contexts, including underrepresented communities and minorities. While conducting randomized controlled trials on specific communities and investigating heterogeneous treatment effects across demographic groups is beyond the scope of our work, our study was deployed globally and included participants of various gender identities, ethnicities, ages, and countries (Methods; Supplementary Figure A.11, A.12). However, this is a critical area of research, and ensuring equitable access to culturally sensitive empathic support requires further investigation.

Our study evaluated a single Human-AI collaboration interface design, and there could have been other potential interface designs, as well. Additionally, as a secondary exploration, we analyzed a classification-based interface design, which provided participants with the option to request automatic expressed empathy scores for their responses, using the empathy classifier developed in Section 2.1 (Supplementary Figure A.6). We assigned this secondary classification-based AI treatment to 10% of the incoming participants at random (N=30). Due to conflicting human and automatic evaluation results, we observed that the effects of this secondary treatment on empathy of participants were ambiguous (Supplementary Figure A.5a, A.5b); however, the design was perceived as being less actionable than our primary rewriting-based interface (Supplementary Figure A.5c). This poses questions on what types of design are optimal and how best to provide feedback.

Finally, we recruited participants from a single platform (TalkLife) and only for providing empathic support in the English language. We further note that this study focuses on empathy expression in peer support and does not investigate long-term clinical outcomes.

## 2.4 Related Work

### Methods to operationalize and measure key mental health constructs

Prior work has proposed AI-based methods to operationalize and measure key mental health constructs including adaptability and efficacy of counselors (Pérez-Rosas et al., 2019), personalized vs templated counseling language (Althoff et al., 2016), psychological perspective change (Althoff et al., 2016; Wadden et al., 2021), staying on topic (Wadden et al., 2021), therapeutic actions (Lee et al., 2019), moments of change (Pruksachatkun et al., 2019), counseling strategies (Pérez-Rosas et al., 2022; Shah et al., 2022), and conversational engagement (Sharma et al., 2020a). Here, we focus on a particular conversation technique, *empathy*, which is key in counseling and mental health support (Castonguay and Hill, 2017; Robert et al., 2011). Previous research has analyzed empathy in health communities (Khanpour et al., 2017), face-to-face therapy (Gibson et al., 2016), motivational interviewing (Pérez-Rosas et al., 2017), and emotionally-grounded conversations (Rashkin et al., 2019). Small-scale studies on manually annotated datasets have also been conducted (Morris and Picard, 2012; Lord et al., 2015). Our work develops computational methods for identifying and improving empathy in text-based, asynchronous mental health support, informed by psychology and psychotherapy research. We learn to improve the expression of empathy in online mental health support conversations through a reinforcement learning method for empathic rewriting.

### Empathic dialogue generation

Our task of empathic rewriting is related to empathic dialogue generation but has a key difference as it involves making empathic changes to *existing* responses instead of generating new responses from scratch. While research on generating empathic dialogue has mainly focused on chit-chat, open-domain conversations (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020), we work on conversations in online mental health support. Moreover, most empathic dialogue generation methods have a tendency to enable empathic conversations through emotional grounding (Rashkin et al., 2019) or emotion mimicking (Majumder et al., 2020). In mental health support, however, communicating the cognitive aspects of empathy, related to understanding the experiences and feelings of others, is more valued by mental health professionals (Truax and

Carkhuff, 1967; Selman, 1980). We extend this work with the task of empathic rewriting and by leveraging both emotional and cognitive aspects of empathy, using a theoretically grounded framework of empathy.

## Text rewriting and AI-assisted systems

Text rewriting is a broad subarea in natural language processing that includes tasks such as style transfer (Shen et al., 2017; Li et al., 2018), content debiasing (Pryzant et al., 2020; Ma et al., 2020), and controllable text generation (Hu et al., 2017; Dathathri et al., 2020; Mai et al., 2020). We propose empathic rewriting as a new text rewriting task in which conversational utterances are rewritten to make them more empathetic. This task presents unique challenges different from other text rewriting tasks: it requires understanding empathy in conversational contexts and leveraging that understanding for making empathic changes while ensuring high conversational quality in terms of language fluency, context specificity, and diversity.

Here, we propose a reinforcement learning (RL) model for the task of empathic rewriting. Previous work has used RL for the task of sentiment transfer (Luo et al., 2019) by only using text generations as actions. Here, we design an RL agent that simultaneously learns to (a) identify positions for making improvements and (b) generate empathic sentences for insertion or replacement at the identified positions. These actions are important because the task of empathic rewriting requires changes that go beyond simple word-level transformations, as common in sentiment transfer tasks (e.g., change "*bland*" to "*delicious*" in "*the food was bland*" for transferring from negative to positive sentiment).

Prior work has built systems that leverage the identification of effective conversational strategies such as asking open-ended questions for training users in counseling (Tanana et al., 2019). Computational methods that can perform empathic rewriting can be used to suggest ways to make conversations more empathic in similar feedback and training systems for mental health support and counseling. In a related context, researchers have built AI tools for writing assistance in negotiations (Zhou et al., 2019), composing emails (Chen et al., 2019), language translation (Santy et al., 2019), creative writing (Clark et al., 2018), and communication of politeness (Fu et al., 2020).

## 2.5 Summary of Contribution to Thesis

In this chapter, we developed and evaluated a Human-AI collaboration system that led to a 19.6% increase in empathy in peer-to-peer conversations overall (Cohen’s  $d = 0.24$ ) and a 38.9% increase in empathy for mental health supporters who experience difficulty in writing responses (Cohen’s  $d = 0.43$ ) in a randomized controlled trial on a large peer-to-peer mental health platform. This system was based on technical advancements in understanding empathy in conversations and rewriting conversations to make them more empathetic. Concretely, we developed a new framework, dataset, and computational method for understanding expressed empathy in text-based, asynchronous conversations on mental health platforms. Our computational approach effectively identified empathy with underlying rationales. Moreover, the identified components were found to be important to mental health platforms. Also, we proposed a new task of empathic rewriting for transforming low-empathy conversational posts in online mental health support platforms to higher empathy. For this task, we developed and trained PARTNER, a reinforcement learning model which makes sentence-level edits to posts for making them empathic. We show that PARTNER can effectively generate more empathic posts and outperforms baseline methods from related tasks.

Our findings demonstrate the potential of feedback-driven, AI-in-the-loop writing systems to empower online peer supporters to improve the quality of their responses without increasing the risk of harmful responses.

## Chapter 3

# Human-AI Collaboration to Empower People Seeking Support

In the previous chapter, we explored how to develop human-AI collaboration systems to support those who provide mental health support. Next, we shift our focus to individuals seeking mental health support and study how human-AI collaboration can support them. We focus on self-guided mental health interventions and investigate the potential of human-AI collaboration systems to make these interventions more effective.

Self-guided mental health interventions, such as tools to journal and reflect on negative thoughts, offer great promise to expand modes of care and help people learn coping strategies (Schleider et al., 2020; Patel et al., 2020; Schleider et al., 2022; Shkel et al., 2023). While these interventions are not intended to replace formal psychotherapy, they provide immediate “on-demand” access to resources that can help develop techniques for mental well-being, especially for those who do not have access to a trained professional, are on waiting lists, or seek to supplement therapy with other forms of care (Adair et al., 2005).

However, developing interventions that individuals can effectively use without the assistance of a professional therapist is challenging (Garrido et al., 2019). Currently, most self-guided interventions simply transform traditional manual therapeutic worksheets into digital online formats with limited instructions and support (Shkel et al., 2023). Using these worksheets without professional support often leads to cognitively demanding and emotionally triggering experiences,

limiting engagement and usage (Garrido et al., 2019; Baumel et al., 2019; Fleming et al., 2018; Torous et al., 2020). For example, a popular self-guided intervention involves independently practicing *Cognitive Restructuring of Negative Thoughts*, an evidence-based, well-established process that helps people notice and change their negative thinking patterns (Beck, 1976; Burns, 1980). However, the practice includes complex steps like *identifying thinking traps* (faulty or distorted patterns of thinking like “*all-or-nothing thinking*”), which pose a significant challenge for many (Beck, 1976; Burns, 1980). Most individuals lack the necessary knowledge or experience to successfully use such interventions independently without explicit training and support. Moreover, analyzing one’s own thoughts, emotions, and behavioral patterns can be emotionally triggering, especially for those actively experiencing distress. Such accessibility barriers inhibit the widespread adoption of self-guided mental health interventions.

AI-based methods may be able to assist individuals in engaging with self-guided mental health interventions to improve intervention accessibility and effectiveness. Specifically, language models can help individuals learn and practice cognitively demanding tasks (e.g., through automatic suggestions on potential thinking traps in a thought). Moreover, the language model support could help users manage emotionally triggering thoughts and potentially enable a reduction in the emotional intensity of negative thoughts.

Previous work exploring self-guided mental health interventions based on human-language model interaction has predominantly been limited to small-scale (Ly et al., 2017) and wizard-of-oz-style research (Smith et al., 2021; Morris et al., 2015; Kornfield et al., 2023; Kumar et al., 2023). These studies were conducted in controlled lab settings and evaluated on online crowd-worker platforms, such as MTurk, which may not accurately represent the people who actively seek mental health care or use such an intervention (Mohr et al., 2017). Much less is known about intervention effectiveness in ecologically valid settings with individuals experiencing mental health challenges and seeking care. This limits our understanding of end-user preferences within these emerging forms of intervention (Mohr et al., 2017; Blandford et al., 2018; Borghouts et al., 2021; Poole, 2013). Furthermore, language models can exhibit biases resulting in highly varied performance across people from diverse demographics and populations (Caliskan et al., 2017; Blodgett et al., 2020; Lin et al., 2022). There is a need to investigate and improve the equity of language modeling-based interventions.

In this work, we study supporting self-guided mental health interventions with human-AI collaboration. Specifically, we take *Cognitive Restructuring*, an evidence-based self-guided intervention to help overcome negative thinking, as a case study. We design and evaluate a novel language model-based method for cognitive restructuring. We conduct an ecologically valid and large-scale randomized field study on Mental Health America ([screening.mhanational.org](https://screening.mhanational.org); MHA; a popular mental health website that hosts mental health tools and resources) with over 15,000 participants. We investigate its impact on people seeking mental health care, and evaluate and improve its equity across key subpopulations.

### 3.1 Cognitive Restructuring of Negative Thoughts Through Human-AI Collaboration

Negative thoughts are a natural part of human cognition. However, for people experiencing mental health challenges, such thoughts are often entrenched, automatic, and emotionally triggering, making it difficult to overcome them in-the-moment (Beck, 1976). An evidence-based, well-established therapeutic intervention to overcome negative thoughts is *Cognitive Restructuring*. Cognitive Restructuring fosters awareness of and methods for changing negative thinking patterns (Beck, 1976; Beck et al., 2015). Cognitive Restructuring has been proven to be an effective treatment strategy for psychological disorders, especially anxiety and depression (Clark, 2013). It is a process that is central to Cognitive Behavioral Therapy (Beck, 1976), a modality of treatment which has been demonstrated to be as effective as, or more effective than, other forms of psychological therapy or psychiatric medications (Hofmann et al., 2012; Butler et al., 2006).

An individual initiates this process by writing the negative thought they are struggling with. Next, they try to *identify any potential thinking traps* (biased or irrational patterns of thinking) in their thought. Thinking Traps, alternatively known as *cognitive distortions*, refer to biased or irrational patterns of thinking that lead individuals to perceive reality inaccurately (Beck, 1976; Ding et al., 2022). These typically manifest as exaggerated thoughts, such as making assumptions about what others think (“*Mind reading*”), thinking in extremes (“*All-or-nothing thinking*”), jumping to conclusions based on one experience (“*Overgeneralizing*”), etc. Finally, the individual *write a reframed thought* which involves appropriately addressing their underlying thinking traps

and coming up with a more balanced and helpful perspective on the situation.

As an example, consider a PhD student who has struggled with their research project and starts to worry, “*I’ll never complete my PhD*”. Some possible thinking traps that this thought is falling into include *Catastrophizing* (thinking of the worst-case scenario) and *Fortune telling* (trying to predict the future). Addressing these thinking traps, a possible way in which the student can reframe the thought could be to say, “*I’m imagining the worst-case scenario. This project did not work out, but I was able to formulate meaningful hypotheses for my next research project.*”

Prior NLP research has developed methods for a range of text reframing tasks like sentiment and empathy rewriting (Reif et al., 2022; Sharma et al., 2021) and more recently, positive reframing (Ziems et al., 2022). However, little is known about how to develop cognitive restructuring methods that automatically generate highly effective reframed thoughts that can assist individuals in their cognitive restructuring process.

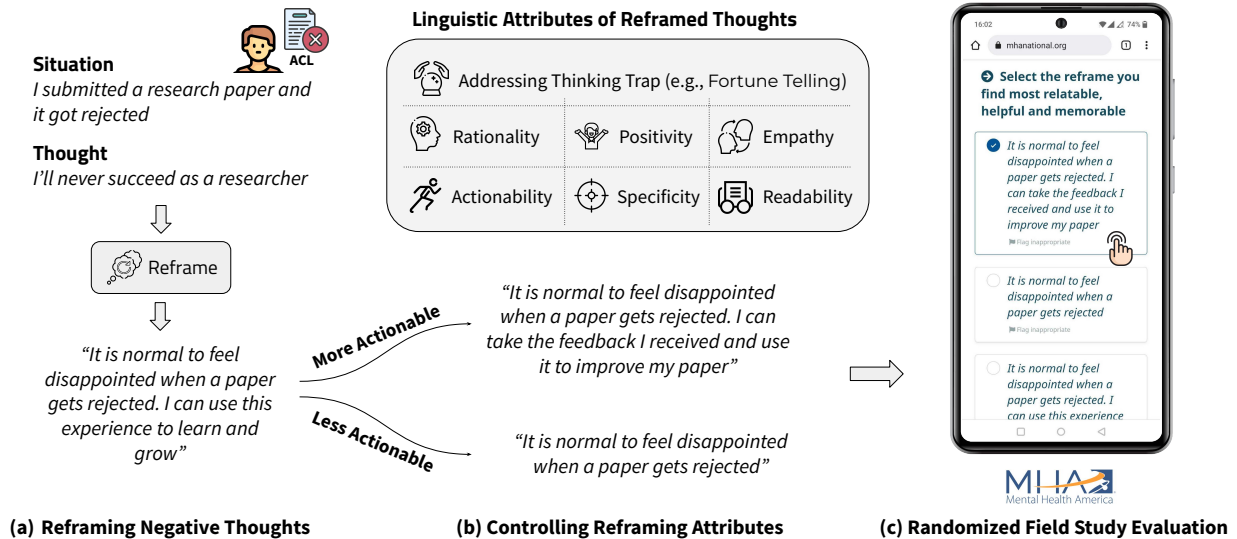
Here, we conduct a study of how language models can be used to reframe negative thoughts (Figure 3.1). We study ways in which a negative thought can be reframed, how LMs can be utilized to perform this reframing, and what types of reframes are preferred by people who experience negative thoughts.

First, in collaboration with clinical psychologists and mental health professionals, we develop a new conceptual framework for characterizing the ways in which a thought might be reframed. We synthesize the most prominent cognitive restructuring processes used in therapy and define seven linguistic attributes of reframed thoughts: whether the reframe *addresses “thinking traps”* (faulty or distorted patterns of thinking), whether it is *rational, positive, empathic, actionable, specific* and *readable*. Building on prior research, we develop automated metrics to measure these attributes and establish construct validity by correlating them with judgements from mental health practitioners.

Next, to develop models for the cognitive restructuring task, we collect and share<sup>1</sup> a dataset from mental health practitioners and clinical psychology graduate students. The dataset includes 600 situations and thoughts with expert-suggested reframes as well as annotations of the proposed reframing attributes. Using this dataset, we develop a retrieval-enhanced in-context learning method (Brown et al., 2020) to *generate* reframed thoughts and to *control* their linguistics-

---

<sup>1</sup>Our code and datasets are available at [github.com/behavioral-data/Cognitive-Reframing](https://github.com/behavioral-data/Cognitive-Reframing).



**Figure 3.1:** (a) We consider the task of reframing negative thoughts with different, more hopeful thoughts using LMs; (b) Different perspectives on a situation may result in different reframes. Here, we propose a framework of seven reframing attributes (see gray box). Given a reframed thought, we control each attribute (e.g., *actionability*) to generate reframes that score higher or lower on that attribute (e.g., *more or less actionable*); (c) We deploy this model on Mental Health America, a large U.S. national mental health website ([bit.ly/changing-thoughts](https://bit.ly/changing-thoughts)) and conduct a randomized field study with 2,067 participants. We suggest LM-generated reframes to MHA visitors and assess which reframing attributes are desirable and what constitutes a reliable, helpful and memorable reframe.

tic attributes. We show that this method achieves the highest overlap with expert-suggested reframes and the highest reliability and helpfulness ratings based on evaluation from mental health experts, when compared to popular NLP baselines.

We investigate which reframing attributes are desirable and what constitutes a reliable, helpful, and memorable reframe. In collaboration (and co-authorship) with mental health experts, and after appropriate ethical review, we deploy a month-long randomized field study on Mental Health America (MHA; a popular website that shares mental health resources and tools online), with 2,067 participants with informed consent. We ask MHA visitors to describe situations and negative thoughts they are experiencing and then suggest LM-generated reframed thoughts with varying linguistic attributes. We find that highly specific and highly empathic reframing is the most preferred and highly specific and actionable reframing is considered the most helpful and memorable. However, we find that reframes that are highly positive are less preferred. These findings provide key implications for cognitive restructuring of negative thoughts and for the use of Human-AI collaboration in this process.

### 3.1.1 Problem Definition and Goals

We work on the task of *Cognitive Restructuring*. Given a situation  $S_i$  and a negative thought  $T_i$ , the task is to generate a reframed thought  $R_i$ .

Psychotherapy literature (Beck, 1976) highlights several desirable outcomes for a successful restructuring: **(a) Reduction in Emotion Intensity**: the restructuring must help the individual *reduce* the intensity of their negative emotion, **(b) Reframe Relatability**: the reframed thought must be *relatable* to them, **(c) Reframe Helpfulness**: the reframe must *help* them overcome the negative thought, **(d) Reframe Memorability**: the reframe must be *memorable* the next time a similar negative thinking pattern emerges, and **(e) Skill Learnability**: the restructuring must help them *learn* how to overcome negative thoughts.

Here, we aim to understand what constitutes successful cognitive restructuring and how language models can assist people in this process. Toward this goal, we characterize the linguistic attributes of reframed thoughts, develop methods to generate reframes and to measure and control their attributes, and investigate which linguistic attributes are related to different reframing outcomes.

### 3.1.2 Framework of Linguistic Attributes of Reframed Thoughts

We draw from clinical therapy practices and collaborate with mental health experts (some of whom are co-authors) to develop a framework of linguistic attributes of reframed thoughts. We illustrate these attributes with the following example for all reframes below – Situation: “*I submitted a research paper and it got rejected;*” Thought: “*I’ll never succeed as a researcher.*”

**Addressing Thinking Traps.** Negative thinking often falls into common patterns, called “*thinking traps*.” Also called *cognitive distortions*, these include exaggerated and biased patterns of thinking which cause individuals to perceive reality inaccurately (Beck, 1976; Ding et al., 2022). Common thinking traps include: assuming what others think (“*Mind reading*”), thinking in extremes (“*All-or-nothing thinking*”), focusing on the worst-case scenario (“*Catastrophizing*”), trying to predict the future (“*Fortune telling*”), etc. See Appendix B.3 for the full list.

A reframe may or may not directly address one or more of the thought’s thinking traps. A reframe like “*I don’t know what the future will bring*” directly addresses the thinking trap “*Fortune*

telling,” whereas a reframe like *“I will surely become a successful researcher”* does not address this thinking trap but rather continues to express it.

**Rationality.** Another strategy to reframe a thought is to reflect on evidence for and against it and reason about what these evidence imply (Beck, 1976). For example, the rejection of the paper is evidence of having the thought *“I’ll never succeed as a researcher.”* However, the evidence against this thought could be that acceptance or rejection of one paper does not make someone a failure, which may lead to a reframe *“Just getting one paper rejected doesn’t define my failure.”* A rational reframe is guided by such strong evidence whereas an irrational reframe is based on unrealistic assumptions.

**Positivity.** A reframe of a negative thought tries to emphasize the positive perspectives of the situation but different reframes may have different levels of positivity. An overly positive reframe like *“I’m going to win best paper awards for every paper from now on”* exaggerates the positive perspectives, which is likely to set the person up for disappointment rather than success (Dember and Penwell, 1980). On the other hand, a balanced response like *“I may or may not succeed, but I’ll keep trying”* considers both positive and negative perspectives of the situation.

**Empathy.** It can be helpful to acknowledge the feelings caused by negative thoughts (Allen and Leary, 2010; Robert et al., 2011). A reframe may express empathy or self-compassion by validating how one is feeling. E.g., *“It is okay to feel disappointed when a paper gets rejected.”*

**Actionability.** To encourage pleasant emotions, one commonly used therapeutic approach is Behavioral Activation (Dimidjian et al., 2011; Burkhardt et al., 2021). This involves engaging in behaviors or actions that may help in overcoming negative thoughts. A reframe may suggest specific actions (e.g., *“I can take the feedback I received and use it to improve my paper”*), may not suggest specific actions but be actionable (e.g., *“I can use this experience to learn and grow”*) or may not be actionable at all (e.g., *“I may or may not become a successful researcher”*).

**Specificity.** A reframe may specifically address the situation and the thought (e.g., *“One paper rejection doesn’t define my failure as a researcher”*) or may be generic enough to be applicable to a wide range of negative situations and thoughts (e.g., *“I’m going to succeed”*). While a specific reframe may be more helpful in-the-moment, a generic reframe could be effective for recurring thoughts, which are frequently a result of the “core” beliefs that a person holds (Beck, 2005;

David et al., 2009).

**Readability.** The linguistic reasoning capabilities of individuals may be different (e.g., across age groups or education levels) (Kaplan et al., 1995). Accordingly, a reframe may either be simple or complex to read (e.g., *“I’ll do well in the future”* vs. *“I’m resolute in my ambition to succeed”*).

### 3.1.3 Data Collection

To facilitate computational methods for cognitive reframing, we collect a dataset of reframed thoughts, annotated with their linguistic attributes.

#### Curated Situations & Negative Thoughts

We start by curating data sources for situations and negative thoughts.

**Thought Records Dataset (Burger et al., 2021).** This dataset contains hypothetical and real-world situations, thoughts and emotional processes reported by crowdworkers on Amazon Mechanical Turk. We manually curate 180 pairs of diverse situations with negative thoughts from this dataset.

**Mental Health America (MHA).** Situations and thoughts from crowdworkers may not reflect the broad range of mental health challenges that people face in real-life. To incorporate more real-world situations and thoughts, we ran a survey on the MHA website ([screening.mhanational.org](https://screening.mhanational.org)). MHA visitors (who typically use the website for screening of mental illnesses) were asked to describe any negative thoughts and the associated situations they were struggling with. We manually curate 120 pairs of self-reported situations and thoughts to ensure broad coverage of relevant topics based on high diversity and manual filtering.

#### Annotation Task and Procedure

Reframing negative thoughts is a cognitively difficult process that requires practice and training, making crowdwork data collection approaches challenging. To ensure high-quality reframes and annotations, we recruited 15 current mental health practitioners and clinical psychology graduate students with significant practical experience in cognitive reframing.<sup>2</sup> For each (situation,

---

<sup>2</sup>For recruitment, we advertised our study through university mailing lists and newsletters of a mental health organization. Recruited experts were paid @ 37.5 USD / hr.

thought) pair in our data source, we ask them to (1) write two different reframed thoughts, (2) annotate the thinking traps addressed by each reframed thought and (3) compare the two reframes and choose the one that is more rational, more positive, more actionable, more empathic, more specific and more readable. In total, we collect 600 reframed thoughts with annotations on their linguistic attributes. We share this dataset publicly at [github.com/behavioral-data/Cognitive-Reframing](https://github.com/behavioral-data/Cognitive-Reframing).

## Ethics and Safety

Our data collection was conducted after a review and approval by the University of Washington’s Institutional Review Board and informed participants about study purpose, risks, and data collection. All participants were 18 or older, provided informed consent, and were given access to a crisis hotline. We do not assess any clinical outcomes.

### 3.1.4 Method

We design automated metrics for measuring linguistic attributes, and develop methods to generate reframed thoughts and to control their attributes.

## Measuring Reframing Attributes

**Addressing Thinking Traps.** Given a situation  $S_i$ , a negative thought  $T_i$  and a reframed thought  $R_i$ , our goal is to identify the set of thinking traps addressed by the reframed thought. We approach this as a multi-label classification task, and fine-tune a GPT-3 model<sup>3</sup> on the expert-annotated thinking trap labels collected in Section 3.1.3.

**Rationality.** Rationality is the quality of being guided by reasons (Danielson et al., 2004). Here, we operationalize rationality of a reframed thought  $R_i$  as its *reasoning strength* and ask the following two questions: (1) What might be the reasoning behind  $R_i$ ?; (2) Are the reasons sound? To understand the reasoning behind  $R_i$ , we develop an *abductive explanation* based method (Peirce, 1974; Bhagavatula et al., 2020; Jung et al., 2022). For a given  $(S_i, T_i)$ , we use a language model

---

<sup>3</sup>We use `text-davinci-003` as our GPT-3 model for all experiments.

to generate (a) the most plausible explanations that *support*  $\mathbf{R}_i$  and (b) the most plausible explanations that *refute* it. Moreover, to check if the explanations are sound, we recursively generate explanations behind the explanations to test their reasoning strength (Appendix B.4). Let  $sup(\cdot)$  be a generator function that generates explanation *supporting* a reframe and let  $ref(\cdot)$  be a generator function that generates explanation *refuting* a reframe. Then, we recursively define reasoning strength  $RS(\mathbf{S}_i, \mathbf{T}_i, \mathbf{R}_i)$  as

$$\begin{aligned} & \left( P(\mathbf{R}_i = \text{sound} | \mathbf{S}_i, \mathbf{T}_i) \times \mathbb{E}_{r \sim sup(\cdot)} [RS(\mathbf{S}_i, \mathbf{T}_i, r)] \right) \\ & - \left( P(\mathbf{R}_i = \text{flawed} | \mathbf{S}_i, \mathbf{T}_i) \times \mathbb{E}_{r \sim ref(\cdot)} [RS(\mathbf{S}_i, \mathbf{T}_i, r)] \right) \end{aligned}$$

To design the explanation generator functions,  $sup(\cdot)$  and  $ref(\cdot)$ , we leverage in-context learning (Brown et al., 2020). In collaboration with mental health experts, we design 10 demonstration examples of situations, thoughts and reframed thoughts with explanations that support (“*This reframed thought is sound because...*”) and refute (“*This reframed thought is flawed because...*”) a particular reframe. We use these examples to prompt GPT-3. Moreover, to estimate the probabilities  $P(\mathbf{R}_i = \text{sound})$  and  $P(\mathbf{R}_i = \text{flawed})$ , we use the token probability of generating “*sound*” and “*flawed*” respectively, given  $\mathbf{S}_i, \mathbf{T}_i, \mathbf{R}_i$  and the text “*This reframed thought is*” as input to GPT-3.<sup>4</sup>

**Positivity.** To measure the positivity of the generated reframed thought, we use a RoBERTa-based sentiment classifier fine-tuned on the TweetEval benchmark (Barbieri et al., 2020).

**Empathy.** To measure empathy, we build upon the empathy classification model presented in Sharma et al. (2020b). This RoBERTa-based model leverages a theoretically-grounded framework of empathy consisting of three empathy communication mechanisms (emotional reactions, interpretations, and explorations) and predicts empathy levels in mental health conversations on a scale from 0 to 6. Here, we further fine-tune this model on the domain of reframed thoughts through a manually labeled dataset of 300 reframed thoughts with empathy labels (labeled by one author with expertise in empathy in mental health context).

**Actionability.** To measure actionability, we hypothesize that an actionable reframe is one that either (1) *suggests a concrete action* or (2) *does not suggest a concrete action but is easy to act upon*.

---

<sup>4</sup>We experimented with different alternatives for “*sound*” and “*flawed*” and observed similar results.

We cast action concreteness as a binary classification task: given reframe  $\mathbf{R}_i$ , predict  $\text{contains\_action}(\mathbf{R}_i) \in \{0, 1\}$ . We make few-shot predictions by prompting GPT-3 with 10 examples of reframed thoughts paired with actionability ratings from Section 3.1.3 (details in Appendix B.1).

To determine the ease with which  $\mathbf{R}_i$  can be acted upon, we examine the next set of actions entailed by  $\mathbf{R}_i$ . Our hypothesis is that a *diverse* next action set may indicate ambiguity which might be less actionable, whereas a *coherent* next action set may indicate clarity which might be more actionable. Here, we instruct GPT-3 to generate  $k = 5$  next action candidates given a reframed thought (instruction prompting; zero-shot). We compute the next action coherence — denoted  $\text{next\_action\_coherence}(\mathbf{R}_i)$  — by embedding each of the  $k$  action candidates using RoBERTa (Liu et al., 2019) and computing the average pairwise cosine similarity. Higher similarity indicates greater coherence among the possible next actions. Our overall actionability measurement is defined as  $\text{contains\_action}(\mathbf{R}_i) + \text{next\_action\_coherence}(\mathbf{R}_i)$ .

**Specificity.** Following prior work (Xu et al., 2018; Sharma et al., 2021), we measure specificity using sentence embedding similarity between the reframed thought  $\mathbf{R}_i$  and the concatenation of the situation  $\mathbf{S}_i$  and the thought  $\mathbf{T}_i$  (using RoBERTa embeddings (Liu et al., 2019)).

**Readability.** We employ the commonly used Coleman-Liau Index (CLI) metric (Coleman and Liau, 1975) which assesses readability based on the character and word structure within a sentence. The Coleman-Liau Index is calculated as  $0.0588L - 0.296S - 15.8$ , where  $L$ : average number of letters per 100 words;  $S$  is the average number of sentences per 100 words.

## Reframe Generation

In-context learning methods can learn to generalize NLP tasks from a handful of examples (*few-shot learning*) or from hand-written instructions alone (*instruction prompting*) (Brown et al., 2020). However, through a qualitative analysis of 100 manually written situations and thoughts, we found that a simple in-context learning method with a fixed set of examples often failed to appropriately reframe situations and thoughts for which no relevant in-context examples were provided (e.g., someone with anxiety having “*racing thoughts*”).

To appropriately reframe thoughts related to a range of situations and thoughts, we de-

Attribute	Pearson Correlation
Addressing Thinking Traps	0.680***
Rationality	0.448**
Positivity	0.550***
Empathy	0.575***
Actionability	0.647***
Specificity	0.427**
Readability	0.331*

**Table 3.1:** Correlation of our proposed attribute measures by with human judgments from mental health experts. \*:  $p < 0.05$ ; \*\*:  $p < 0.001$ ; \*\*\*:  $p < 10^{-5}$ .

velop a retrieval-based in-context learning method (Liu et al., 2022b). For each situation  $S_i$  and negative thought  $T_i$ , we retrieve  $k$ -similar examples from our dataset (Section 3.1.3). We first encode situations and thoughts using RoBERTa embeddings. Then, we retrieve  $k$  examples,  $\{(s_1, t_1), \dots, (s_k, t_k)\}$ , from our dataset based on the top- $k$  values of  $\text{cosine\_sim}(\text{concat}(s, t), \text{concat}(S_i, T_i))$ . We choose  $k = 5$  (Appendix B.1).

### Controlling Linguistic Attributes of Generated Reframes

While our proposed method allows us to generate a single reframe, it does not directly give us control over its linguistic attributes beyond mimicking the retrieved examples (Section 3.1.2). Here, we intend to vary the linguistic attributes of the reframes.

A reframed thought may or may not address a thinking trap in the original thought  $T_i$ . Here, we generate two reframes  $R_i^{(tt,Y)}$  and  $R_i^{(tt,N)}$ , one that addresses the thinking traps in  $T_i$  and another that does not address it.<sup>5</sup> We extract two separate sets of in-context examples from our dataset – those that address at least one thinking trap and those that do not (as collected in Section 3.1.3). We use those examples to prompt GPT-3 to generate  $R_i^{(tt,Y)}$  and  $R_i^{(tt,N)}$ .

Moreover, a reframed thought may have high or low rationality, positivity, empathy, actionability, specificity and readability values. For these six attributes, given a reframe  $R_i$  and a linguistic attribute  $a$ , we generate two *additional* reframes  $R_i^{(a,H)}$  and  $R_i^{(a,L)}$ , one that scores higher on attribute  $a$  and another that scores lower on it (e.g., higher or lower actionability). To accomplish this, recall that each (situation, thought) pair from Section 3.1.3 is annotated with

<sup>5</sup>If a thought exhibits multiple thinking traps, we check if the reframe addresses at least one of them.

Model	Automatic			Human		
	BLEU	R-1	R-L	BScore	Rel.	Help.
Retrieval Only	21.6	18.8	14.2	86.7	2.58	3.14
Pos. Reframing	24.4	23.6	17.6	87.6	2.67	2.40
DialoGPT	22.5	17.4	13.5	86.3	2.49	3.21
T5	24.9	23.4	17.8	87.2	2.51	3.30
GPT-3 Only	25.0	23.9	18.0	88.3	2.97	3.98
Our Model	<b>27.8</b>	<b>26.0</b>	<b>19.9</b>	<b>88.6</b>	<b>3.10</b>	<b>4.11</b>

**Table 3.2:** Automatic and Human Evaluation Results. R-1: ROUGE-1; R-L: ROUGE-L; BScore: BertScore; Rel.: Relatability; Help.: Helpfulness.

two reframes and that the reframes are compared along each linguistic attribute. For a human-annotated instance  $j$ , let  $\mathbf{R}_j^{*(a,H)}$  and  $\mathbf{R}_j^{*(a,L)}$  be the reframes judged to be high and low on attribute  $a$ , respectively. To generate  $\mathbf{R}_i^{(a,H)}$  from  $\mathbf{R}_i$ , we prompt GPT-3 with in-context examples  $\{\mathbf{R}_j^{*(a,L)} \rightarrow \mathbf{R}_j^{*(a,H)}\}_{j=1}^k$ , using  $k = 5$ . Similarly, to generate  $\mathbf{R}_i^{(a,L)}$  from  $\mathbf{R}_i$ , we prompt GPT-3 with examples  $\{\mathbf{R}_j^{*(a,H)} \rightarrow \mathbf{R}_j^{*(a,L)}\}_{j=1}^k$ .

### 3.1.5 Experiments and Results

We assess the construct validity of proposed linguistic attributes and evaluate the performance of the reframe generation model.

#### Construct Validity of Linguistic Attributes

We validate our proposed linguistic attribute measures by correlating them with the human judgments of mental health experts, as obtained in Section 3.1.3. We find a strong Pearson correlation for addressing thinking traps (0.680\*\*\*) and actionability (0.647\*\*\*), a moderate correlation for rationality (0.448\*\*), positivity (0.550\*\*\*), empathy (0.575\*\*\*) and specificity (0.427\*\*), and a weak correlation for readability (0.331\*) (Table 3.1).<sup>6</sup>

<sup>6</sup>\*,  $p < 0.05$ ; \*\*,  $p < 0.001$ ; \*\*\*,  $p < 10^{-5}$

## Reframe Generation Performance

We use both automatic and human evaluation to assess the performance of our proposed reframe generation model as developed in Section 3.1.4.

**Experimental Setup.** We use top- $p$  sampling with  $p = 0.6$  for text generation (Holtzman et al., 2020). We split the 600 expert-annotated examples (Section 3.1.3) into train and test using a 70:30 split.

**Baselines.** (1) *Retrieval Only* – For a test input, we retrieve the training set example with the highest cosine similarity based on RoBERTa embeddings; (2) *Positive Reframing* – We reuse the BART-based positive reframing model from Ziems et al. (2022); (3) *DialoGPT* – GPT-2 adapted to dialogue (Zhang et al., 2020); (4) *T5* – Text-to-text transfer LM (Raffel et al., 2020);<sup>7</sup> (5) *GPT-3 Only* – We randomly retrieve 5 examples from our training set and use them to prompt GPT-3.

**Automatic Evaluation.** We examine the semantic similarity between the model outputs and the ground truth reframings in the above-created test split. We use BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-L (Lin, 2004) and the BERTScore (Zhang et al., 2019b) metrics. We find that our proposed model has an 11.2% higher BLEU score and 9.7% higher ROUGE scores than the next best-performing baselines – GPT-3 Only and Positive Reframing (Table 3.2).

**Human Evaluation.** We assess the two key reframing outcome metrics of *reliability* (how reliable would a reframed thought be) and *helpfulness* (how helpful would a reframed thought be in overcoming negative thoughts). We recruit three mental health practitioners. We ask them to rate the models’ outputs on test set examples based on their reliability and helpfulness on a 1 to 5 scale. We find that our proposed model achieves the highest reliability and helpfulness ratings (Table 3.2). Surprisingly, the Positive Reframing method showed the least helpfulness and low reliability, indicating that just reframing negative thoughts based on positivity may not be highly reliable and helpful.

---

<sup>7</sup>Training DialoGPT and T5 on 600 samples only may be challenging. Here, we use an *overgeneration* strategy – Starting from our collected data samples, we utilize the pattern replication capabilities of GPT-3 to generate 10,000 more examples, similar to Liu et al. (2022a).

## 3.2 Randomized Field Study on a Large Mental Health Platform

Next, we study how our proposed cognitive restructuring model may assist people in overcoming negative thoughts.

We conduct an ecologically valid and large-scale randomized field study on Mental Health America ([screening.mhanational.org](https://screening.mhanational.org); MHA; a popular website that hosts mental health tools and resources) with over 15,000 participants (Section 3.2.1). We examine the design of our tool, investigate its impact on people seeking mental health care, and evaluate and improve its equity across key subpopulations. Specifically, we address the following research questions:

**RQ1 – Design.** How can we design a self-guided cognitive restructuring intervention that is supported through human-AI collaboration?

**RQ2a – Overall Effectiveness.** To what extent does human-AI collaboration based cognitive restructuring help individuals in alleviating negative emotions and overcoming negative thoughts?

**RQ2b – Impact of Design Hypotheses.** What is the impact of individual design hypotheses on the intervention effectiveness?

**RQ3a - Preference.** What types of reframed thoughts do people prefer?

**RQ3b - Linguistic Attributes.** How do the linguistic attributes of reframed thoughts relate to the desired outcomes of cognitive restructuring?

**RQ4 – Equity.** How equitable is the intervention and what strategies may improve equity?

To address these research questions, we first formulate the design hypotheses for this intervention through qualitative feedback from participants of early prototypes of the system and brainstorming with mental health professionals (Section 3.2.2). We hypothesize that assisting users in cognitively and emotionally challenging processes, contextualizing thought reframes by reflecting on situations and emotions, integrating psychoeducation, facilitating interactive refinement of reframes, and ensuring safety will result in a more effective human-AI collaboration system for cognitive restructuring. Taking these hypotheses into consideration, we design a new system that uses a language model to support people through various steps of cognitive restructuring, including the *identification* of thinking traps in thoughts and the *reframing* of negative thoughts. Our language model suggests possible thinking traps a given thought may have, as

well as suggests possible ways of reframing negative thought (Section 3.2.2).

After systematic ethical and safety considerations, active collaboration with mental health professionals, advocates, and clinical psychologists, and IRB review and approval, we deploy this system on MHA. We utilize a mixed-methods study design, combining quantitative and qualitative feedback, to assess the outcomes of this system on the platform visitors. We find that 67.64% of participants experience a positive shift (i.e., reduction) in their emotion intensity and 65.65% report helpfulness in overcoming their negative thoughts through the use of our system. Moreover, participants indicate that the system assists them in alleviating cognitive barriers by simplifying task complexity and emotional barriers by providing a less triggering experience (Section 3.2.3). Also, enabling participants to iteratively improve reframes by seeking additional reframing suggestions leads to a 23.73% greater reduction in the intensity of negative emotions. Moreover, participants who choose to make their reframes actionable report superior outcomes compared to those who make them empathic or personalized (Section 3.2.4).

To address the needs of individuals from diverse demographics and subpopulations, it is critical to develop equitable solutions. Here, we evaluate the performance of our system across people of different demographics and subpopulations (Section 3.2.7). The intervention is found to be more effective for individuals identifying as females, older adults, individuals with higher education levels, and those struggling with issues, such as parenting and work. However, it is found to be less effective for individuals identifying as males, adolescents, those with lower education levels, and those struggling with issues such as hopelessness or loneliness. We investigate the potential benefits of customizing the intervention for adolescents, who we found to have one of the largest disparities in our intervention outcomes. We find that making the language model-generated suggestions simpler and more casual leads to a 14.44% increase in reframe helpfulness for adolescents in the age group of 13 to 14.

We discuss the implications of our study for the use of human-language interaction in the development of self-guided mental health interventions, emphasizing the need for effective personalization, appropriate levels of interactivity with the language model, equity across subpopulations, and its safety (Section 3.2.8).

### 3.2.1 Study Overview

Our study was conducted over a nine-month process involving iterative ideation, prototyping, and evaluation in collaboration with mental health experts (some of whom are co-authors).

#### Study Design

**Platform.** We used Mental Health America (MHA) for our study. We deployed our system on this platform, where it was hosted with appropriate ethical considerations and participant consent process (see Privacy, Ethics, and Safety), along with other existing self-guided mental health tools.

**Participants.** Many MHA visitors are interested in mental health resources including self-guided systems. Our study participants were visitors to MHA, who chose to use our system and provided informed consent to participate in the research study. Our study comprised a total of 15,531 participants.<sup>8</sup> We carried out several experiments in parallel, including randomized trials with independent participants. The number of participants involved in each experiment is represented by “N” values throughout. Participants were 13 years and older (also, see Privacy, Ethics, and Safety below for a discussion on our reasons for including minors).

**Tasks and Procedures.** We started our study by formulating design hypotheses for developing self-guided cognitive restructuring through human-language model interaction. This was done through feedback from users of early prototypes of the system, brainstorming with mental health experts, and leveraging pertinent findings from prior work (Section 3.2.2). Based on these design hypotheses, we developed our final system for human-language model interaction powered cognitive restructuring (Section 3.2.2).

We evaluated the effectiveness of our system on several psychotherapy based metrics through a field study on the MHA platform (Section 3.2.3). In addition, we evaluated the importance of our design hypotheses by conducting randomized trials on the platform, explicitly ablating specific design features, such as removing psychoeducation from the system (Section 3.2.4). Finally,

---

<sup>8</sup>Overall, we had 43,347 participants in total who consented initially, but 27,816 (64.17%) dropped out before completing the outcome survey. Analysis in the work only include the 15,531 (35.83%) participants who completed the outcome survey. Note that dropout rates of online self-guided tools are commonly between 70% to 99.5% (Karyotaki et al., 2015; Fleming et al., 2018).

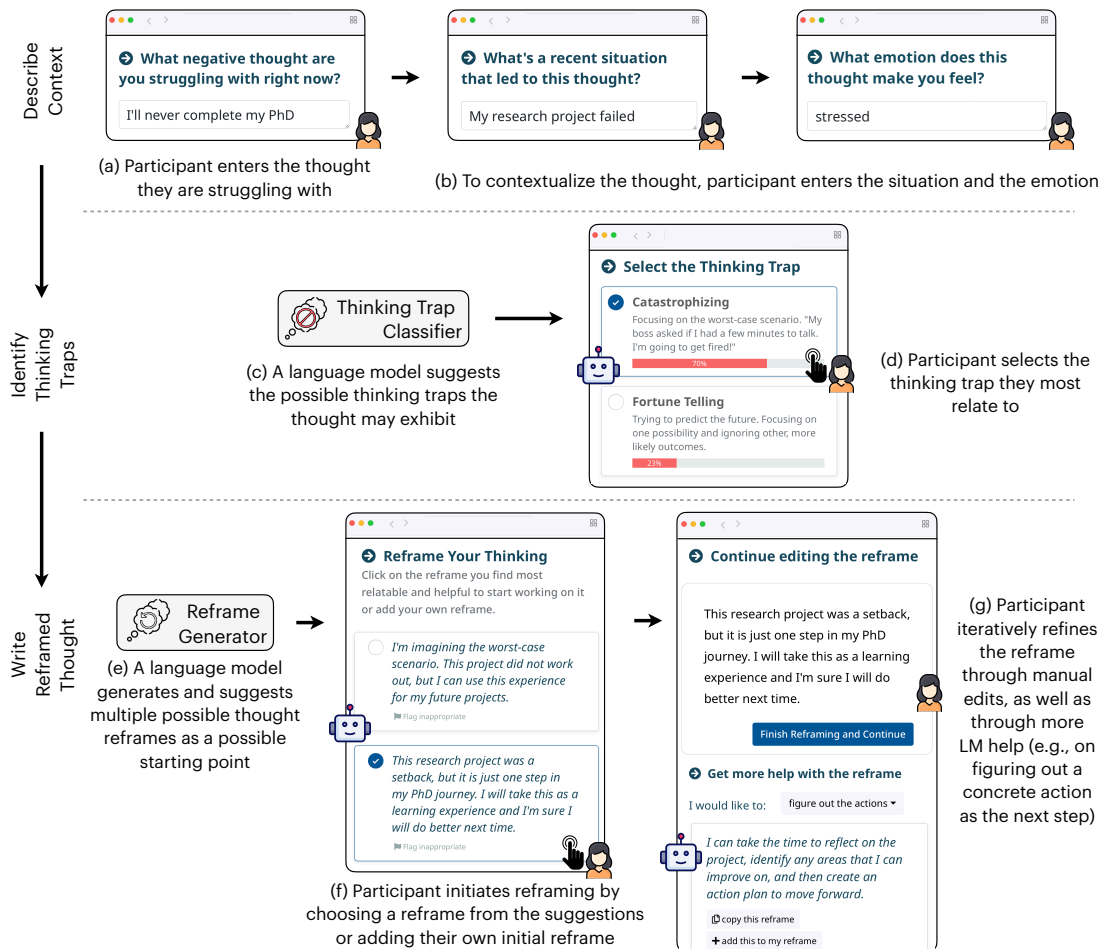
we analyzed and improved the equity of our system (Section 3.2.7).

**Privacy, Ethics, and Safety.** We designed and conducted our field study after carefully reviewing the potential benefits and risks to participants in consultation and collaboration with mental health experts. Our study, including the participation of minors, was approved by our Institutional Review Board. We included minors (those aged between 13 and 17) in this study as they represent a large, key demographic on MHA and were already frequently using similar self-guided interventions outside of our study. Therefore, we concluded that disallowing minors access to some self-guided tools would not decrease risk but potentially make an effective intervention inaccessible to a large fraction of users. We obtained informed consent from adults and informed assent from minors, ensuring they were fully aware of the study’s purpose, risks, and benefits (Appendix Figure B.6 and Figure B.7). Participants were informed that they would interact with an AI-based model that automatically generates reframed thoughts, without any human supervision. Further, appropriate steps were taken to avoid harmful content generation (Section 3.2.2), but participants were informed about the possibility that some of the generated content may be upsetting or disturbing. Also, participants were given access to a crisis hotline. Examples in this work have been anonymized using best privacy practices (Matthews et al., 2017).

For minors, we requested and received a waiver of parental permission from the IRB. This is because discussing mental health issues with parents may pose additional risks including discomfort in disclosing psychological distress (Smith et al., 2022a; Samargia et al., 2006) and reduced autonomy (Wilson and Deane, 2012). Therefore, obtaining parental permission is generally avoided in similar studies (Schleider et al., 2022). Also, given our online setting, obtaining parental permission is impractical and logistically difficult as we are not directly interacting with the participants. We instead obtained the assent of minors as approved by the IRB.

### **3.2.2 RQ1: How can we design a self-guided cognitive restructuring intervention that is supported through human-language model interaction?**

Here, we design a novel system for human-language model interaction based self-guided cognitive restructuring.



**Figure 3.2:** We design a human-language model interaction based system for self-guided cognitive restructuring of negative thoughts. The system involves (a-b) describing the context by participants, (c-d) LM-assisted identification of thinking traps, and (e-g) LM-assisted writing of reframed thoughts.

## Design Hypotheses

The design of our system was based on hypotheses that were formulated by incorporating qualitative feedback from users of early prototypes of the system, brainstorming on design decisions with mental health experts, and leveraging relevant insights from previous mental health studies. Here, we briefly describe the design hypotheses that surfaced through this process. Our primary contributions include the design (Section 3.2.2) and evaluation (Section 3.2.3) of a system based on these hypotheses. In addition, we evaluate the impact of  $H_2$ ,  $H_3$ , and  $H_4$  through randomized trials (Section 3.2.4). Moreover, we evaluate  $H_1$  and  $H_5$  through qualitative analysis of participant feedback (Sections 3.2.3 and 3.2.4 respectively).

**H<sub>1</sub>: Assisting participants in processes that are cognitively and emotionally challenging may improve intervention effectiveness.** Previous research suggests that restructuring negative thoughts can be *cognitively* and *emotionally* challenging (Shkel et al., 2023). The limited availability of mental health professionals and resources often acts as barriers to accessibility, typically resulting in a lack of knowledge and exposure to therapeutic processes among people. Moreover, the entrenched nature of thoughts makes them difficult to overcome in the moment. In fact, many participants during our design exploration phase highlighted these challenges (e.g., a participant wrote, “*I struggle with formulating these thought → situation → thinking trap → reframe scenarios by myself*”; another participant wrote, “*I find it hard to think of these reframes myself in the moment*”; another participant wrote, “*I have a hard time figuring out with cognitive distortions I’m using*”). Therefore, we aim to guide participants through the identification of thinking traps and assist them in writing effective reframes. Here, we leverage language models to achieve this, as detailed in Section 3.2.2. Section 3.2.3 evaluates the effectiveness of this assistance.

**H<sub>2</sub>: Contextualizing thought reframes through situations and emotions may improve intervention effectiveness.** Cognitive Behavioral Therapy posits that our thoughts are shaped by our situations, subsequently influencing our beliefs and emotions (Beck, 1976). Therefore, the ability to reflect on situations and emotions and recognize their connection with negative thoughts could be beneficial in accurately assessing thinking traps and writing effective reframes. Our system enabled participants to contextualize their thoughts by answering additional questions related to their situations and emotions (Section 3.2.2). However, it is well established that introducing additional burdens in the form of questions may lead to increased dropout, a core challenge in digital mental health (Baumel et al., 2019; Torous et al., 2020). Section 3.2.4 describes the randomized trial that assesses the impact of this contextualization on the tradeoff between intervention effectiveness and overall participant engagement.

**H<sub>3</sub>: Integrating psychoeducation may improve intervention effectiveness.** Cognitive Restructuring is a skill people can learn. Learning this skill enables participants to identify their negative thoughts, assess the thinking traps they often fall into, and develop the ability to reframe them into something more hopeful in the moment (Hundt et al., 2013; Strunk et al., 2014). However, acquiring this skill is not straightforward and typically requires comprehensive psychoeducation

coupled with practice (e.g., one participant wrote *“I don’t think that everyone knows about thinking traps and the types or kinds of thinking traps there is, so I think there should be a description or definition about each thinking traps”*). Here, we introduced participants to different thinking traps through definitions and examples, along with strategies to reframe their specific thinking traps (Section 3.2.2). Section 3.2.4 details the randomized trial that evaluates the impact of integrating psychoeducation on skill learnability.

**H<sub>4</sub>: Facilitating interactive refinement of reframes may improve intervention effectiveness.**

A key component of our design involves the suggestion of automatically generated reframes through a language model (H<sub>1</sub>; Section 3.2.2). However, we found that participants of early prototypes of the system desired the ability to have varied AI suggestions. One participant said, *“[I want] more choices and variation in the reframing”*. Another participant said *“I wish it included actionable insights.”* Moreover, many participants desired the ability to interactively refine the reframes. One participant said, *“I need to write the information and revise it as many times as possible.”* Another participant said, *“It could be more interactive, also go more in depth.”*

Therefore, our design included the option for iterative editing and updating of reframes. This enabled participants to seek more specific suggestions from the language model, including *“making it more relatable to their situation”*, *“figuring out the next steps and actions”*, and *“feeling more supported and validated”* (Section 3.2.2). Section 3.2.4 evaluates the effects of this interactive interaction through a randomized trial.

**H<sub>5</sub>: Mechanisms to avoid unsafe content generation and flag inappropriate content may make the intervention more safe.** The need for safety considerations is crucial when intervening in high-stakes settings like mental health (Li et al., 2020; Martinez-Martin et al., 2018). There is a risk that AI might inadvertently harm, rather than help, individuals coping with mental health challenges. Therefore, a key goal was to ensure safety and minimize risks. Our design included mechanisms to avoid unsafe content generation and flag inappropriate content, as described in Section 3.2.2. We discuss the content that was flagged by participants in Section 3.2.4.

## System Design

In our system (Figure 3.2), we guided participants through a five-step process. This included (1) describing the thought, (2) detailing the situation, (3) reflecting on the emotion, (4) identifying the thinking traps, and (5) finally, reframing the thought.

**Step I, II, and III: Participant describes the context.** On selecting to use the system and after consenting to participate in our study (Appendix Figure B.6 and Figure B.7), a participant first articulates the thought they are struggling with (e.g., *“I’ll never complete my PhD”*; Figure 3.2a). Next, to contextualize their thought ( $H_2$ ; Figure 3.2b), the participant describes a recent situation that may have led to this thought (e.g., *“My research project failed”*). Additionally, the participant reflects on the emotion they are currently experiencing along with its intensity (e.g., *“stressed”*; 9 out of 10).

**Step IV: LLM-assisted selection of thinking traps.** The next step is to identify thinking traps. The typical process of identifying thinking traps involves participants navigating through a list to single out possible traps in their thoughts, which we identified as a cognitive and emotional barrier ( $H_1$ ). Here, we use a language model to assist participants in the identification of thinking traps among 13 common thinking traps (see Appendix Table B.3).

For this, we rank the thinking traps for the given thoughts using a language model and show those to the participants along with their individual estimated likelihoods (e.g., *“Catastrophizing – 70%; Fortune Telling – 23%; Overgeneralizing – 7%”*; Figure 3.2c). In order to incorporate psychoeducation, we provide definitions and examples of these thinking traps ( $H_3$ ; Appendix Table B.3). We use the GPT-3 model (Brown et al., 2020) finetuned over a dataset of thinking traps developed in Section 3.1.4. This model achieves a top-1 accuracy of 62.98% on the 13-class thinking trap classification problem.

The participant selects one or more thinking traps from this ranked list that they most closely identify with based on their thinking pattern (e.g., *“Catastrophizing”*; Figure 3.2d).

**Step V: LLM-assisted writing of reframes.** Finally, the participant writes a reframed thought addressing their thinking traps. Writing reframes to negative thoughts while maintaining composure in the moment is challenging ( $H_1$ ), therefore, we use the retrieval-enhanced in-context learning method developed in Section 3.1.4 to assist participants.

Using this language model, we generate multiple suggestions for possible thought reframes and show them to the participant as a potential starting point (Figure 3.2e).<sup>9</sup> The multiple suggestions are aimed at offering varying perspectives to the participant’s original thought (e.g., *“I’m imagining the worst-case scenario. This project did not work out, but I can use this experience for my future projects.”*, *“This research project was a setback, but it is just one step in my PhD journey. I will take this as a learning experience and I’m sure I will do better next time.”*, *“I am disappointed that my research project failed, but I can still complete my PhD if I keep working hard and don’t give up.”*<sup>10</sup>). In addition, to incorporate psychoeducation, we provide instructions on ways in which the participant can reframe the specific thinking traps selected by them in the previous step (e.g., *“Tips to overcome catastrophizing”*; H<sub>3</sub>; Appendix Table B.3).

The participant initiates reframing by choosing a reframe from the initial suggestion list or by writing a reframe on their own (Figure 3.2f). The participant then iteratively refines this reframe through manual edits, as well as through additional, optional help from the language model. For this, we provide participants with an option to seek more specific suggestions from the language model, including *“making it more relatable to their situation”*, *“figuring out the next steps and actions”*, and *“feeling more supported and validated”* (H<sub>4</sub>; Figure 3.2g). For the option selected by the participant, we generate additional suggestions using the language model. Participants can either copy these additional suggestions, add them to their initial reframe, replace their initial reframe with it, or use it as an inspiration. Also, see Appendix Figure B.13 for the detailed interface.

**Safety considerations.** To minimize harmful outputs generated by language models (H<sub>5</sub>), we combined classification-based content filtering with rule-based content filtering. We use a classification based content filtering system provided by Azure OpenAI ([bit.ly/azure-content-filter](https://bit.ly/azure-content-filter)) which identifies and filters out content related to “hate”, “sexual”, “violence”, and “self-harm” categories. In addition, we developed a rule-based method to filter out any generated content that contained words or phrases related to suicidal ideation or self-harm. To achieve this, we created a list of 50 regular expressions (e.g., to identify phrases like *“feeling suicidal”*, *“want to die”*, and

---

<sup>9</sup>To generate multiple suggestions, we perform top-p sampling (Holtzman et al., 2020) multiple times. We show three suggestions by default, but participants are provided with an option to seek more suggestions if needed.

<sup>10</sup>The reframing suggestions in this example have been generated by our GPT-3 based model.

“harm myself”) based on suicidal risk assessment lexicons such as the one by Gaur et al. (2019). A language model-generated reframe suggestion that matched any of the regular expressions was filtered out and not suggested to the participants. Also, participants were given the option to flag inappropriate reframing suggestions through a “Flag inappropriate” button (Figure 3.2f; Section 3.2.4).

We deployed this system on the MHA platform and studied its effectiveness with platform visitors (see Section 3.2.1). We make the code used to design the system publicly available at [github.com/behavioral-data/Self-Guided-Cognitive-Restructuring](https://github.com/behavioral-data/Self-Guided-Cognitive-Restructuring).

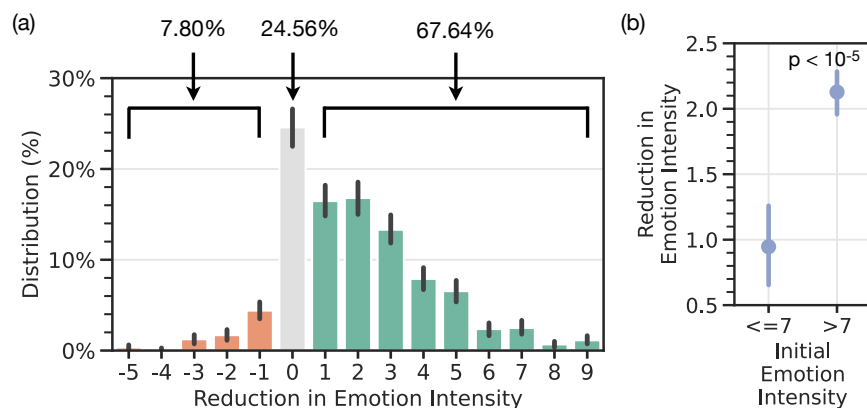
### 3.2.3 RQ2a – To what extent does our intervention help individuals in alleviating negative emotions and overcoming negative thoughts?

We used a mixed-method approach to evaluate the effectiveness of our system. Here, we first describe the different quantitative and qualitative measures used in our study, followed by the evaluation of our system on these metrics.

#### Quantitative Measures

Drawing from various metrics prevalent in the cognitive behavioral therapy literature, we conducted a comprehensive analysis of our system. We assessed the effects of our system on the participants’ emotions, the efficacy of the reframes they wrote, and learnability of the skill:

1. **Reduction in Emotion Intensity:** Intensity of the participant’s emotion before the system use – Intensity of the participant’s emotion after the system use. We collected the emotion associated with the participant’s negative thought before the system use (“*What emotion does this thought make you feel?*”) and collected its intensity both before and after the system use (“*How strong is your emotion? (1 to 10)*”).
2. **Reframe Relatability:** After the system use, we asked the participant: “*How strongly do you agree or disagree with the following statement? – I believe in the reframe I came with*” (1 to 5; 1: Strongly Disagree; 5: Strongly Agree).
3. **Reframe Helpfulness:** After the system use, we asked the participant: “*How strongly do you*



**Figure 3.3:** (a) Reduction in emotion intensity of participants before and after using the system (emotion scale: 1 to 10). We found that 67% of the participants reported having a positive reduction in (negative) emotions (N=1,922). (b) Participants with higher emotion intensity before using the system reported a higher reduction in emotion intensity post the system usage (N=1,922). Error bars represent 95% bootstrapped confidence intervals.

agree or disagree with the following statement? – The reframe *helped* me deal with the thoughts I was struggling with” (1 to 5; 1: Strongly Disagree; 5: Strongly Agree).

4. **Reframe Memorability:** After the system use, we asked the participant: “How strongly do you agree or disagree with the following statement? – I will *remember* this reframe the next time I experience this thought” (1 to 5; 1: Strongly Disagree; 5: Strongly Agree).

5. **Skill Learnability:** After the system use, we asked the participant: “How strongly do you agree or disagree with the following statement? – By doing this activity, I *learned* how I can deal with future negative thoughts” (1 to 5; 1: Strongly Disagree; 5: Strongly Agree).

## Qualitative Measures

We also collected subjective feedback from participants. At the end of the system usage, we asked an optional open-ended question “We would love to know your feedback. What did you like or dislike about the tool? What can we do to improve?”

## Results: Quantitative

**67.64% of the participants reported a positive change in emotions.** We assessed the difference in the intensity of participants’ self-reported negative emotions before and after utilizing our system

Outcome Measure	Mean	Std
Reduction in Emotion Intensity (-10 to 10)	1.90	1.29
Reframe Relatability (1 to 5)	3.84	1.17
Reframe Helpfulness (1 to 5)	3.33	1.35
Reframe Memorability (1 to 5)	3.52	1.36
Skill Learnability (1 to 5)	3.39	1.39

**Table 3.3:** Mean and standard deviation of the five quantitative measures as reported by participants.

Outcome Measures	Initial Emotion Intensity	
	$\leq 7$	$> 7$
Reduction in Emotion Intensity (-10 to 10)	0.95	2.13
Reframe Relatability (1 to 5)	3.98	3.65
Reframe Helpfulness (1 to 5)	3.49	3.09
Reframe Memorability (1 to 5)	3.67	3.33
Skill Learnability (1 to 5)	3.57	3.14

**Table 3.4:** Participants with higher emotion intensity before using the system reported a higher reduction in emotion intensity post the system usage. Participants with higher initial emotion intensities reported worse reframing outcomes, suggesting that writing effective reframes and learning the cognitive restructuring skill was harder when individuals were emotionally agitated (N=1,922).

(N=1,922). Figure 3.3a shows the distribution. Our findings revealed a positive emotional shift in 67.64% (1,300) of the participants, while 24.56% (472) of the participants reported no change in their emotion intensity. A small 7.80% (150) of the participants reported a negative shift in their emotions, with the majority (72%; 108) of them experiencing a relatively minor negative shift of -1.

**Participants with higher initial emotion intensity experienced a greater improvement in emotions.** We checked the effects of our system on participants with different initial emotion intensities. We found that participants with more intense initial emotions ( $>7$  out of 10) reported 124.21% more substantial positive shifts in their emotional state than those with less intense initial emotions (2.13 vs. 0.95;  $\leq 7$  out of 10; N=1,922;  $p < 10^{-511}$ ; Fig 3.3b). Psychotherapy research suggests that a higher intensity of negative moods and depression is associated with stronger negative cognition and maladaptive thoughts (Beevers et al., 2007). Our findings indicate that participants with greater initial emotional intensity could have a greater benefit from a cogni-

<sup>11</sup>We use a Two-sided student's t-test for all statistical tests in this work.

tive restructuring intervention like ours, potentially due to the positive effects it has on their cognition, which in turn may positively effect mood and emotion.

**Majority of participants found the reframes believable, helpful, and memorable.** We evaluated the effectiveness of the reframes that people are able to write using our system (N=1,922). Overall, we found that 80.49% of the participants found the reframes relatable to them, 65.65% of participants found the reframes helpful in overcoming negative thoughts, and 70.49% of participants found the reframes memorable or easy to remember. Further investigating people with different emotional intensities, we found that people with higher initial emotional intensities (> 7 out of 10) reported 8.29% lower reframe relatability (3.65 vs. 3.98;  $p < 10^{-5}$ ), 11.46% lower reframe helpfulness (3.09 vs. 3.49;  $p < 10^{-5}$ ), 9.26% lower reframe memorability (3.33 vs. 3.67;  $p < 10^{-5}$ ), and 12.04% lower skill learnability (3.14 vs. 3.57;  $p < 10^{-5}$ ) than those with lower initial emotional intensities ( $\leq 7$  out of 10; N=1,922; Table 3.4). This suggests that when individuals are emotionally agitated, it is harder to come up with effective reframes and learn cognitive restructuring.

**Most participants found the system helpful in learning cognitive restructuring.** We assessed how effectively our system can be used to learn the skill of managing negative thoughts. We found that 67.38% of the participants reported that the system helped them in learning how to deal with negative thoughts.

We also report mean and standard deviations of the quantitative outcome measures in Table 3.3.

## **Results: Qualitative**

On analyzing the qualitative feedback from participants, we observed that participants highlighted three key ways in which they found assistance from our system ( $H_1$ ).

First, many participants indicated that the system helped them overcome cognitive barriers, especially when they “feel stuck”, and doing this exercise is “difficult”, “on their own” and “in the moment.” A participant wrote, “*My own reframes are difficult, and AI gives multiple other perspectives to consider.*” Also, some participants reported that it helped them find “the right words” or “ideas to start with.” A participant wrote, “*Thank you for helping me to find the right*

*words to clearly reframe a negative thought and how to apply the thought to my own thinking processes.” Another noted, “I appreciated that the option of having the AI tool walk you through the reframing process step by step (e.g., by choosing the negative thought you may be experiencing + giving possible reframing ideas to start with/add more details to).”*

Second, participants expressed how the system enabled a less emotionally triggering experience. One participant wrote, *“I felt in control and more comforted that I can handle difficult situations with confidence.”* Another participant wrote, *“This activity let me calm down...”*. Another participant noted, *“...this made the process much less daunting...”*. This is perhaps consistent with the quantitative findings on reduced emotion intensity (Section 3.2.3).

Third, participants valued that the system allowed them to explore multiple viewpoints. One participant wrote, *“...After reading several reframes and looking over them I realized that there are many options, many positive sides.”* Another participant wrote, *“I felt reassured to see multiple views, and reflect upon them...”*

Overall, these results suggest that there are opportunities to assist participants in cognitively challenging and emotionally triggering psychological processes through human-language model interaction.

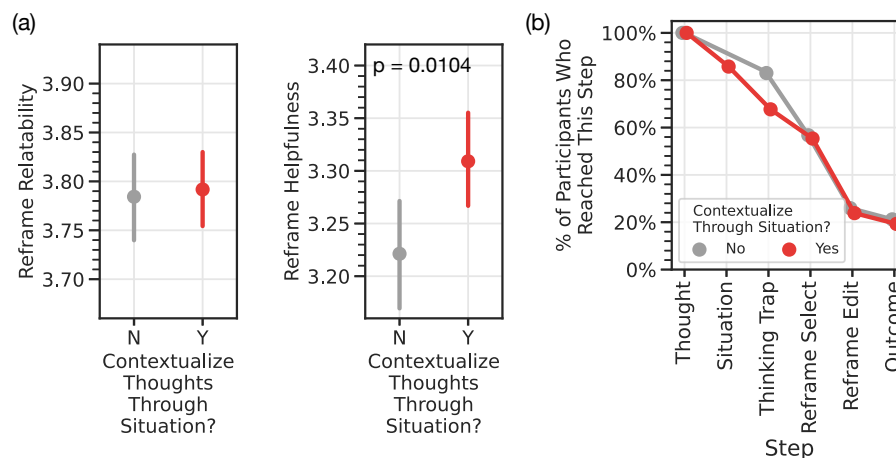
### **3.2.4 RQ2b – What is the impact of individual design hypotheses on the effectiveness of the intervention?**

Here, we studied the impact of individual design hypotheses (Section 3.2.2) on the effectiveness of the intervention and overall participant engagement. To facilitate this, we deployed different design variations of our system by ablating specific design features (e.g., one variation that includes psychoeducation and another variation that removes it). For each design ablation, we conducted randomized controlled trials in which incoming participants were randomly assigned one of the two design variations. To measure the impact, we evaluated the difference in outcomes between participants involved in the two design variations. In this section, we report the results from ablating contextualization (**H<sub>2</sub>**), psychoeducation (**H<sub>3</sub>**), and interactivity (**H<sub>4</sub>**).

## Contextualizing through Situations Improves Reframe Helpfulness

Reflecting on situations and emotions and understanding their connection with negative thoughts can help in writing more personalized and effective reframes ( $H_1$ ). However, asking participants for additional information like descriptions of a relevant situation and emotion can potentially increase dropout which could prevent successful outcomes. Therefore, the reflection process comes with a tradeoff with higher participant burden and higher dropout rates. Here, we conducted two different randomized trials. One where we enabled contextualization through situations to half of the participants at random. Another where we enabled contextualization through emotions to half of the participants at random.

We found that contextualizing participant thoughts through their situations led to 2.80% more helpful reframes and similar levels of reliability (3.31 vs. 3.22;  $N=1,636$ ;  $p = 0.0192$ ; Figure 3.4a). This indicates the benefits of increased reflection in self-guided mental health interventions. Also, while typically an increased information request is correlated with a higher dropout rate, we found that a similar number of participants reached the end of the tool, regardless of the additional information requested (Figure 3.4b).



**Figure 3.4:** Randomized controlled trial to estimate the effects of contextualizing thoughts through situation ( $N=1,636$ ). (a) Contextualizing participant thoughts through their situations led to 2.80% more helpful reframes (3.31 vs. 3.22;  $p = 0.0104$ ) but did not lead to more reliable reframes. (b) Asking for additional context did not lead to a lower completion rate. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .

Surprisingly, participants who contextualized their thoughts through emotions reported 3.86% lower levels of reliability (3.87 vs. 3.72;  $N=4,016$ ;  $p < 0.001$ ; Appendix Figure B.8). This may be

because our language model does not incorporate emotions while identifying thinking traps or suggesting reframes, due to the lack of relevant cognitive restructuring dataset containing self-reported emotion annotations. Consequently, this could possibly lead participants to develop unwarranted expectations where they anticipate their emotional states to be addressed in the reframing suggestions, even when they are not. Note that this is different for descriptions of situations, which the language model does take into account and typically reflects in the generated reframes.

Qualitative feedback from participants indicated that they desired the inclusion of these steps as it helped them better process their thoughts (e.g., a participant wrote *“What made it especially helpful was being able to contextualize my feelings, which I feel allows for a more relatable reframe”*).

### **Integrating Psychoeducation has Limited Impact on Overall Effectiveness**

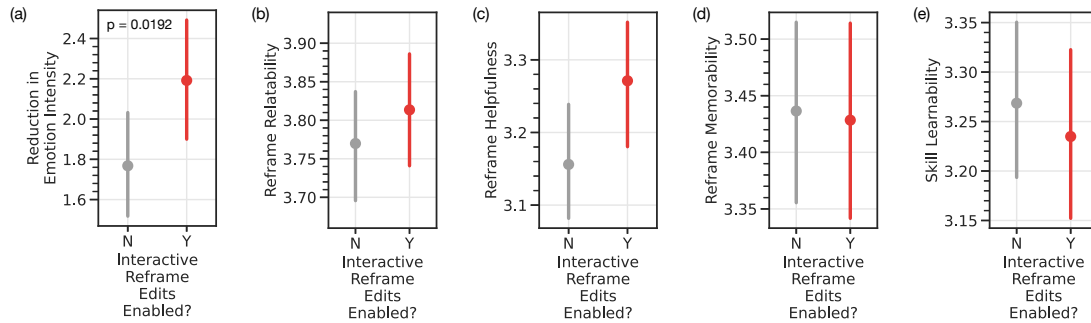
Providing psychoeducation with the intervention may help people learn the cognitive restructuring skill more effectively. While users of early prototypes expressed interest in integrating psychoeducation (**H<sub>3</sub>**; Section 4.1), our randomized trial indicated that it did not lead to significant quantitative improvement in outcomes including skill learnability as self-reported by participants (at  $\alpha = 0.05$ ;  $N=1,850$ ; Appendix Figure B.8).

Nevertheless, qualitative feedback from participants indicated that they found the provided definitions, examples, and strategies helpful (e.g. one participant wrote, *“I like how the tool provided explanations”*; another participant wrote, *“I like the simple explanations and examples for each thought trap”*).

### **Increased Interactivity with the Language Model is Associated with Improved Outcomes**

We provided participants with an option to seek more specific suggestions from the language model, including *“making the reframe more relatable to their situation”*, *“figuring out the next steps and actions”*, and *“feeling more supported and validated”* (**H<sub>4</sub>**; Figure 3.2g; Appendix Figure B.13).

In our randomized trial, where only half of the participants at random were given this option, we found that having this option available led to a 23.73% greater reduction in emotion intensity

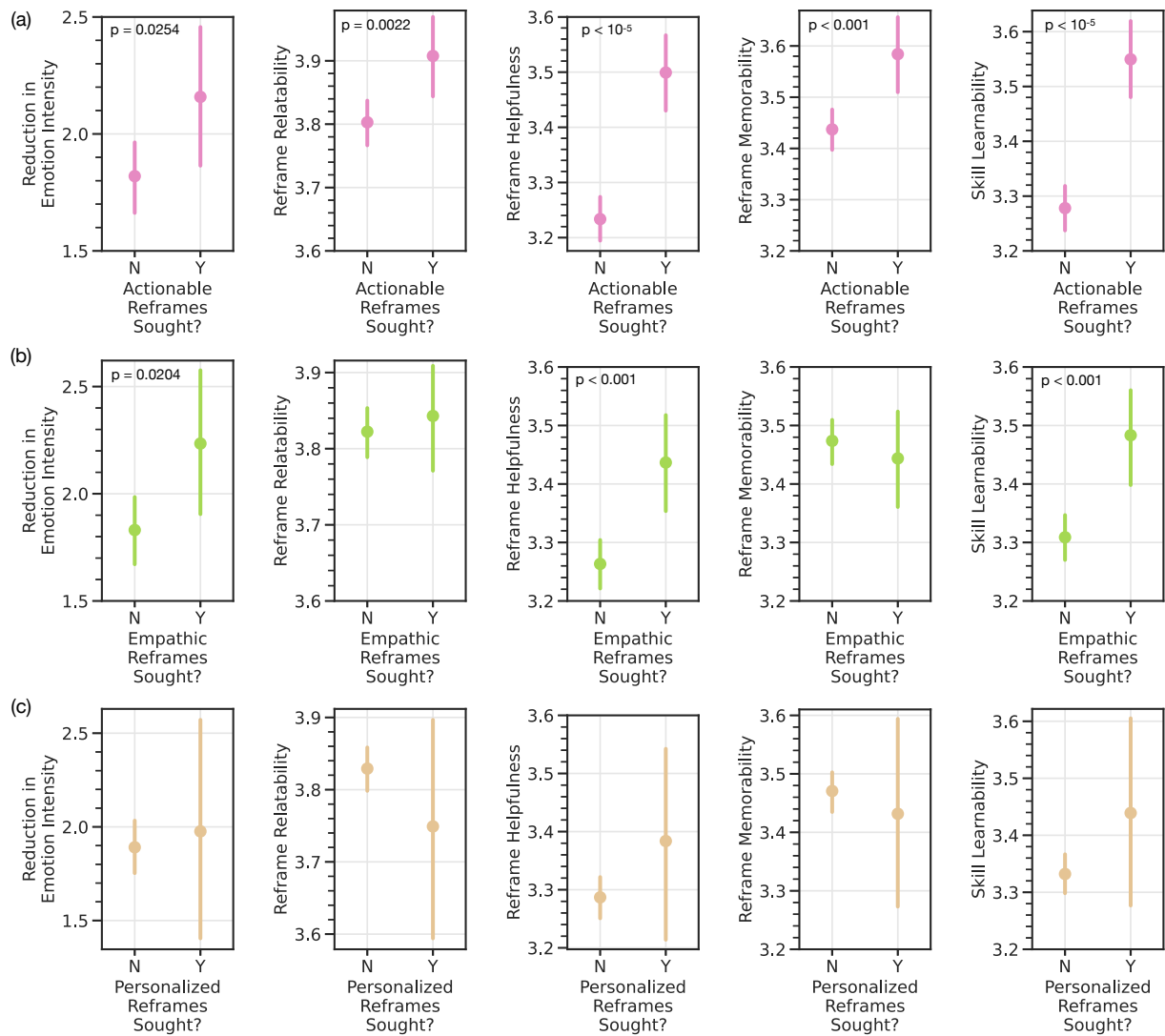


**Figure 3.5:** Randomized controlled trial to estimate the effects of enabling participants to iteratively edit reframes through increased interaction with the language model (N=2,165). 38% of participants chose to use this intervention. (a) Having the option of interactive reframe edits available to participants led to a 23.73% greater reduction in emotion intensity (2.19 vs. 1.77). (b-e) However, it did not lead to significant differences in other outcomes (at  $\alpha = 0.05$ ). Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .

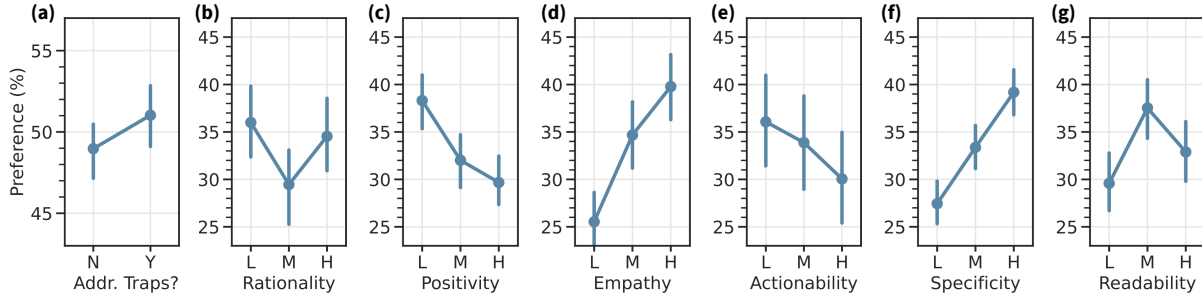
(2.19 vs. 1.77;  $p = 0.0192$ ; N=2,165) and insignificant differences in other outcomes (Figure 3.5). Also, some participants appreciated having this option. One participant wrote, “I’m glad there was an option to get supported and validated”.

Further, among the study participants who were provided this option, we observed that 38.60% of them made use of it. Those who chose to use it to further interact with the language model to seek additional reframing suggestions of specific types (actionable, empathic, or personalized) reported 5.57% higher reframe helpfulness (3.41 vs. 3.23;  $p < 10^{-5}$ ) and 4.86% higher skill learnability (3.45 vs. 3.29;  $p < 0.001$ ) than participants who did not use it (N=992; Appendix Figure B.10). Moreover, those who chose to make their reframes actionable during this step (by choosing the option, “I want to figure out the next steps and actions”) reported significantly superior effectiveness across all five outcomes than those who did not (Figure 3.6a). Prior work has highlighted the importance of behavioral activation which involves engaging in behaviors or actions that may help in overcoming negative thoughts (Dimidjian et al., 2011; Burkhardt et al., 2021). Our work shows that participants explicitly seeking out actionable reframed thoughts are more likely to report better outcomes.

Moreover, those who chose to make their reframe empathic reported a 21.86% higher reduction in emotion intensity (2.23 vs. 1.83;  $p = 0.0204$ ), 5.52% higher reframe helpfulness (3.44 vs. 3.26;  $p < 0.001$ ) and 5.14% higher skill learnability (3.48 vs. 3.31;  $p < 0.001$ ) and no significant differences based on reframe relatability and reframe memorability (at  $\alpha = 0.05$ ; N=992; Fig-



**Figure 3.6:** Participants were provided with an option to seek one or more of the following types of more specific suggestions from the language model – actionable, empathic, or personalized. Among those who used any of the options (N=992), (a) those who chose to make their reframes actionable experienced superior effectiveness across all five outcomes; (b) those who chose to make their reframes empathic reported 21.86% higher reduction in emotion intensity (2.23 vs. 1.83), 5.52% higher reframe helpfulness (3.44 vs. 3.26), and 5.14% higher skill learnability (3.48 vs. 3.31) and no significant differences based on reframe reliability and reframe memorability (at  $\alpha = 0.05$ ); (c) those who chose to make their reframes personalized reported no significant differences in outcomes (at  $\alpha = 0.05$ ). Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .



**Figure 3.7: Which linguistic attributes of reframed thoughts do people prefer?** For a given situation and thought from MHA visitors, we show them multiple LM-generated reframes with variance across a randomly selected attribute (e.g., low, medium, and high actionability). We find that highly empathic and highly specific reframings are more preferred. On the other hand, reframes with high positivity are less preferred. N: No; Y: Yes; L: Low; M: Medium; H: High. Error bars represent 95% bootstrapped confidence intervals.

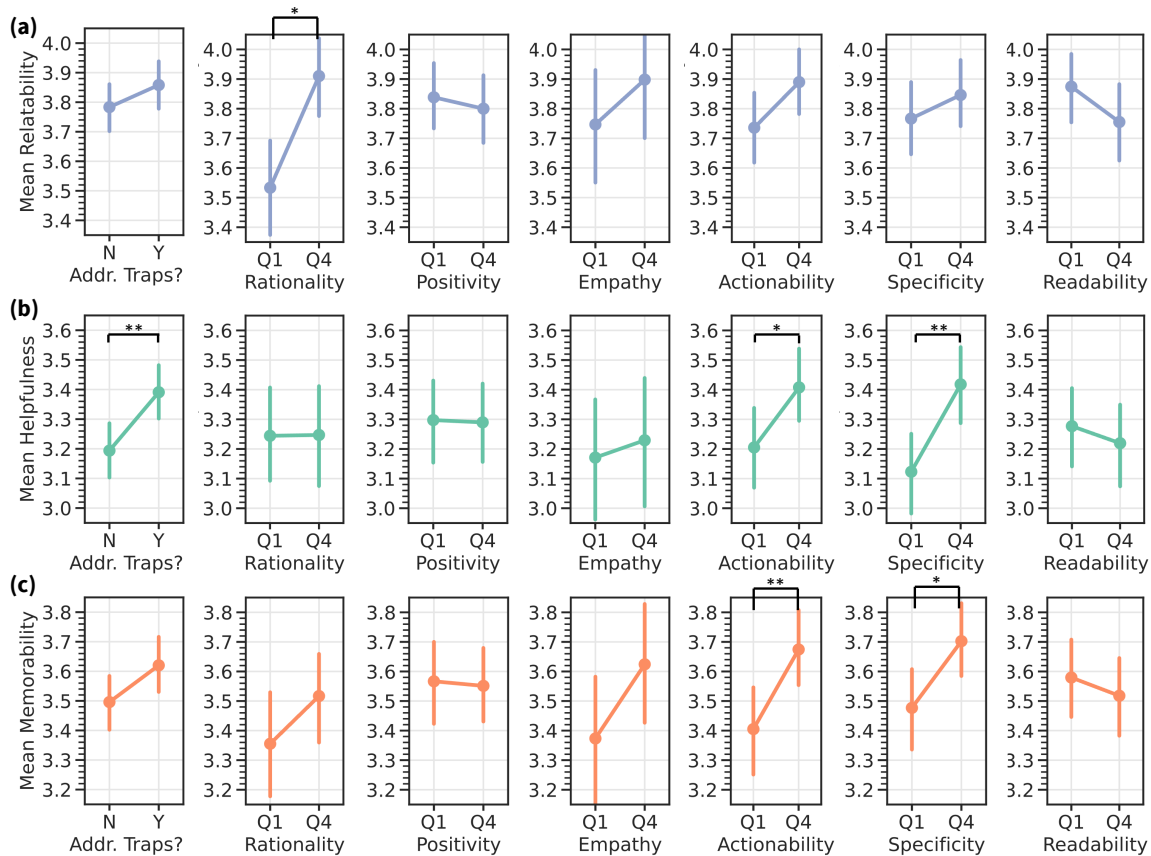
ure 3.6b). We did not find significant differences based on whether a participant chose to make a reframe personalized or not (at  $\alpha = 0.05$ ;  $N=992$ ; Figure 3.6c).

### 3.2.5 RQ3a - What types of reframed thoughts do people prefer?

To understand which reframing attributes people prefer, we suggest multiple LM-generated reframes which vary across our attribute values. Given a situation and thought, we start by generating one reframed thought using our model. Next, we randomly select an attribute (e.g., actionability) and vary the first reframe based on it (e.g., to generate two additional reframes with higher or lower actionability) using our proposed controllable text generation method (Section refsubsec:control-reframe). Figure 3.7 reveals key differences between the linguistic attributes of reframes that people select and prefer:

**(1) Highly empathic and specific reframings are preferred more.** We find that highly empathic reframes are preferred 55.7% more frequently than reframes with lower empathy (39.7% vs. 25.5%;  $p < 10^{-5}$ ); highly specific reframes are preferred 43.1% more frequently than reframes with lower specificity (39.2% vs. 27.4%;  $p < 10^{-5}$ ). Prior work has shown the importance of empathy and less “templated” responses in mental health support conversations (Althoff et al., 2016). Here, we show that empathy and specificity of LM-generated reframes may support people in reframing negative thoughts.

**(2) Overly positive reframes are preferred less.** On the other hand, reframes with high positivity



**Figure 3.8: Which linguistic attributes are associated with desired cognitive reframing outcomes?** For a given situation and thought, we show one LM-generated reframe to MHA participants and ask them to rate it on reliability, helpfulness and memorability on a 1 to 5 scale. For each linguistic attribute, we compare the first (Q1) and the fourth quartile (Q4). We find that **(a)** reframes that have higher rationality are more reliable; **(b)** reframes that address thinking traps, have higher actionability or higher specificity are more helpful; **(c)** reframes that have higher actionability or higher specificity are more memorable. Error bars represent 95% bootstrapped confidence intervals. \*:  $p < 0.05$ ; \*\*:  $p < 0.01$ .

are preferred 22.7% less frequently than reframes with lower positivity (29.6% vs. 38.3%;  $p < 10^{-5}$ ). This may be because adopting an overly positive reframed thought may be challenging for individuals who are already experiencing emotionally triggering negative thoughts (Dember and Penwell, 1980).

Participants also prefer medium-readability reframes over very simple or very complex reframes, perhaps because their language is balanced for a wider audience.

### 3.2.6 RQ3b - How do the linguistic attributes of reframed thoughts relate to the desired outcomes of cognitive reframing?

We assess what characterizes a reframe that is relatable, helpful and memorable. We show a *single* model-generated reframe to the participants and ask them to rate it on a 5-point Likert scale (Likert, 1932) with regards to the three outcome measures (1: Strongly Disagree; 5: Strongly Agree). We do not provide participants in this experiment with a choice of multiple reframes to avoid any selection effects (Section refsubsec:reframing:design:preference). Figure 3.8 offers key insights on which attributes of reframed thoughts are related to different desired outcomes:

**(1) Reframes that are more rational are more *relatable*.** We find that reframes that have higher rationality are 10.8% more relatable than lower rationality reframes (3.91 vs. 3.53;  $p < 0.05$ ). This may be because higher rationality reframes, by definition, are more likely to be based on reasons and are less likely to make unrealistic assumptions, making them easier to relate to.

**(2) Reframes that address thinking traps and are more actionable and specific are more *helpful*.** Reframes that address thinking traps are 6.3% more helpful than reframes that do not address them (3.39 vs. 3.19;  $p < 0.01$ ). Such reframes specifically challenge the cognitive biases in thinking patterns (e.g., “*Fortune Telling*”; Appendix B.3), which has shown to be more effective in dealing with negative thoughts in psychotherapy research (Beck, 1976; Burns, 1980). Moreover, we find that reframes with higher actionability are 6.6% more helpful than lower actionability reframes (3.41 vs. 3.20;  $p < 0.05$ ) and reframes with higher specificity are 9.6% more helpful than lower specificity reframes (3.42 vs. 3.12;  $p < 0.01$ ).

**(3) Reframes that are more actionable and more specific are more *memorable*.** We find that reframes with higher actionability are 7.9% more memorable than lower actionability reframes (3.67 vs. 3.40;  $p < 0.01$ ) and reframes with higher specificity are 6.3% more memorable than lower specificity reframes (3.70 vs. 3.48;  $p < 0.05$ ).

Issues	Reduction in Emotion Intensity	Reframe Relatability	Reframe Helpfulness	Reframe Memorability	Skill Learnability	N
Body Image	1.42	3.89	3.20	3.49	3.38	71
Dating & Marriage	2.05	3.79	3.20	3.47	3.33	328
Family	1.99	3.78	3.26	3.35	3.34	170
Fear	1.63	3.53	3.07	3.26	3.20	123
Friendship	1.91	3.65	3.20	3.48	3.20	159
Habits	1.72	3.98	3.50	3.52	3.57	42
Health	2.36	3.91	3.45	3.77	3.47	53
Hopelessness	1.11	3.41	2.66	3.06	2.84	70
Identity	2.54	4.00	3.55	3.64	3.09	11
Loneliness	1.56	3.43	2.74	3.03	2.77	146
Money	1.71	3.73	2.80	3.43	3.17	30
Parenting	2.06	4.19	3.69	3.97	3.61	36
School	1.94	3.79	3.20	3.34	3.13	181
Tasks & Achievement	1.65	3.56	3.04	3.23	2.99	232
Trauma	1.42	3.33	2.58	3.00	2.50	12
Work	2.31	3.89	3.54	3.77	3.58	258

**Table 3.5:** Effectiveness of our system across different issues expressed by participants. Numbers highlighted in **green** indicate outcomes that are significantly better than the population mean ( $p < 0.05$ ). Numbers highlighted in **red** indicate outcomes that are significantly worse than the population mean ( $p < 0.05$ ). We found that participants who expressed *Parenting* and *Work* related issues reported better outcomes than the population means. Moreover, participants who expressed *Hopelessness*, *Loneliness*, and *Tasks & Achievement* related issues reported worse outcomes.

### 3.2.7 RQ4 – How equitable is the intervention and what strategies may improve its equity?

Next, we assess how equitable our intervention is across the issues expressed by participants (Section 3.2.7) and across participant demographics (Section 3.2.7). Moreover, we work towards improving equity of our system by improving its effectiveness for a specific subpopulation experiencing one of the worst outcome disparities, adolescents (Section 3.2.7).

#### Assessing Outcomes across Issues

To better determine the effectiveness of our intervention in various scenarios, we assessed the outcomes across different types of situations and thoughts that individuals might experience. We characterized participants' situations and thoughts based on the broader issues that they relate

to. In collaboration with mental health experts (some of whom are co-authors), we manually labeled 500 thoughts and situations to identify the potential issues that they are associated with. The result of this iterative open-ended coding process was a set of 16 different issues expressed by participants. These include *Body Image*, *Dating & Marriage*, *Family*, *Fear*, *Friendship*, *Habits*, *Health*, *Hopelessness*, *Identity*, *Loneliness*, *Money*, *Parenting*, *School*, *Tasks & Achievement*, *Trauma*, and *Work*. See Appendix Table B.1 for their definitions and examples.

We used this dataset to finetune a GPT-3 model (`text-davinci`), which achieved an accuracy of 73.00% on a held-out set (random performance 6.25%). We used this model to analyze the outcomes for people experiencing different issues and to identify the issues where our intervention performed better or worse (Table 3.5).

We found that participants expressing *Hopelessness* and *Loneliness* related thoughts reported worse outcomes relative to other issues. Participants with *Hopelessness* (e.g., “*I will never be better*”) reported 41.27% lower reduction in emotion intensity (1.11 vs. 1.89;  $p = 0.0220$ ), 8.33% lower reframe relatability (3.42 vs. 3.72;  $p = 0.0200$ ), 16.61% lower reframe helpfulness (2.66 vs. 3.19;  $p < 0.001$ ), 10.53% lower reframe memorability (3.06 vs. 3.42;  $p = 0.0169$ ), and 12.07% lower skill learnability (2.84 vs. 3.23;  $p = 0.0119$ ) than the population means. Moreover, participants with *Loneliness* (e.g., “*I feel like no one is with me*”) reported 8.45% lower reframe relatability (3.43 vs. 3.72;  $p = 0.0031$ ), 14.11% lower reframe helpfulness (2.74 vs. 3.19;  $p < 0.001$ ), 11.40% lower reframe memorability (3.03 vs. 3.42;  $p < 0.001$ ), and 14.24% lower skill learnability (2.77 vs. 3.23;  $p < 0.001$ ) than the population means. These differences could suggest that thoughts related to some issues are more challenging to overcome than others (as also suggested in psychology theory (Hawkley and Cacioppo, 2010; Heinrich and Gullone, 2006; Beck et al., 1975) ) or that they represent a different subpopulation. However, we also found a lower reduction in emotion intensity (i.e., an outcome measured pre- and post-intervention), suggesting that our system might have had greater difficulty in assisting these issues. In fact, some participants commented that the reframing suggestions did not work well for issues that were too complex and nuanced. One participant wrote, “*It might be too simple for more complicated problems.*” Another participant wrote, “*More complex problems need more precise results in my opinion.*” Some participants thought that the suggestions to such complex problems were “*superficial*”, “*artificial*,” or “*hard to relate to.*” Future iterations of the system could benefit from designing more sophisticated language

modeling solutions for complex issues.

We also observed *Tasks & Achievement* related thoughts (e.g., “I can’t finish my work”) to have 4.30% lower reframe relatability (3.56 vs. 3.72;  $p = 0.0299$ ), 5.56% lower reframe memorability (3.23 vs. 3.42;  $p = 0.0274$ ), and 7.42% lower skill learnability (2.99 vs. 3.23;  $p = 0.0071$ ). Qualitative feedback from participants with such thoughts revealed that they often sought concrete actions beyond what the reframe suggestions could offer.

Moreover, those who used our system for *Parenting* and *Work* reported significantly better outcomes than the population means. Those with *Parenting* issues reported 12.63% higher reframe relatability (4.19 vs. 3.72;  $p = 0.0102$ ), 15.67% higher reframe helpfulness (3.69 vs. 3.19;  $p = 0.0145$ ), and 16.08% higher reframe memorability (3.97 vs. 3.42;  $p = 0.0080$ ). And those with *Work* issues reported 22.22% higher reduction in emotion intensity (2.31 vs. 1.89;  $p = 0.0197$ ), 4.57% higher reframe relatability (3.89 vs. 3.72;  $p = 0.0198$ ), 10.97% higher reframe helpfulness (3.54 vs. 3.19;  $p < 0.001$ ), 10.23% higher reframe memorability (3.77 vs. 3.42;  $p < 0.001$ ), and 10.84% higher skill learnability (3.58 vs. 3.23;  $p < 0.001$ ).

## Assessing Outcomes across Demographics

Language modeling interventions are known to be biased toward people of specific demographics. For interventions targeting mental health, previous research has found that language models are likely to perpetuate social stereotypes, for example, under-emphasizing men’s mental health (Lin et al., 2022). Broadly, this corresponds to the principle of demographic parity in the fairness in machine learning literature (Caliskan et al., 2017; Mehrabi et al., 2021; Blodgett et al., 2020).

Here, we studied the difference in outcomes of our intervention across participants of different demographics. We asked participants to optionally provide demographic information, including age (ranges between 13 to 65+), gender (Female, Male, or Non-Binary), race/ethnicity (American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Middle Eastern or North African, Native Hawaiian and Pacific Islander, White, More than One, or Other), and education levels (Middle School, High School, Undergraduate, Graduate, or Doctorate).

Table 3.6 reports the outcome differences. We found that participants aged 17 or younger

Demographics	Reduction in Emotion Intensity	Reframe Relatability	Reframe Helpfulness	Reframe Memorability	Skill Learnability	N
<b>Age</b>						
13–14	1.84	3.64	2.94	3.03	2.99	146
15–17	1.64	3.50	2.68	3.07	2.83	149
18–24	1.98	3.78	3.13	3.48	3.15	247
25–34	2.00	3.89	3.32	3.70	3.40	179
35–44	2.15	3.97	3.40	3.69	3.51	109
45–54	1.64	3.96	3.32	3.64	3.56	71
55–64	1.88	3.96	3.46	3.96	3.21	32
65+	1.20	4.00	3.38	3.88	4.13	8
<b>Gender</b>						
Female	1.92	3.78	3.21	3.44	3.26	646
Male	2.19	3.74	2.95	3.38	3.04	258
Non-Binary	1.94	3.76	3.30	3.46	3.22	54
<b>Race/Ethnicity</b>						
AIAN	2.17	2.50	2.67	3.17	2.67	6
Asian	1.91	3.79	3.08	3.43	3.12	216
Black / African Am.	2.43	3.85	3.30	3.62	3.49	47
Hispanic or Latino	1.97	3.91	3.43	3.59	3.47	76
MENA	1.90	3.78	2.90	2.98	2.80	50
NHPI	2.00	4.40	4.00	3.60	3.20	5
White	2.05	3.73	3.12	3.47	3.22	438
More than One	2.83	3.84	3.29	3.24	3.16	38
Other	0.78	3.75	3.06	3.29	3.02	48
<b>Education</b>						
Middle School	1.80	3.58	2.89	2.96	2.79	120
High School	1.80	3.65	2.97	3.31	3.09	313
Undergraduate	2.04	3.79	3.19	3.54	3.28	239
Graduate	2.30	3.98	3.47	3.69	3.52	211
Doctorate	1.46	4.21	2.96	3.93	3.07	28

**Table 3.6:** Effectiveness of our system across different demographic population. Numbers highlighted in green indicate outcomes that are significantly better than the population mean ( $p < 0.05$ ). Numbers highlighted in red indicate outcomes that are significantly worse than the population mean ( $p < 0.05$ ). AIAN: American Indian or Alaska Native; MENA: Middle Eastern or North African; NHPI: Native Hawaiian and Pacific Islander. We found that adolescents, males, and those with middle school education reported worse outcomes. Moreover, adults (age  $\geq 25$ ) and those with graduate and doctorate education reported better outcomes.

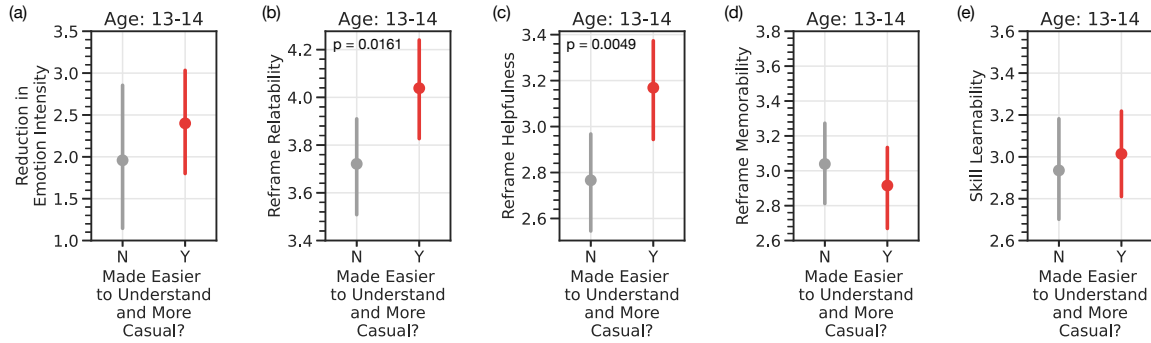
reported 4.84% lower reframe relatability (3.54 vs. 3.72;  $p = 0.0091$ ), 10.03% lower reframe helpfulness (2.87 vs. 3.19;  $p < 0.001$ ), 10.23% lower reframe memorability (3.07 vs. 3.42;  $p < 0.001$ ), and 10.22% lower skill learnability (2.90 vs. 3.23;  $p < 0.001$ ) compared to the population mean. On the other hand, those aged 25 or above reported significantly better outcomes overall. This suggests that our intervention is less effective for adolescents and more effective for adults. Section 3.2.7 explores improving the effectiveness of our intervention for adolescents.

Moreover, we found that male participants reported 7.52% lower reframe helpfulness (2.95 vs. 3.19;  $p = 0.0162$ ), and 5.88% lower skill learnability (3.04 vs. 3.23;  $p = 0.0364$ ) than the population mean. This is consistent with prior work that shows that language models are likely to be disparate toward men's mental health (Lin et al., 2022). Race or ethnicity of the participants was not consistently associated with better or worse outcomes. However, those who identified their race as "Other" reported a 58.73% lower reduction in emotion intensity than the population mean (0.78 vs. 1.89;  $p = 0.0058$ ). Moreover, those who identified their race as "Middle Eastern or North African" reported 12.87% lower reframe memorability (2.98 vs. 3.42;  $p = 0.0180$ ) and 13.31% lower skill learnability (2.80 vs. 3.23;  $p = 0.0207$ ).

Finally, based on education levels, participants with a "Middle School" education reported 9.40% lower reframe helpfulness (2.89 vs. 3.19;  $p = 0.0249$ ), 13.45% lower reframe memorability (2.96 vs. 3.42;  $p < 0.001$ ), and 13.62% lower skill learnability (2.79 vs. 3.23;  $p < 0.001$ ). On the other hand, those who identified as "Graduate" reported 6.99% higher reframe relatability (3.98 vs. 3.72;  $p = 0.0020$ ), 8.78% higher reframe helpfulness (3.47 vs. 3.19;  $p < 0.001$ ), 7.89% higher reframe memorability (3.69 vs. 3.42;  $p = 0.0019$ ), and 8.98% higher skill learnability (3.52 vs. 3.23;  $p = 0.0017$ ). Note that age and education are strongly correlated (pearson's correlation = 0.62), especially for younger participants that did not yet have the time to advance their education, suggesting that the relationship between education and outcomes may be at least partially explained by age.

### **Improving Intervention Equity by Improving the Experience of Adolescents**

Because intervention effectiveness differs significantly across people's issues and demographics, it is crucial to identify solutions that improve intervention equity. This may require adapting the



**Figure 3.9:** Randomized controlled trial to estimate the effects of making reframes easier to understand and more casual on adolescents in age group 13 to 14 (N=148). Adolescents reported 8.60% higher reframe relatability (4.04 vs. 3.72) and 14.44% higher reframe helpfulness (3.17 vs. 2.77) if they were suggested easier to understand and more casual reframes compared to instances where such reframes were not suggested. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .

intervention to different subpopulations.

Here, we performed one specific experiment to study how our language model-based intervention may be adapted to make it more equitable. We particularly focused on teenagers and adolescents whom we found to have one of the largest outcome discrepancy.<sup>12</sup>

Research suggests that current treatment methods are often structurally incompatible with the ways adolescents engage with, or wish to engage with, mental health care (Kruzan et al., 2022a). Although adolescents are more likely to use self-guided mental health interventions (Schleider et al., 2020), our analysis suggests that our human-language modeling based intervention may be less effective for this demographic (Section 3.2.7). Given the escalating youth mental health crisis (Avenevoli et al., 2015), it is important to develop solutions that bridge this gap. To achieve this, we tried and identified the challenges that uniquely affect adolescents.

We hypothesized that the linguistic complexity of our system may affect its performance among adolescents. Research in sociolinguistics has shown that language use varies with age (Barbieri, 2008). On analyzing the reading complexity of the reframed thoughts authored by participants of different age groups, we found that those between the ages of 13 and 17 tend to write thoughts and reframes with the lowest levels of reading complexity (based on the Coleman–Liau Index (Coleman and Liau, 1975); Appendix Figure B.11). Therefore, we tried to reduce the

<sup>12</sup>While we also observed a large outcome discrepancy across educational attainment, this can largely be explained through age (as a 15 year old almost certainly did not yet have a chance to complete a college education yet).

reading complexity of the reframing suggestions to adolescents. For this, given a reframing suggestion, we asked the GPT-3 language model (Brown et al., 2020) to make it easier to understand and more casual (using the prompt, “*Revise the following text to make it easy to understand for a 5th grader. Also, make it more casual: {reframe}*”), similar to other efforts targeting people of different subpopulations (e.g., for scientific communication (August et al., 2022)). Also, see Appendix Table B.2 for examples that illustrate this rewriting process.

Figure 3.9 reports the results of a randomized trial that only provides these easier to understand and more casual reframing suggestions to half of the participants at random. We found that adolescents in the age group 13 to 14 reported 8.60% higher reframe relatability (4.04 vs. 3.72;  $p = 0.0161$ ) and 14.44% higher reframe helpfulness (3.17 vs. 2.77;  $p = 0.0049$ ) when they were suggested reframes with lower reading complexity (N=148). Moreover, adolescents in the age group 15 to 17 reported 15.58% higher reframe helpfulness (3.19 vs. 2.76;  $p = 0.0042$ ) when they were suggested such reframes (N=174). We did not find significant differences for adults ( $\geq 18$ ) through this intervention (N=760; Appendix Figure B.12). This suggests that a simpler and more casual language might be beneficial to many. However, based on qualitative feedback, certain adult participants expressed a preference for a less casual language. Future work could explore how to accommodate such individual preferences.

### 3.2.8 Discussion

#### Supporting the Learning and Practice of Self-Guided Interventions

Our work demonstrates how language modeling interventions can support mental health. Approximately 20% of people worldwide are experiencing mental health problems, but less than half receive any treatment (WHO, 2022; Olfson, 2016). Due to widespread clinician shortages, lengthy waiting lists, and lack of insurance coverage, many vulnerable individuals have limited access to therapy and counseling. In addition, mental health issues are heavily stigmatized, which frequently prevents individuals from seeking appropriate care (Sickel et al., 2014).

Effective self-guided mental health interventions could rapidly increase access to care (Schleider et al., 2020; Patel et al., 2020; Schleider et al., 2022; Shkel et al., 2023). However, despite their

inherent promise, the wide-scale implementation of these interventions remains a challenge owing to the cognitive and emotional challenges that they pose (Shkel et al., 2023; Garrido et al., 2019). Most interventions that digitally facilitate self-guided interventions simply transform traditional manual therapeutic worksheets into digital online formats (Shkel et al., 2023). These provide limited instructions and support, which affects user engagement and usage (Garrido et al., 2019; Baumel et al., 2019; Fleming et al., 2018; Torous et al., 2020). Other studies have used wizard-of-oz methods to assist users (Ly et al., 2017; Smith et al., 2021; Morris et al., 2015; Kornfield et al., 2023; Kumar et al., 2023). However, the controlled research setting of these studies limit their ecological validity, thereby limiting our understanding of user preferences when systems are deployed in real world (Mohr et al., 2017; Blandford et al., 2018; Borghouts et al., 2021; Poole, 2013).

Here, we contribute the design of a novel system for human-language model interaction-based self-guided cognitive restructuring of negative thoughts. We conduct a large-scale, randomized, empirical studies in an ecologically informed setting to understand how people with lived experience of mental health interact with it.

Our findings open up opportunities for improved learning and practicing of key mental health strategies and coping skills. Moreover, these interventions could complement traditional treatment options, e.g., by being accessible to users when they have difficulties finding a therapist, or in between sessions.

## Implications on the Design of Self-Guided Mental Health Intervention

Several of our design hypotheses (Section 3.2.2) were observed to improve intervention outcomes. These include personalizing the intervention to the participant, facilitating iterative interactivity with the language model, and pursuing equity, all of which may generalize to support other self-guided mental health interventions.

**Personalization.** Effectively supporting humans through self-guided interventions necessitates personalization (Kornfield et al., 2022). Our design incorporated personalization of reframes by not only seeking more information from the participants in the form of their situations, but also integrating it into the suggestions generated by the language model. We found that this form of

personalization led to more helpful reframes than an intervention without it (Section 3.2.4). This was potentially beneficial in generating suggestions that were more realistic and made fewer assumptions about the participant. Moreover, it emphasized the benefits of increased self-reflection by participants, particularly when thinking about the situation associated with negative thoughts.

On the other hand, when we solicited *emotions* from the participants and failed to incorporate them into the generated suggestions, we observed a significant *decrease* in helpfulness (Section 3.2.4). This can likely be attributed to unrealistic expectations set up by our intervention, where participants might presume that their emotional states will be addressed in the reframing suggestions, even when that is not the case.

This shows the significance of personalizing language modeling suggestions when developing such self-guided interventions. Note that we did not find significant improvements based on whether participants explicitly sought specific suggestions to make a reframe personalized (Section 3.2.4).

**Interactivity.** In our intervention, participants who interacted more actively with the language model (by seeking additional suggestions) achieved better outcomes (Section 3.2.4). This highlights the significance of designing interventions that enable a better interaction between the participant and the language model.

Moreover, qualitative feedback from participants revealed that many of them anticipated a continuous, back-and-forth interaction with the language model. Some participants desired a more in-depth exploration of thoughts. One participant wrote, “*I like it but it needs to go deeper with the thoughts.*” Our design facilitates the iterative refinement of reframes. However, a more in-depth exploration could potentially involve addressing new thoughts that arise during the writing of a reframe for the original thought, or even tackling multiple related thoughts at once. Future iterations of the system could work towards a design that is capable of processing multiple thoughts in parallel.

Some participants even expected a chat-like interaction, probably influenced by their experiences with popular systems like OpenAI’s ChatGPT ([chat.openai.com](https://chat.openai.com)). While such mechanisms may offer greater interactivity, their open-ended and uncontrollable nature creates challenges in making them conform to well-established therapeutic processes like cognitive restructuring

(Stade et al., 2024; De Choudhury et al., 2023; Li et al., 2020; Tate et al., 2023). In evidence-based cognitive behavioral therapy, cognitive restructuring is typically exercised in a very structured manner (e.g., by asking a specific set of questions in a sequence). Our work shows the promise of extending such exercises with language models with similar digital interfaces.

While it is possible to replicate these steps within a pure chatbot paradigm, the nature of chat-based interactions can become more complex and may differ from the current methods employed in these interventions. For example, reviewing multiple thinking traps or integration of expandable psychoeducation content is more complex in a purely chat-based interface. Due to the similarity of our intervention with well-established therapy exercises and worksheets, our interface is likely to be a lower burden than a new chat interface. Reducing participant burden has often been associated with improved engagement and completion outcomes (Garrido et al., 2019; Baumel et al., 2019; Fleming et al., 2018; Torous et al., 2020). In particular, this has been observed many times in our studied population (MHA platform visitors with lived experience in mental health that are not driven by study compensation incentives). Therefore, appropriate care is required when designing interventions that offer the right trade-off between interactivity and principled adherence to theory.

Another key aspect related to interactivity is over-reliance on the language model assistance. Taking away any “productive struggle” and doing the restructuring *for* the user without fostering reflection and independent practice is likely unproductive. This relates to “desirable difficulty” that differentiates true, long-term “learning” of a skill from the short-term “performance” during skill acquisition (Guadagnoli and Lee, 2004; Bjork, 2017; Bjork and Bjork, 2020). A longer-term goal would be to modulate the difficulty of the self-guided intervention relative to the skill level of the user of the intervention such that the user can build their skills optimally, which forms an interesting direction of future research.

To assess if users are being over-reliant on a system like ours or being able to learn the skill through it, one could observe if users progressively apply the skill they are being taught in their daily lives. This would involve asking users whether they caught themselves thinking negatively, recognized the negative thinking patterns, and reframed the thought in-the-moment while they were having the thought. There exist standardized measures such as the “Competencies of Cognitive Therapy Scale” that operationalize this type of assessment (Strunk et al., 2014). While ours

was a short single-use intervention, this kind of assessment requires a more longer term study, which may require different incentives, recruiting, and platforms and was therefore outside the scope of this work.

**Equity.** We found that our intervention was less effective for adolescents, males, and individuals with lower levels of education (Section 3.2.7). This is consistent with prior research, which has revealed a bias in language models used in mental health contexts toward similar demographics (Lin et al., 2022). These findings highlight the importance of adapting self-guided mental health interventions utilizing AI models to suit the needs of different demographics and key subpopulations.

Our work proposes a rewriting-based method to achieve this goal (Section 3.2.7). This could involve identifying the specific challenges associated with intervening in certain populations (e.g., reading complexity for adolescents) and then, designing appropriate solutions to address those challenges (e.g., lowering the reading complexity).

## **Limitations**

Our evaluation was confined to a population from a single platform. Yet, our sample size was relatively large and diverse. However, only a few participants over the age of 65 and identifying as American Indian or Alaska Native and Native Hawaiian or Pacific Islander (Table 3.6). While outcomes varied especially across age rather than race/ethnicity, more work is needed to identify and evaluate opportunities for culturally responsive interventions. The focus of this study was limited to short-term outcomes, necessitating further research to evaluate the long-term effects on participants. Still, our study contributes multiple large-scale randomized trials to inform the design and efficacy of digital mental health interventions. Our outcomes rely significantly on the quality of language model-generated thinking traps and reframes and therefore could change as language models improve. However, the algorithms and models used for generation in this work represent the current state-of-the-art.

### 3.3 Related Work

Our work builds upon previous research on digital mental health interventions, AI for mental health, and the design of human-AI collaboration and text reframing systems.

#### Digital Mental Health Interventions

The critical gap between the overwhelming need for and limited access to mental health care has prompted clinicians, technologists, and advocates to develop digital interventions that provide accessible care for all. Several efforts have concentrated on facilitating digital, text-based supportive conversations, through peer-to-peer support networks, such as TalkLife ([talklife.com](http://talklife.com)) and Supportiv ([supportiv.com](http://supportiv.com)), as well as through on-demand talk therapy platforms like Talkspace ([talkspace.com](http://talkspace.com)), BetterHelp ([betterhelp.com](http://betterhelp.com)), and SanVello ([sanvello.com](http://sanvello.com)). Researchers have conducted randomized controlled trials to study the efficacy of these interventions compared to traditional methods of care, such as in-person counseling and worksheet-based skill practices (Hull and Mahan, 2017; Song et al., 2023; Moberg et al., 2019).

Another key focus in this pursuit has been the development of self-guided mental health interventions (Schleider et al., 2020; Patel et al., 2020; Schleider et al., 2022; Shkel et al., 2023). These interventions are designed in various forms, such as “*Do-It-Yourself*” apps to improve mental health “in-the-moment” of crisis, self-help tools for learning and practicing therapeutic skills, and more. Popular examples include self-guided meditation such as Headspace ([headspace.com](http://headspace.com)) or Calm ([calm.com](http://calm.com)). Researchers have also explored the design of apps to track mood changes (Schueller et al., 2021), emotion regulation (Smith et al., 2022b) and to combat loneliness (Boucher et al., 2021). Kruzan et al. (2022b) studied the process of online self-screening of mental illnesses and its role in help-seeking. Howe et al. (2022) designed and evaluated a workplace stress-reduction intervention system and found that high-effort interventions reduced the most stress. Also, Cognitive Behavioral Therapy (CBT) is an evidence-based, well-established psychological treatment (Beck, 1976). Researchers have designed digital self-guided interventions that streamline elements of CBT like cognitive restructuring of negative thoughts by transforming traditional manual worksheets into digital online formats (Rennick-Egglestone et al., 2016; Shkel et al., 2023). Other work has focused on digital mental health interventions based on Dialectical Behavioral

Therapy (DBT) (Schroeder et al., 2018, 2020).

Our work extends this literature by investigating how to design and evaluate self-guided digital mental health interventions that leverage human-AI collaboration. We take CBT-based cognitive restructuring as a case study and design and evaluate a system that assists individuals in restructuring of negative thoughts through language models in a large-scale randomized field study.

## **AI for Mental Health**

Our work is related to the growing field of research in Artificial Intelligence (AI) and Natural Language Processing (NLP) for mental health and wellbeing.

Several human-AI collaboration tools for mental health assist support providers, e.g., clinicians (Tanana et al., 2019; Shen et al., 2020) or peer supporters as in the case of the work presented in this thesis (Chapter 2). This work provides insights on how human-AI collaboration may directly support people struggling with mental health challenges through cognitive restructuring. Computational work on cognitive restructuring has relied on small-scale crowdsourcing studies (Smith et al., 2021; Morris et al., 2015). Our work develops scalable methods for cognitive restructuring and conducts a randomized field study on a large mental health platform.

In their human-centered study, Kornfield et al. (2022) sought to understand the adoption of automated text messaging tools. The study revealed that the participants were interested in making the tools more personalized, favored varying levels of engagement, and wanted to explore a broad range of concepts and experiences. The design of our self-guided intervention builds on these findings aiming to facilitate intervention personalization through situations and emotions of participants, iterative engagement with the language model, and to improve the equity of the intervention across different participant issues and demographics.

Prior work has also studied mental health bias in language models. Lin et al. (2022) investigated gendered mental health stigma present in masked language models and showed that models captured social stereotypes, such as the perception that men are less likely to seek treatment for mental illnesses. We study the disparities of our intervention's effectiveness among individuals from diverse demographics and facing different issues. We also propose a way of

improving the interventions for a key subpopulation.

More broadly, our work relates to the growing body of research in NLP for mental health and psychological well-being (Althoff et al., 2016; Sharma and De Choudhury, 2018; Gaur et al., 2019; Lee et al., 2019; Miner et al., 2019; Pendse et al., 2020; Pérez-Rosas et al., 2019; Pruksachatkun et al., 2019; Yang et al., 2019; Zhang et al., 2019a; Jaidka et al., 2020; Saha and Sharma, 2020; Wadden et al., 2021; Welch et al., 2020; Zhang and Danescu-Niculescu-Mizil, 2020; Lahnala et al., 2021; Naseem et al., 2022; Pérez-Rosas et al., 2022; Shah et al., 2022; Shen et al., 2022; Stewart et al., 2023).

## **Design of Human-AI Collaboration and Text Reframing Systems**

Broadly, our work relates to the design of human-AI collaboration systems that facilitate an interactive setting in which humans can effectively engage with language models to accomplish real-world tasks (Amershi et al., 2019; Lee et al., 2022). Examples include systems for creative writing (Clark et al., 2018), programming (e.g., CoPilot ([github.com/features/copilot](https://github.com/features/copilot))), and brainstorming ideas (e.g., Jasper ([jasper.ai](https://jasper.ai))). Our work studies how such human-AI collaboration systems can support self-guided mental health interventions.

Prior text reframing research has developed methods for related tasks including style, sentiment, and politeness transfer (Reif et al., 2022; Madaan et al., 2020) as well as positive reframing (Ziems et al., 2022). Our work develops text-reframing methods for cognitive restructuring and demonstrates that linguistic attributes of addressing thinking traps, rationality, actionability, specificity, and readability are critical to high-quality reframes.

### **3.4 Summary of Contribution to Thesis**

In this chapter, we designed a human-AI collaboration system that leverages language models to support people through various steps of cognitive restructuring. We defined a framework of seven linguistic attributes of cognitive restructuring, developed automatic metrics to measure these attributes, and validated their measurements with mental health experts. We collected and shared a dataset of 600 situations, thoughts, and reframes from mental health experts and used it to train a retrieval-enhanced in-context learning model based on GPT-3. Through a series of field

studies and randomized trials on a large mental health website, we evaluated this system with 15,531 participants. Our findings demonstrated the effectiveness of this system in helping people reduce the intensity of their negative emotions and effectively reframing negative thoughts. Also, we proposed and validated various design hypotheses including contextualizing people's thoughts through their situations and facilitating iterative interaction with the language model. We found that people struggling with negative thoughts prefer reframes that are highly empathic or specific, but do not prefer reframes that are highly positive. Moreover, we assessed the equity of our system across people with different issues and people of different demographics, and improved equity by demonstrably improving the experience of adolescents through lowering the reading complexity of the language model suggestions.

## Chapter 4

# Evaluating AI-based Mental Health Interventions

In the previous two chapters, we have demonstrated the potential of human-AI collaboration systems to empower people seeking and providing mental health support through in-the-moment feedback. Next, we work towards developing methods to evaluate these systems.

As shown in Chapters 2 and 3, human-AI collaboration techniques have the potential to augment and empower both support providers and support seekers. AI-based techniques are also being developed to train therapists (Tanana et al., 2019), assess their adherence to treatment (Flemotomos et al., 2022), and streamline administrative tasks such as note-taking (Krishna et al., 2021). Traditionally, such processes can be slow, expensive, labor-intensive, and require extensive expert instruction (Tanana et al., 2019; Moyers et al., 2005). By training AI systems on thousands of therapy conversations, they can be used to simulate virtual clients with a wide range of psychological profiles, allowing therapists to practice their skills in a safe and controlled environment. Moreover, such AI systems may provide feedback on key therapy skills and help assess the quality of care. This ultimately could enhance the efficiency and effectiveness of the way in which therapy is delivered.

Simultaneously, AI-based self-guided interventions for mental health are increasingly becoming available, offering innovative tools and techniques for individuals seeking support. Specialized therapy chatbots like Woebot ([woebothealth.com](https://woebothealth.com)), Wysa ([wysa.com](https://wysa.com)), and Sonia ([sonia-](https://sonia-ai.com)

health.com) are being developed to provide personalized support at scale. Interestingly, many individuals are also using general purpose LLMs like ChatGPT for therapeutic purposes, even when not explicitly designed for it, sharing anecdotal LLM examples that resemble therapist behaviors (Al-Sibai, 2023; Reddit-1, 2023; Reddit-2, 2023; Twitter, 2023).

However, mental health experts have cautioned about the potential risks of using AI-based systems in mental health contexts. Both general-purpose and mental health-specific AI systems have been scrutinized for generating biased, harmful, or factually incorrect information (Blodgett et al., 2020; Lin et al., 2022; Ganguli et al., 2022), potentially undermining their effectiveness in supporting both therapists and clients. Moreover, exposing vulnerable individuals to untested interventions, such as under-development chatbots, raises potential risks and ethical concerns. Therefore, evaluating and addressing the potential risks and ethical concerns is critical to ensure the safe and effective integration of these LLMs into existing and new mental health interventions. To achieve this, experts have called for developing mechanisms for continuous monitoring and evaluation of AI systems used within a high-stakes domain like mental health to ensure responsible usage and adherence to established interventions (Stade et al., 2024; De Choudhury et al., 2023; Li et al., 2020; Tate et al., 2023; Van Veen et al., 2024).

Currently, there is a significant lack of research and evidence demonstrating the effectiveness, safety, and feasibility of AI systems in providing mental health interventions. Established interventions like Cognitive Behavioral Therapy (CBT) and Motivational Interviewing (MI), even in computer-assisted formats, are widely recognized as evidence-based practices for a variety of mental health conditions due to extensive research and evaluation (Beck, 1976; Miller and Rollnick, 2012; Wright et al., 2019). There are well-defined guidelines on how clinicians should implement them (Beck, 1976; Miller and Rollnick, 2012). While high-quality AI-delivered CBT or MI *could* potentially be effective, we currently lack methods to evaluate this technology. There is a risk of AI exhibiting clinically contraindicated behaviors (e.g., offering solutions before reflecting on client's feelings and experiences (Cochran and Cochran, 2015)). Without systematic large-scale evaluation frameworks, we cannot determine the presence or frequency of such undesirable behaviors. Furthermore, the circumstances under which LLMs might prioritize certain behaviors like emotional reflecting on client emotions or problem-solving, and whether it aligns with high-quality evidence-based practices like CBT or MI. Understanding the behavior of LLMs

is particularly important in high-stakes settings such as mental health where undesirable behavior could have severe consequences on vulnerable clients. Moreover, identifying the desirable and undesirable behaviors can help identify the potential benefits and risks, which can inform adoption, iterative development, and continuous monitoring.

Here, we propose two different approaches to evaluate AI-based interventions for mental health. First, we demonstrate how clinical trials can be conducted to effectively evaluate short-term and long-term outcomes, equity, and safety of AI-based mental health interventions. We evaluate the AI-based cognitive restructuring intervention developed in Chapter 3, comparing it against traditional interventions including an equivalent non-AI tool and psychoeducation only in an ecologically valid setting. In collaboration with mental health experts, we establish the relevant outcome measures and develop a guiding framework for the essential ethical and safety considerations. We present the conceptual model that outlines this clinical trial, followed by the results demonstrating key similarities and differences between the various intervention groups (Section 4.1).

Next, we conduct a proof-of-concept study proposing a framework for evaluating LLMs in mental health support. We develop BOLT, a novel, general computational framework to systematically study the **Behavior Of LLMs** in the context of their application in Therapy. We quantitatively measure LLM behavior across 13 psychotherapeutic approaches with in-context learning methods. Then, we compare the behavior of LLMs against high- and low-quality human therapy. Our analysis based on Motivational Interviewing therapy reveals that LLMs often resemble behaviors more commonly exhibited in low-quality therapy rather than high-quality therapy, such as offering a higher degree of problem-solving advice when clients share emotions. However, unlike low-quality therapy, LLMs reflect significantly more upon clients' needs and strengths (Section 4.2).

## 4.1 A Clinical Trial of AI-Based Mental Health Intervention

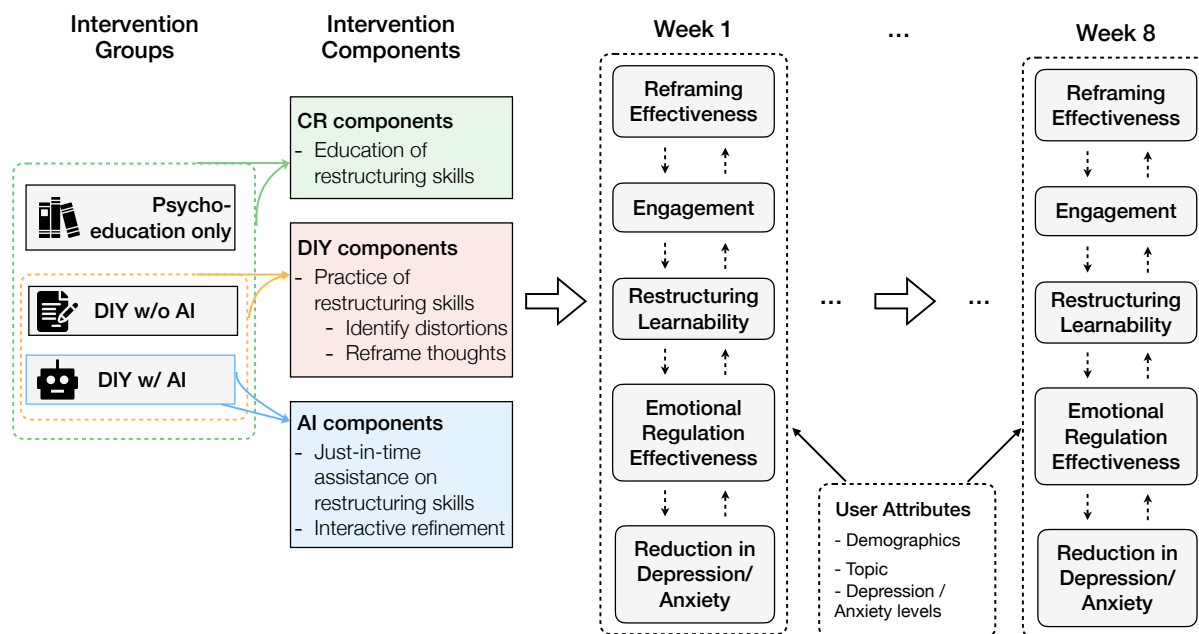
Clinical trials involving human participants are the gold standard for assessing the efficacy and safety of new health interventions. Various types of psychotherapy, such as Cognitive Behavioral Therapy (CBT) (Beck, 1976), Dialectical Behavior Therapy (DBT) (Linehan, 2014), or Motivational

Interviewing (MI) (Miller and Rollnick, 2012), have been evaluated in clinical trials to determine their efficacy in addressing specific mental health concerns like depression, anxiety, or PTSD (Arch et al., 2012; Silverman et al., 1999; Linehan et al., 2015; Burke et al., 2003). As new AI-based interventions are developed for mental health, such clinical trials become relevant for expanding treatment options, advancing relevant science, and improving patient care.

While AI-based mental health interventions are increasingly being developed, there is limited research directly comparing them with similar non-AI interventions through clinical trials. We do not have a clear understanding of how AI-based interventions stack up to traditional approaches in terms of effectiveness, user engagement, or their ability to help individuals learn key mental health skills. Moreover, there is a notable absence of research evaluating key long-term effects of AI-based interventions like weeks- and months-long changes in symptoms of depression or anxiety and long-term skill acquisition. This necessitates evaluating outcomes from a single user over a long duration and assessing repeated intervention usage, which is non-trivial. Also, the use of AI systems comes with harms and risks they inadvertently perpetuate (De Choudhury et al., 2023; Huang et al., 2023; Duffourc and Gerke, 2023; Weidinger et al., 2021). The absence of clinical trials means that there is a significant knowledge gap regarding the safety of AI-based mental health interventions, as well as how such safety measures should be established and evaluated in the critical high-risk setting of mental health. Additionally, AI systems are prone to biases that can lead to significant variance in performance across diverse demographics and populations. The extent to which AI-based interventions will have consistent effects across different subpopulations remains unknown.

Here, we demonstrate how clinical trials can effectively evaluate short-term and long-term outcomes, equity, and safety of AI-based mental health interventions. We use the AI-based cognitive restructuring intervention we developed in Chapter 3 as a case study. We conduct a clinical trial in an ecologically valid setting to compare our AI-based self-guided cognitive restructuring tool (DIY w/ AI) against an equivalent non-AI tool (DIY w/o AI) and psychoeducation only (Psychoeducation Only) control conditions (see Figure 4.1 for an overview).

We discuss the intervention groups that we intend to compare in this clinical trial, the relevant outcome measures, and the essential ethical and safety considerations. We then present the conceptual model that outlines this clinical trial, followed by the results demonstrating key



**Figure 4.1:** A conceptual model outlining our study. We conducted a between-subjects study where participants were divided into DIY w/ AI, DIY w/o AI, or Psychoeducation Only groups. All three groups had education components that included the education of restructuring skills. DIY w/o AI and DIY w/ AI groups have DIY components that include the practice of restructuring skills including identifying distortions and reframing thoughts. Finally, DIY w/ AI group has AI components that include just-in-time assistance on restructuring skills and interactive refinement. We study how these different interventions may affect outcomes both short-term and long-term in a study spanning multiple weeks. We also study how user attributes like demographics and depression/anxiety levels have an influence on outcomes.

similarities and differences between the various intervention groups.

#### 4.1.1 Intervention Groups for Randomized Controlled Trials

We implemented a between-subjects study design in which each participant was randomly assigned to one of the following three conditions:

**(1) Psychoeducation Only.** This group focuses on education components, teaching participants cognitive restructuring skills through traditional psychoeducation materials. Participants assigned to this group receive a PDF containing definitions and examples of various cognitive distortions, along with ways in which those thinking traps can be overcome, curated by mental health experts based on psychotherapy manuals.

**(2) DIY w/o AI.** This group provides participants with a DIY tool to practice cognitive restructuring. Participants use an online tool to guide them through the different steps of cognitive restructuring including identifying thinking traps and writing reframed thoughts. In addition, the DIY tool is integrated with the same education components as the Psychoeducation Only group.

**(3) DIY w/ AI.** This group additionally integrates the DIY tool from the DIY w/o AI group with AI components. Using the language model developed in Section 3.1, participants are provided with suggestions on the potential cognitive distortions that their thoughts entail and multiple possible thought reframes as possible starting points.

#### 4.1.2 Study Measures to Assess Short-term and Long-term Effects

We collaborate with mental health experts to identify and define key short-term and long-term outcomes that should be evaluated in this clinical trial:

**(1) Reframing Effectiveness.** The effectiveness of each reframe written by participants:

1. *Reframe Relatability:* I believe in the reframe I came up with (5: Strongly agree; 4: Somewhat agree; 2: Somewhat disagree; 1: Strongly disagree)
2. *Reframe Helpfulness:* The reframe helped me deal with the thought I was struggling with (5: Strongly agree; 4: Somewhat agree; 2: Somewhat disagree; 1: Strongly disagree)

3. *Reframe Memorability*: I will remember this reframe the next time I experience this thought (5: Strongly agree; 4: Somewhat agree; 2: Somewhat disagree; 1: Strongly disagree)
4. *Skill Learnability*: By doing this activity, I learned how I can deal with future negative thoughts (5: Strongly agree; 4: Somewhat agree; 2: Somewhat disagree; 1: Strongly disagree)

**(2) Emotional Regulation.** The effects that each reframe had in regulating the beliefs and emotions of the participants in-the-moment:

1. *Belief Change*: After doing this activity, how strongly do you believe in your original thought? (1 to 7)
2. *Emotion Change*: After doing this activity, how strong is your emotion? (1 to 7)

**(3) PHQ-9.** Questionnaire to assess the severity of depression. It asks: Over the last 2 weeks, how often have you been bothered by any of the following problems?

1. Little interest or pleasure in doing things (0 to 3)
2. Feeling down, depressed, or hopeless (0 to 3)
3. Trouble falling or staying asleep, or sleeping too much (0 to 3)
4. Feeling tired or having little energy (0 to 3)
5. Poor appetite or overeating (0 to 3)
6. Feeling bad about yourself or that you are a failure or have let yourself or your family down (0 to 3)
7. Trouble concentrating on things, such as reading the newspaper or watching television (0 to 3)
8. Moving or speaking so slowly that other people could have noticed. Or the opposite being so fidgety or restless that you have been moving around a lot more than usual (0 to 3)
9. Thoughts that you would be better off dead, or of hurting yourself (0 to 3)

**(4) GAD-7.** Questionnaire to assess the severity of anxiety. It asks: Over the last 2 weeks, how often have you been bothered by any of the following problems?

1. Feeling nervous, anxious, or on edge (0 to 3)
2. Not being able to stop or control worrying (0 to 3)
3. Worrying too much about different things (0 to 3)
4. Trouble relaxing (0 to 3)
5. Being so restless that it is hard to sit still (0 to 3)
6. Becoming easily annoyed or irritable (0 to 3)
7. Feeling afraid, as if something awful might happen (0 to 3)

**(5) Emotional Mechanisms.** The participant's level of hope, motivation, confidence, and willingness towards mental health goals or actions:

1. How hopeful do you feel about the future? (1 to 7)
2. How hopeful do you feel about managing negative thoughts? (1 to 7)
3. How worried do you feel about managing negative thoughts? (1 to 7)
4. How motivated are you to keep working on managing negative thinking? (1 to 7)
5. How confident are you in your ability to identify thinking traps? (1 to 7)
6. How confident are you in your ability to reframe negative thoughts? (1 to 7)
7. How willing are you to commit to engaging in this practice three times a week after this study? (1 to 7)
8. How likely are you to recommend it to a friend? (1 to 7)

**(6) Competencies of Cognitive Therapy Scale (Skill Acquisition).** Assessing the participants' cognitive therapy skills (Strunk et al., 2014)

1. *CCTS 20*: I caught myself thinking negatively, recognized the negative bias, and re-evaluated the situation. (1 to 7)
2. *CCTS 28*: When I got upset, I took time to step back from a situation and consider that my negative thoughts might be inaccurate. (1 to 7)
3. *CCTS 21*: I often caught myself thinking in an irrational way and I actively worked to develop more rational views. (1 to 7)
4. *CCTS 6*: When my negative thoughts and emotions really bothered me, I had a specific action plan of things I could do to cope. (1 to 7)
5. *CCTS 24*: When I blamed myself for something bad that happened, I took time to consider other factors that may have been involved. (1 to 7)
6. *CCTS 11*: When I found myself upset about something, I took note of what I was thinking and worked to develop a more balanced view. (1 to 7)

### 4.1.3 Ethical and Safety Considerations

The systematic ethical and safety considerations in this clinical trial were based on a principle-based ethics framework, following [Coghlan et al. \(2023\)](#), [Floridi and Cowls \(2022\)](#), and [Beauchamp and Childress \(2001\)](#).

Here, we discuss the five common principles derived from these frameworks, providing systematic guidance on how to responsibly navigate potential risks while developing and evaluating AI-based mental health interventions.

**(1) Non-maleficence.** *Avoid causing physical, social, or mental harm to participants.* Our tool was co-designed with mental health experts, patient advocates, and clinicians to identify any potential risks early. We further studied the safety implications prior to deployment. For this, we developed a realistic sandbox environment to ensure thorough testing instead of immediate deployment. Our safety testing involved examining system inputs that could potentially produce harmful outputs. Our safety filtering mechanisms were updated accordingly to address the potentially harmful LM-generated content (e.g., by adding specific regular expressions and using

a content moderation API). Additionally, throughout the study, participants were given access to a crisis hotline and were able to quit at any point.

**(2) Beneficence.** *Ensure that interventions do good or provide real benefit to participants.* This involved co-designing with mental health experts, patient advocates, and clinicians to identify opportunities to benefit. This led us to focus on cognitive restructuring, which is a well-established, evidence-based intervention that has been shown to positively impact people’s mental well-being. Our initial study on Mental Health America demonstrates that our system positively impacts emotional intensity for 67% of participants and helps 65% overcome negative thoughts (Section 3.2).

**(3) Respect for Autonomy.** *Respect participants’ values and choices.* We prioritized human agency and initiative by enabling participants to maintain control over what they want to write. Our language model initially offered suggestions on how to reframe thoughts. However, further interaction with the model was only possible when explicitly requested by the participant. Collection of demographic and outcome data was intentionally left optional.

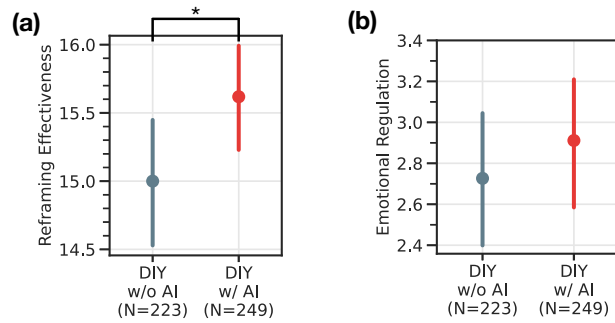
**(4) Justice.** *Treat participants without unfair bias, discrimination or inequity.* Our clinical trial incorporated a broad set of participants with varied demographics, including underrepresented subpopulations. We explicitly evaluated the equity of positive outcomes and identified key subpopulations where the intervention was found to be less effective.

**(5) Explicability.** *Provide users with sufficient transparency about the nature and effects of the technology, and be accountable for its design and deployment.* We sought informed consent from participants and were transparent about the use of AI, risks, and data use (Appendix Figure B.6 and Figure B.7). We also continuously monitored quantitative and qualitative feedback for potential concerns, which also informed our design hypotheses. Finally, we make our code publicly available at [github.com/behavioral-data/Self-Guided-Cognitive-Restructuring](https://github.com/behavioral-data/Self-Guided-Cognitive-Restructuring).

#### 4.1.4 Study Procedure

Our clinical trial was conducted over a period of 8 weeks and consisted of several phases.

During the initial 4 weeks, participants completed weekly questionnaires assessing their depression (PHQ-9), anxiety (GAD-7), emotional mechanisms, and cognitive therapy competence



**Figure 4.2:** We found that DIY w/ AI led to significantly higher reframing effectiveness compared to DIY w/o AI during the first usage (Cohen’s  $d = 0.18$ ). We also found that DIY w/ AI led to similar emotional regulation as DIY w/o AI (Cohen’s  $d=0.07$ ). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals. \*:  $p < 0.05$

(CCTS). Those in the DIY w/ AI and DIY w/o AI groups were also required to use their respective online tools at least three times per week, while the Psychoeducation Only group received weekly psychoeducation materials.

In week 5, participants completed an exit survey to gather qualitative feedback on their overall experience.

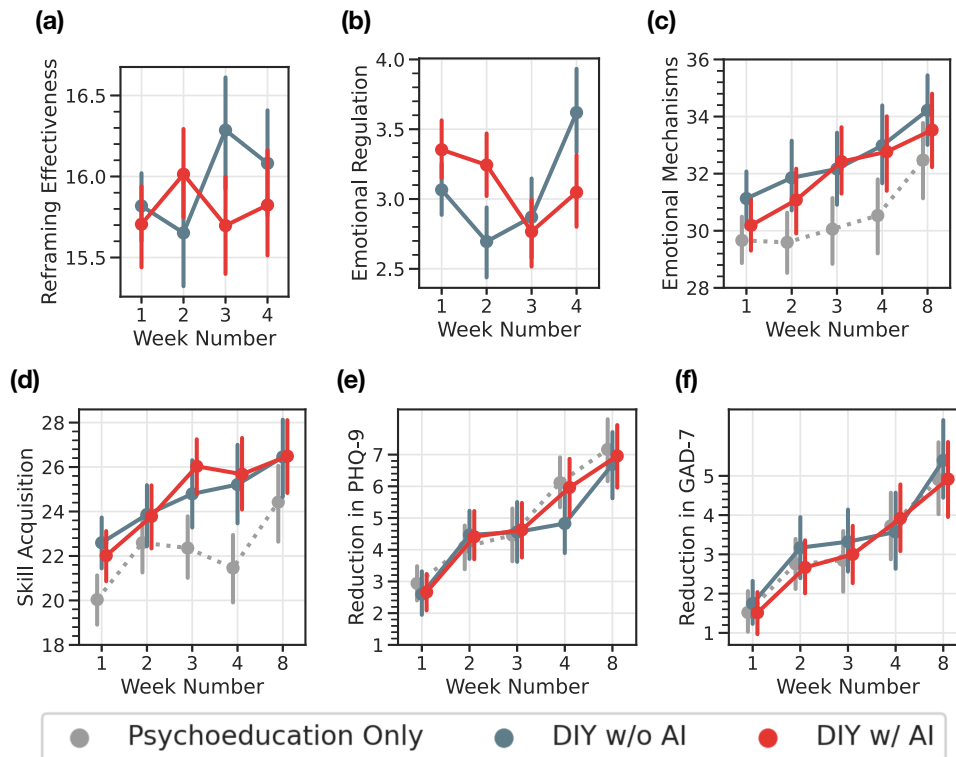
Finally, in week 8, we conducted a follow-up survey to assess whether the effects were sustained.

#### 4.1.5 Results

**Increased short-term effectiveness when using AI.** First, we examined the differences between DIY w/ AI and DIY w/o AI when the DIY component of cognitive restructuring is used for the first time. Since these individuals are likely unfamiliar with cognitive restructuring techniques, their understanding of how to practice them effectively may be limited. Therefore, AI-based assistance could offer valuable assistance in navigating the various steps, such as identifying thinking traps and writing effective reframes.

We found that participants in the DIY w/ AI group reported significantly higher reframing effectiveness during the first usage compared to participants in the DIY w/o AI group (Cohen’s  $d = 0.18$ ;  $p$ -value = 0.047; Figure 4.2a). Also, we found that participants in both groups reported similar levels of regulation in emotions (Cohen’s  $d = 0.07$ ;  $p$ -value = 0.426; Figure 4.2b).

This suggests that integrating AI into self-guided mental health tools could significantly im-



**Figure 4.3:** We assess the outcomes of DIY w/ AI, DIY w/o AI, and Psychoeducation Only groups over a period of 8 weeks. We found that participants in the DIY w/ AI and DIY w/o AI conditions demonstrated similar levels of Emotional Mechanisms, Skill Acquisition, Reduction in PHQ-9, and Reduction in GAD-7. However, participants in the Psychoeducation Only condition demonstrated significantly lower emotional mechanisms and skill acquisition levels than participants in DIY w/ AI and DIY w/o AI conditions. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

prove initial reframing effectiveness.

### Increased long-term engagement and similar levels of long-term effectiveness when using AI.

While an intervention might initially regulate emotions and reduce the symptoms of a mental illness, it is the sustained impact over time that truly reveals its effectiveness. Without sustained change, the symptoms are likely to resurface. Here, through a controlled and ecologically valid study, in which engage users for 8 weeks, we assess the long-term outcomes of DIY w/ AI, DIY w/o AI, and Psychoeducation Only groups.

We found that AI-based self-guided cognitive restructuring intervention significantly improved long-term engagement, as measured by the number of times cognitive restructuring is practiced, as compared to the non-AI intervention. Concretely, we found that participants in the DIY w/ AI used the DIY tool significantly more than participants in the DIY w/o AI condition (Cohen's  $d = 0.13$ ;  $p$ -value  $< 0.001$ ).

Furthermore, participants in the DIY w/ AI condition demonstrated similar long-term effectiveness as DIY w/o AI condition, including similar levels of skill acquisition (Cohen's  $d = 0.027$ ;  $p$ -value = 0.613), similar levels of emotional mechanisms (Cohen's  $d = -0.064$ ;  $p$ -value = 0.236), and similar reductions in depression (Cohen's  $d = 0.060$ ;  $p$ -value = 0.266) and anxiety (Cohen's  $d = -0.043$ ;  $p$ -value = 0.429) as measured over weeks 1 to 8 (Figure 4.3c-f).

Moreover, they demonstrated similar levels of reframing effectiveness (Cohen's  $d = -0.022$ ;  $p$ -value = 0.443) and emotional regulation (Cohen's  $d = 0.018$ ;  $p$ -value = 0.535) aggregated over weeks 1 to 8 (Figure 4.3a-b).

Participants in Psychoeducation Only demonstrated significantly lower emotional mechanisms and skill acquisition levels than participants in DIY w/o AI and DIY w/ AI groups suggesting that the practice of restructuring through a DIY tool is much more effective.

These findings suggest that AI is particularly helpful in improving engagement, a key outcome in mental health interventions. Also, AI and non-AI interventions demonstrate similar long-term effectiveness. While AI might be assumed to be "doing the thinking for you", participants who used AI actually improved their cognitive restructuring skills and achieved similar positive mental health outcomes in the long term as those who did not use AI.

**Heterogeneous treatment effects across subpopulations.** AI-based interventions are known to be disparate across different populations. Here, we found that people of certain demographics found AI to be more effective than others.

We found significant differences between the effects of AI based on the age and initial PHQ-9 and GAD-7 levels of the participants (Table 4.1). We found that young adults (age 18 to 24) assigned to DIY w/ AI reported significantly higher levels of reframing effectiveness (Cohen's  $d = 0.23$ ;  $p$ -value = 0.001), emotional regulation (Cohen's  $d = 0.69$ ;  $p$ -value =  $6.09 * 10^{-20}$ ), and emotional mechanisms (Cohen's  $d = 0.28$ ;  $p$ -value = 0.038) compared to young adults assigned to the DIY w/o AI.

We also found differences based on the initial PHQ-9 and GAD-7 levels reported at the start of the study. Participants with medium depression (PHQ-9 from 15 to 19) assigned to DIY w/ AI reported significantly higher levels of reframing effectiveness (Cohen's  $d = 0.19$ ;  $p$ -value = 0.0003), emotional regulation (Cohen's  $d = 0.22$ ;  $p$ -value =  $2.98 * 10^{-5}$ ), and reduction in PHQ-9

Population	Reframing Effectiveness	Emotional Regulation	Emotional Mechanisms	Skill Acquisition	Reduction in PHQ-9	Reduction in GAD-7	N
<b>Age</b>							
18-24	0.23	0.69	0.28	0.22	-0.13	-0.25	356
25-34	-0.09	-0.26	-0.19	0.01	0.10	-0.09	607
35-44	-0.12	0.03	-0.21	-0.05	0.21	0.03	564
45-54	0.14	0.14	0.19	0.05	-0.12	-0.18	401
55-64	-0.21	-0.30	-0.47	-0.11	0.21	0.72	171
65+	-0.12	-0.56	-0.40	-0.12	-0.62	-0.60	57
<b>Initial PHQ-9</b>							
10-14	-0.01	-0.09	-0.06	-0.05	-0.03	-0.26	584
15-19	0.19	0.22	0.05	0.06	0.24	0.13	704
20-27	-0.11	-0.01	-0.11	0.07	-0.02	0.03	796
<b>Initial GAD-7</b>							
10-14	0.03	-0.01	0.08	0.30	-0.00	0.00	707
15-21	-0.16	-0.05	-0.25	-0.16	0.07	0.01	1121

**Table 4.1:** Effect sizes (Cohen’s d) comparing outcomes between DIY w/ AI participants and DIY w/o AI participants, across various subpopulations. Numbers highlighted in green indicate outcomes where DIY w/ AI is significantly better than DIY w/o AI ( $p < 0.05$ ). Numbers highlighted in red indicate outcomes where DIY w/ AI is significantly worse than DIY w/o AI ( $p < 0.05$ ). We found that young adults (age 18 to 24) assigned to DIY w/ AI reported significantly higher levels of reframing effectiveness and emotional regulation compared to young adults assigned to the DIY w/o AI.

(Cohen’s  $d = 0.24$ ;  $p$ -value = 0.012) compared to participants with medium depression assigned to the DIY w/o AI. On the other hand, participants with severe depression (PHQ-9 from 20 to 27) assigned to DIY w/ AI reported significantly lower levels of reframing effectiveness (Cohen’s  $d = -0.11$ ;  $p$ -value = 0.024) compared to participants with severe depression assigned to the DIY w/o AI. Moreover, participants with severe anxiety (GAD-7 from 15 to 21) assigned to DIY w/ AI reported significantly lower levels of reframing effectiveness (Cohen’s  $d = -0.16$ ;  $p$ -value =  $4.62 \times 10^{-5}$ ), emotional mechanisms (Cohen’s  $d = -0.25$ ;  $p$ -value = 0.0007), and skill acquisition (Cohen’s  $d = -0.16$ ;  $p$ -value = 0.028) compared to participants with severe anxiety assigned to the DIY w/o AI.

### Analysis of the Language Model Generated Content that was Flagged Inappropriate

We implemented a feature for participants to report any inappropriate content generated by the language model. This was achieved by including a distinct “Flag inappropriate” button

for each generated reframe (Section 3.2.2). Overall, we found that 0.65% (301 out of 46,593) of the reframing suggestions shown were flagged. After conducting a qualitative review of these flagged reframes, we found a few dozen instances where the model's suggestions repeated the negative factors/sentiment described by the participant in the original thought which may inadvertently reinforce negative beliefs about oneself (e.g., "I may be a "failure", but I'm still trying my best." in response to the thought "I'm a failure"). Note that in most cases repeating parts of the participant's thought or situation helps to validate their experience and emotional reaction and personalize their reframe. Therefore, this highlights the importance of effectively differentiating which aspects of the participant's thoughts to re-state in the reframing suggestions. Future work should look more closely at how to facilitate this differentiation.

Nevertheless, for many of the 301 flagged instances, because all participants were shown three reframes as starting points, participants were able to select a different reframing suggestion from the three options presented to them, eventually reporting favorable outcomes. We also checked user dropout between users who flagged content vs. those who did not and found no significant differences. Interestingly, the dropout rate was slightly lower among users who flagged content, at 38.2%, compared to 46.5% for those who who did not. Notably, none of the flagged instances had references to suicidal ideation or self-harm, suggesting that the safety mechanisms designed to address these concerns were likely effective (Section 3.2.2).

## 4.2 A Computational Framework for Behavioral Assessment of LLM Therapists

The previous section demonstrates how AI-based mental health interventions can be evaluated with human participants through clinical trials. However, the rapid pace at which LLMs are being developed and iteratively improved also necessitates the exploration of *automated* evaluation methods to keep up with advancements. Here, we study how to conduct automatic, systematic evaluation of AI-based mental health interventions.

Ever since Eliza, a mock psychotherapist, was introduced in the 1960s (Weizenbaum, 1966), the prospect of therapy chatbots that can offer mental health support has intrigued clinicians, researchers, as well as the general public. The recent advent of ChatGPT and the rapid progress

in large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023; Bubeck et al., 2023) have amplified the excitement. This growing excitement is driven by a stark reality: Hundreds of millions of people do not or cannot receive mental health care for various reasons, including a shortage of professionals, lack of insurance coverage, high costs, and stigma surrounding therapy (Hosny and Aerts, 2019; WHO, 2022; Olfson, 2016; Sickel et al., 2014). As a result, individuals and researchers have started using LLMs as a potential solution to bridge this critical access gap (Hamilton, 2023; Broderick, 2023; Stade et al., 2024; Youper, 2023; Sharma et al., 2023a; Lin et al., 2024). However, mental health experts have raised concerns and caution about using LLMs in therapeutic contexts without appropriate evaluation (Stade et al., 2024; De Choudhury et al., 2023; Li et al., 2020; Tate et al., 2023; Van Veen et al., 2024).

As such rapidly evolving trends continue in the critical high-stakes setting of mental health, our field lacks methods to systematically assess LLMs used for mental health support. At a fundamental level, our understanding of the *behaviors* exhibited by LLMs, i.e., ways in which they respond to clients seeking mental health support, is significantly limited. It remains unclear under what circumstances LLMs emphasize certain behaviors like reflecting on client emotions or problem-solving and to what degree. Understanding the behavior of LLMs is particularly important in high-stakes settings such as mental health where undesirable behavior could have severe consequences on vulnerable clients. Moreover, identifying the desirable and undesirable behaviors can help identify the potential benefits and risks, which can inform adoption and iterative development.

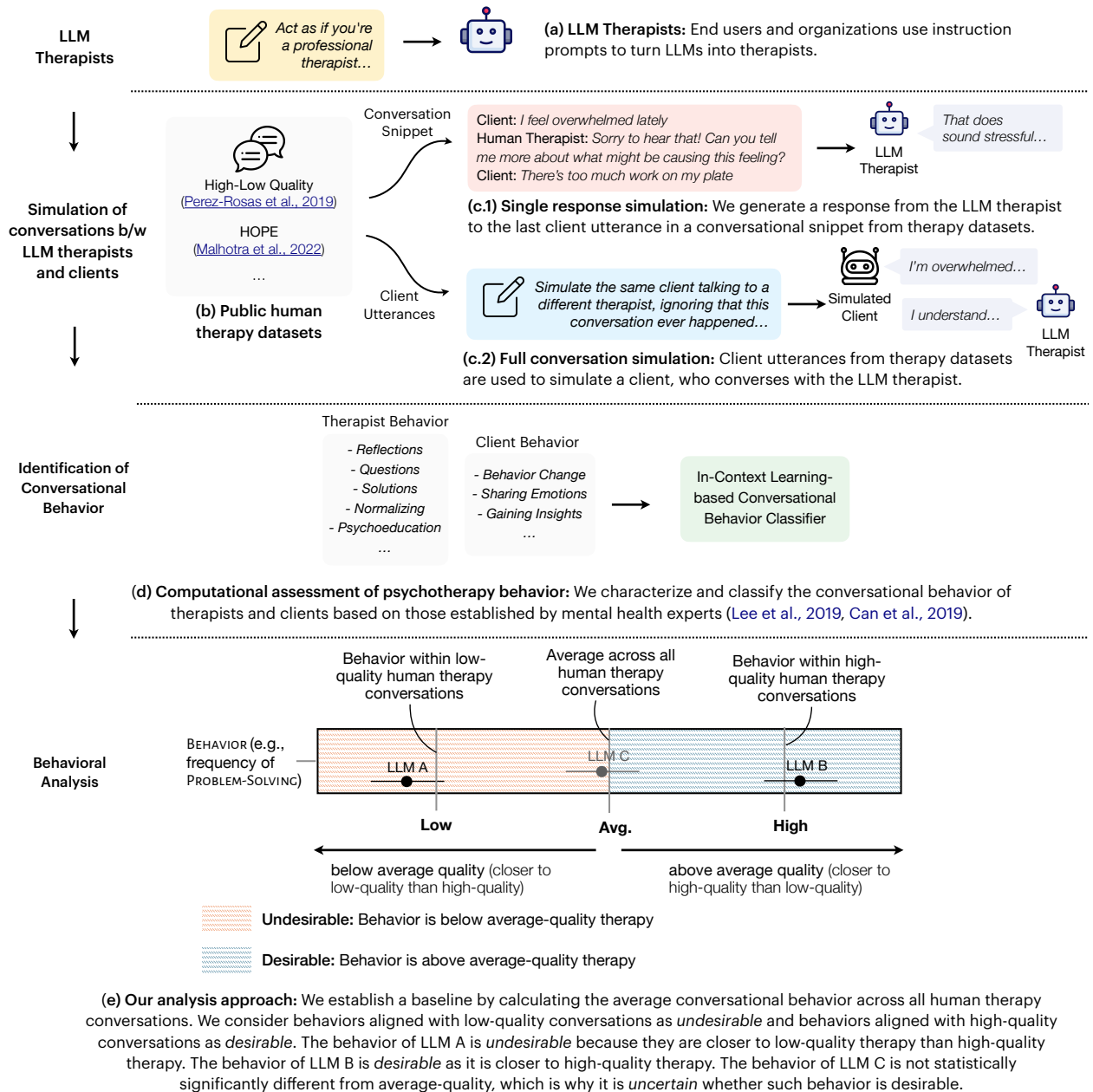
Such contexts highlight the critical need to evaluate whether LLM-based interventions are feasible, effective, and safe in mental healthcare, how they compare to established interventions, and whether they would ultimately be helpful to patients and clinicians. However, evaluating LLMs being used for mental health support poses significant ethical and technical challenges. Directly recruiting clients struggling with mental health issues and asking them to interact with untested LLMs for evaluation poses significant potential risks. It exposes these vulnerable populations to potentially harmful LLM generations. Also, the evaluation necessitates a large-scale assessment of different types of open-ended texts that LLMs may generate, in response to a diverse range of clients, for different types of support, and at different stages of support, making the behavioral assessment complex and multi-faceted. Traditional evaluation methods, such as

the Cognitive Therapy Rating Scale (CTRS) (Vallis et al., 1986), outline desirable and undesirable therapist behaviors. However, they have been designed for manual assessment of human therapists. Due to the complexity of such scales, it is difficult to adapt them to automated, large-scale LLM evaluation. Therefore, efforts to translate these guidelines into actionable metrics for automatic, large-scale evaluations have been limited, with existing work focusing only on specific therapy types (Flemotomos et al., 2022).

In this work, we conduct a proof-of-concept study proposing a framework for evaluating LLMs in mental health support. We develop BOLT, a novel, general computational framework to systematically study the **Behavior Of LLMs** in the context of their application in **Therapy**. Given an LLM (e.g., GPT-4) that is being used as a therapist (Section 4.2.1), we aim to quantitatively assess the ways in which it responds when supporting clients with mental health issues (Section 4.2.2). Subsequently, we aim to compare the behavior of *LLM therapists* against that of high- and low-quality human therapists, and study whether their behavior can be modulated into more desirable behaviors (Section 4.2.2).

To analyze the behavior of LLM therapists, we need conversations between LLM therapists and clients. BOLT simulates conversations between simulated clients and LLMs, leveraging client-human therapist conversations from existing public datasets of therapy conversations to ensure high fidelity (Section 4.2.2). Next, to assess the conversational behavior of therapists and clients in these conversations, BOLT leverages a prompting-based classifier that uses psychotherapy-based definitions and in-context examples to identify the psychotherapy techniques underlying all utterances. The techniques, which are operationalized based on those established by mental health experts, include REFLECTIONS, QUESTIONS, SOLUTIONS, NORMALIZING, and PSYCHOEDUCATION for therapists and BEHAVIOR CHANGE, SELF-DISCLOSURE OF AFFECT OR EXPERIENCE, and GAINING INSIGHTS for clients (Section 4.2.2).

Furthermore, BOLT establishes behaviors that are representative of high-quality therapy and low-quality therapy sessions. We use high-quality and low-quality conversations as annotated by previous NLP for Mental Health research (Pérez-Rosas et al., 2019) based on established principles of Motivational Interviewing. We analyze these conversations using our prompting-based behavior classifier to identify specific therapist and client behaviors associated with high-quality and low-quality Motivational Interviewing. This allows us to distinguish between potentially



**Figure 4.4:** Overview of BOLT, a computational framework that enables systematic assessment of the behavior of LLM therapists and compares them to high- and low-quality human therapy.

desirable and undesirable behaviors (Section 4.2.2).

Finally, BOLT assesses the behavior of these LLM therapists based on the *frequency* of their behavior, the *temporal order* in which the behavior is expressed in conversations, their *adaptability* to different behaviors from clients, and the expressed *linguistic attributes*. For each of these assessments, we make comparisons with average-, high- and low-quality conversations from human therapists.

To illustrate the utility of BOLT, we develop a case study applying it to Motivational Interviewing (MI), a popular, well-established therapeutic approach (Miller and Rollnick, 2012). In this study, we evaluate the behavior of four popular, state-of-the-art LLM variants – GPT-4 (OpenAI, 2023), GPT-3.5-turbo (Brown et al., 2020), Llama2-70b, and Llama2-13b (Touvron et al., 2023), when providing MI. As different therapy frameworks (e.g., MI vs. Problem-Solving Therapy (Haley, 1992)) might have varying nuances regarding what is considered high-quality therapy, an assessment within a specific therapeutic approach is required in order to interpret the findings in a principled manner and at high granularity. However, note that BOLT is intentionally designed to be generalizable, with behavioral techniques that are broadly applicable in psychotherapy. Such a design could potentially generalize to other psychotherapy contexts and varied therapeutic frameworks. We demonstrate this through a secondary experiment with other therapeutic approaches (Section 4.2.2).

While anecdotal examples have demonstrated striking similarity to plausible therapist statements, our systematic approach reveals that LLM therapists often more closely resemble the behavior of human therapists in low-quality rather than high-quality sessions.

#### **4.2.1 Bolt: Framework for Assessing Conversational Behavior of LLM Therapists**

##### **LLM Therapists**

A common way in which current LLMs are being used as therapists is through custom “system prompts” that instruct them to function as therapists. BOLT uses a simple “system prompt” which was designed based on commonly reported example prompts by users on mental health subreddits (Reddit-1, 2023; Reddit-2, 2023):

*Act as if you’re a professional therapist. You provide evidence-based therapy to help clients*

*seeking help with mental health challenges. You should maintain your therapist persona while responding. Communicate in a conversational style, mirroring the style of previous therapist responses.*

We make subtle variations to this prompt to adapt it to different therapy types, such as Motivational Interviewing (Appendix Figure C.2). We study the effects of variations in “system prompts” (Section 4.2.2).

## Datasets of Therapy Conversations

Acquiring datasets of therapy conversations poses a significant challenge as they are typically private and rarely shared. Moreover, potential privacy issues may arise when exposing therapy datasets to public LLM APIs as they may contain sensitive client information. To circumvent these potential concerns, BOLT leverages publicly available datasets of therapy conversations.

Publicly available therapy conversation datasets are limited. Here, we use two datasets that carefully preprocess publicly available YouTube videos on therapy. This ensures high-quality transcripts while maintaining the confidentiality of sensitive personal information. We use both datasets with appropriate license and consent as provided by the authors within their terms of use.

**(1) High-and-Low-Quality Therapy Conversation Dataset (High-Low Quality).** The first dataset, created by Perez-Rosas et al. (Pérez-Rosas et al., 2019), comprises 259 therapy conversations, primarily focusing on the evidence-based motivational interviewing (MI) therapy. Based on the MI psychotherapy principles, they evaluate the quality of conversations in the dataset to have 155 high-quality therapy transcripts and 104 low-quality transcripts. We use both high-quality and low-quality therapy conversations by human therapists to understand desirable and undesirable conversational behaviors.

**(2) HOPE Dataset.** The second dataset from Malhotra et al. (Malhotra et al., 2022) was used to study dialogue acts in therapy. This dataset contains 212 therapy transcripts and includes conversations employing different types of therapy techniques (e.g., MI, Cognitive Behavioral Therapy).

Both of these datasets (available in CSV formats with speaker and utterance columns) were

originally sourced from public video-sharing platforms like YouTube. We use these datasets to simulate conversations between LLM therapists and simulated clients. Moreover, we use them to study the behavior of LLM therapists in varied mental health contexts (Section 4.2.2).

## Simulating Conversations between LLM Therapists and Clients

To characterize and analyze the behavior of LLM therapists, we need conversation data between LLM therapists and clients. BOLT uses public therapy datasets to simulate these conversations. We develop two simulation strategies representing common LLM therapist use cases –

**(1) LLM Single Response Simulation.** We use LLMs to write a single response within an ongoing conversation session between a human therapist and a client. Let  $C_i$  be an ongoing conversation between a human therapist and client with utterances  $u_{i,1}, u_{i,2}, \dots, u_{i,j}$  ending with a client utterance  $u_{i,j}$ . We aim to use an LLM  $\mathcal{L}$  to generate a single response  $u_{i,j+1}^{(\mathcal{L})}$  to the client utterance. To simulate such single responses, we use the therapy datasets to create all possible sequences of  $u_{i,1}, u_{i,2}, \dots, u_{i,j}$  ending with a client utterance. For each of these sequences, we use the “system prompt” to generate the next LLM therapist response  $u_{i,j+1}^{(\mathcal{L})}$ . The LLM therapist here responds to a prior, real-world, authentic interaction between a human client and a human therapist. This is helpful in not biasing the LLM based on simulated clients or therapists. Check Appendix Tables of the first part C.28 and second part C.29 for an illustrative example.

**(2) LLM Full Conversation Simulation.** In addition to single-response generation, we also simulate a full, multi-turn conversation session between LLMs and clients through a novel method based on highly consistent simulated clients. Such a simulation is helpful in assessing the behavior of LLM therapists across multiple turns of support to the same simulated client. Importantly, we evaluate the consistency of these simulated conversations. Here, we leverage the utterances from clients in the therapy datasets to simulate artificial clients using GPT-4. For each human-human conversation between a client and a therapist, we provide it as input to GPT-4 and ask it to simulate the same client talking to a different therapist in a parallel universe, ignoring that this conversation ever happened (to avoid GPT-4 continuing from the provided human-human conversation rather than simulating a new similar conversation), but keeping the same conversational style, the manner of addressing certain topics or concerns, and the life events and emotions

being discussed (also see Appendix Figure C.2). Then, we generate a conversation between this artificial client and the LLM therapist. We start the generation by randomly picking between the client or therapist to go first. We let them talk for up to 20 turns (which roughly corresponds to the mean conversation length (20.8 turns) in the two datasets used) or until one of them outputs an end token. See Appendix Tables C.30 and C.31 for illustrative examples.

**Evaluating Simulated Conversations.** We evaluate the consistency of the simulated conversations. We define *consistency* as generating an utterance that is logically and semantically consistent with the previous utterances without conflicting with them. We use a GPT-4-based few-shot prompting method (Ziems et al., 2023) to measure the consistency and coverage of the simulated conversations (this measure achieves an accuracy of 95.83% and macro-f1 of 82.22% in our human evaluation). We find that the simulated conversations have a near-perfect mean consistency of 2.95 (out of 3; std = 0.13). In addition, we qualitatively assess the conversations and find that simulated conversations follow the source conversational snippets in their overall theme and structure.

## Behavioral Techniques in Psychotherapy

To assess the conversational behavior exhibited by therapists and clients during therapy sessions, we characterize the psychotherapeutic approaches underlying utterances. We draw upon the clinically relevant dialogue acts of therapist and client utterances (intentions of the speaker in the utterance (Stolcke et al., 2000)) established by experts in psychotherapy in two previous studies (Lee et al., 2019; Cao et al., 2019).

We discuss the conversational behaviors in the following. Then we describe our method for automatically identifying these behaviors.

**Conversational Behavior of Therapists** For characterizing the conversational behavior of therapists, we focus on 13 major psychotherapeutic approaches based on techniques identified in Lee et al. (Lee et al., 2019) and Cao et al. (Cao et al., 2019) organized into five semantically related categories:

(a) **Reflections.** A foundational tool for any therapist is to understand, restate, and emphasize

the client's feelings and experiences (e.g., *"You are feeling overwhelmed and struggling to find a balance between work and family"*) (Arnold, 2014). Such reflective listening has been shown to help clients gain a deeper understanding of their feelings and experiences and also improve the therapeutic relationship between the therapist and client through increased empathy (Brailion and Taiebi, 2020; Rautalinko et al., 2007). Here, we focus on six different ways of reflecting differentiating between (1) NEEDS, (2) EMOTIONS, (3) VALUES, (4) CONSEQUENCES, (5) CONFLICTS, and (6) STRENGTHS (Appendix Table C.23)

**(b) Questions.** Open-ended questions constitute a fundamental element of therapy, helping clients express emotions and share thoughts and experiences they might not explore otherwise (e.g., *"How were you feeling in that moment?"*) (Lundh, 2019; Rautalinko et al., 2007). Also, expressing an active interest in clients' emotions and experiences enhances relationship forming with clients (Miller et al., 2003; Robert et al., 2011). Here, we focus on open-ended questions differentiating between (7) EXPERIENCES, (8) PERSPECTIVES, and (9) EMOTIONS (Appendix Table C.23).

**(c) Solutions.** To achieve the therapeutic goal effectively, therapists often offer solutions to client's problems by discussing concrete problem-solving approaches and constructing a specific plan of action (Nelson-Jones, 2013; Solomonov et al., 2019; Fawcett and Borck-Jameson, 2014). Here, we focus on these two types of solutions differentiating between (10) PROBLEM-SOLVING, and (11) PLANNING (Appendix Table C.23).

**(d) Normalizing.** Being warm, friendly, and respectful has been found to contribute positively to the relationship (or alliance) between the client and therapist, which can be achieved through acknowledging and validating clients's feelings and experiences, empathizing, and reassuring (e.g., *I hear you, it's perfectly normal to feel overwhelmed...*). Hence, we combine these related concepts into (12) NORMALIZING.

**(e) Psychoeducation.** Therapists commonly educate clients about coping strategies, treatment approaches, symptoms, or diagnosis (e.g., *"Cognitive Behavioral Therapy aims to..."*) (Lundh, 2019). Here, we broadly call this aspect as (13) PSYCHOEDUCATION.

Refer to Appendix Table C.23 for definitions and examples of all 13 therapist conversational behaviors.

**Conversational Behavior of Clients** For a therapist, knowing when to express specific behaviors based on different client behaviors (e.g., when to reflect vs. when to utilize problem-solving) is an important skill. Here, to more thoroughly assess therapist behavior based on how they respond and adapt to clients, we characterize the conversational behavior of clients. We use six types of expressions from clients based on the existing works (Lee et al., 2019; Cao et al., 2019) organized into three semantically related categories. We focus on a combination of client behaviors that include the changes in client behaviors, the emotions and experiences they share, and the insights they gain from therapy (Carey et al., 2007; Tulver et al., 2023):

**(a) Behavior Change.** Some clients may struggle with unhealthy behaviors or symptoms (E.g., *“I’m having trouble sleeping”*). Here, we differentiate between clients expressing (1) CHANGING UNHEALTHY BEHAVIOR and (2) SUSTAINING UNHEALTHY BEHAVIOR (Appendix Table C.24).

**(b) Self-Disclosure of Affect or Experiences.** During the conversation, clients may share their feelings or related experiences (e.g., *“I lost my job due to the pandemic”*). Here, we focus on clients’ expression of (3) SHARING POSITIVE EMOTIONS, (4) SHARING NEGATIVE EMOTIONS, and (5) SHARING EXPERIENCES (Appendix Table C.24).

**(c) Gaining Insights.** Clients often report learning something new about themselves or their situation during therapy (e.g., *“I hadn’t considered how much I avoid confrontations”*). Here, we detect such (6) GAINED INSIGHTS by clients.

Refer to Appendix Table C.24 for definitions and examples of all six client conversational behaviors.

## Identifying Conversational Behavior in Psychotherapy Conversations

BOLT automatically identifies the conversational behavior of LLM therapists based on the simulated conversations. We annotate the therapy datasets with conversational behavior and use it to train and evaluate GPT-based methods.

### Classification Setup

We frame the identification of conversational behaviors as either a single multi-label classification task (identify all possible behaviors in a given utterance at once simultaneously) or a multiple

binary-label classification task (identify independently whether a particular conversational behavior is exhibited by an utterance).

Let  $C_i$  be a conversation between a therapist and a client with utterances  $u_{i,1}, u_{i,2}, \dots$

**Multi-label classification.** For each therapist utterance,  $u^{(t)} \in \{u_{i,t_1}, u_{i,t_2}, \dots\}$ , we aim to identify all possible behaviors exhibited by  $u^{(t)}$  among the 13 different therapist codes. Similarly, for each client utterance,  $u^{(c)} \in \{u_{i,c_1}, u_{i,c_2}, \dots\}$ , we aim to identify all possible behaviors exhibited by  $u^{(c)}$  among the 6 different client codes.

**Binary-label classification.** For each therapist utterance,  $u^{(t)} \in \{u_{i,t_1}, u_{i,t_2}, \dots\}$  and for each therapist behavioral code in the 13 possible therapist codes, we aim to identify whether  $u^{(t)}$  exhibits that behavior. Similarly, for each client utterance,  $u^{(c)} \in \{u_{i,c_1}, u_{i,c_2}, \dots\}$  and for each client behavioral code in the 6 possible client codes, we aim to identify whether  $u^{(c)}$  exhibits that behavior.

### Annotations of Conversational Behavior

The broad range of therapist and client conversational behavior and their complexity make crowdwork data annotation process challenging. Here, two co-authors of this paper with significant expertise in NLP and mental health research, and who were actively engaged in defining the conversational behaviors of therapists and clients, annotated 283 therapist utterances and 91 client utterances (inter-rater agreement, krippendorff's alpha = 0.7834). Our annotated dataset includes conversations from the human therapy datasets as well as the simulated conversations based on LLM therapists.

### Models

We compare three models to identify conversational behaviors in therapist and client utterances. Note that GPT-4-based methods are the state-of-the-art for this task.

**Finetuning (multi-label).** The first method involved fine-tuning GPT-3 series models (GPT-3: davinci-002, GPT-3.5: gpt-3.5-turbo-0613) (Brown et al., 2020) with our annotated dataset described above (Note that GPT-4 did not allow finetuning at the time of the writing of this paper. Finetuning for classifying individual conversational behavior for the binary classification method was prohibitively expensive. Therefore, we instead finetune two multi-label classification models

for the conversation behaviors of therapists and clients, respectively.)

**Prompting / In-context Learning (multi-label).** We facilitate in-context learning by prompting GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) models to identify all possible multiple intents in a single utterance. We explored two variations of prompting – (1) We prompt the models with definitions of different conversational behaviors sourced from prior work (Lee et al., 2019; Cao et al., 2019) (Appendix Tables C.23 and C.24); (2) We prompt the models with definitions as well as in-context examples from our annotated dataset (We tried different numbers of in-context examples and found three examples to work the best). We use the following base instruction prompt for both variations – “*What are all possible conversational behaviors of this utterance.*” Check Appendix Figure C.1 for all employed prompts.

**Prompting / In-context Learning (binary-label).** We create binary questions focusing on one intent at a time with its definition and in-context examples (e.g., “*Classify if the utterance contains {conversational behavior}. Answer in Yes or No*”). Check Appendix Figure C.1 for all employed prompts.

## Experiments and Results

**Experimental Setup.** We create five random train-test splits of our annotated dataset, each with a 60:40 split. We use the training split to finetune GPT-3 models and as demonstration examples for in-context learning. For all models, we report the mean and standard deviation of Macro-Precision, Macro-Recall, and Macro-F1 scores for the 13-class therapist behavior classification task and the 6-class client behavior classification task. Given the limited dataset size, averaging across five random splits ensures a more robust reflection of classification performance. As a baseline for comparison, we assign random conversational behavior to utterances.

**Results.** We find that the prompting-based methods outperform the fine-tuned models for both therapist and client behavior (Table C.22). For therapist behavior, among the different variations of the prompts, using definitions and examples in the multi-label setting achieves the best performance with a macro-F1 score of 57.7% (std = 2.5%) which is 43.6% greater than the random baseline, which attained macro-F1 score of 14.1% (std = 1.3%). For client behavior, using definitions and examples in the binary-label setting achieves the best performance with a macro-F1

score of 76.7% (std = 6.3%). This performance is promising especially when compared to the random baseline, which only obtained a 25.9% macro-F1 score (std = 6.2%).

Our analysis shows that the inclusion of examples in prompts significantly enhances the performance of the classification models. This improvement can be attributed to the fact that examples provide the model with a clearer understanding of the nuanced differences in conversational behaviors, particularly those that are complex, such as REFLECTIONS ON VALUES, REFLECTIONS ON CONSEQUENCES, etc. For therapist behaviors, our best classification model has high accuracy for most of the conversational behaviors ranging from 75.6% (QUESTIONS ON EXPERIENCES) to 95.3% (QUESTIONS ON EMOTIONS). For client behaviors, our model performs well for all conversation behaviors, with accuracy ranging from 81.0% (SHARING NEGATIVE EMOTIONS) to 95.4% (SUSTAINING UNHEALTHY BEHAVIOR). See Appendix Table C.25 and C.26 for individual behavior performances of therapists and clients.

## 4.2.2 Results

### Associating Conversational Behavior with High-Quality and Low-Quality Therapy

Now that we can identify behaviors in conversations, we need to understand which behaviors are high-quality or low-quality. This differentiation is crucial, as it helps us determine desirable and undesirable behaviors. Here, we leverage the High-Low Quality dataset from Perez-Rosas et al. (Pérez-Rosas et al., 2019) containing 259 Motivational Interviewing (MI) conversations annotated as either high-quality or low-quality. We analyze these conversations using our behavior classifier to identify specific therapist and client behaviors associated with high-quality and low-quality MI. Our assessment focuses on the *frequency* of behavior, the *temporal order* in which it's expressed, and its *adaptability* to different client behaviors.

What differences between high- and low-quality human-delivered MI therapy does this dataset reveal? We find that therapists from high-quality human therapy conversations more frequently express REFLECTIONS ON NEEDS, REFLECTIONS ON EMOTIONS, REFLECTION ON VALUES, REFLECTIONS ON CONFLICTS, REFLECTIONS ON STRENGTHS, QUESTIONS ON EXPERIENCES, and QUESTIONS ON EMOTIONS by 5.6%, 8.3%, 3.0%, 3.3%, 6.6%, 9.9%, and 3.2%, as compared to low-quality human therapy. Moreover, therapists from high-quality human therapy conversations less frequently ex-

press REFLECTIONS ON CONSEQUENCES, PROBLEM SOLVING, NORMALIZING, and PSYCHOEDUCATION by 7.7%, 9.2%, 5.4%, and 4.6% than low-quality human therapy.

Moreover, previous psychotherapy research has demonstrated that focusing on solving client problems *before* empathizing may negatively affect relationship forming between therapists and clients (Cochran and Cochran, 2015). Here, we find that therapists from high-quality human therapy conversations respond with REFLECTIONS ON NEEDS 8.8 turns earlier, REFLECTIONS ON EMOTIONS 5.3 turns earlier, REFLECTIONS ON CONSEQUENCES 3.8 turns earlier, REFLECTIONS ON CONFLICTS 8.4 turns earlier, and REFLECTIONS ON STRENGTHS 8.2 turns earlier than low-quality human therapy.

Also, it is important to appropriately respond and adapt to the client's behavior (Bordin, 1979; Horvath and Greenberg, 1989; Miller and Rollnick, 2012). When analyzing adaptability to client behaviors, if clients express GAINED INSIGHTS, therapists from high-quality human therapy respond with 15.0% lower PSYCHOEDUCATION compared to low-quality human therapy. If clients express CHANGING UNHEALTHY BEHAVIOR, therapists from high-quality human therapy respond with 9.6% higher REFLECTIONS ON STRENGTHS than low-quality human therapy, likely to reinforce the positive change and promote self-efficacy in clients. Also, if clients SHARE NEGATIVE EMOTIONS therapists from high-quality human therapy respond with 15.9% higher REFLECTIONS ON EMOTIONS than low-quality human therapy, potentially to foster empathy.

These differences between high-quality and low-quality conversations help us analyze LLMs. We assess how LLM therapists compare against human therapists, determining whether their conversational behavior is closer to high-quality or low-quality human therapists. To establish a baseline, we calculate the average conversational behavior across all human therapy conversations. We consider behaviors aligned with low-quality conversations as below average quality and behaviors aligned with high-quality conversations as above average quality. We then assess the LLM therapists' behavior in relation to this average. This allows us to assess whether the behavior of LLM therapists falls below (*undesirable*) or above (*desirable*) this average, indicating closer alignment with low-quality or high-quality human therapists, respectively. Figure 4.4e provides an overview with example LLM behaviors.

Our null hypothesis is that the behavior of an LLM therapist is similar to that of average-quality human therapy. To test this, we conduct a two-sided Student's t-test. The results reported

below are statistically significant at  $p = \frac{0.05}{m}$ , following Bonferroni correction (m: number of intents tested). We report descriptive statistics in Supplementary Tables C.1-C.6.

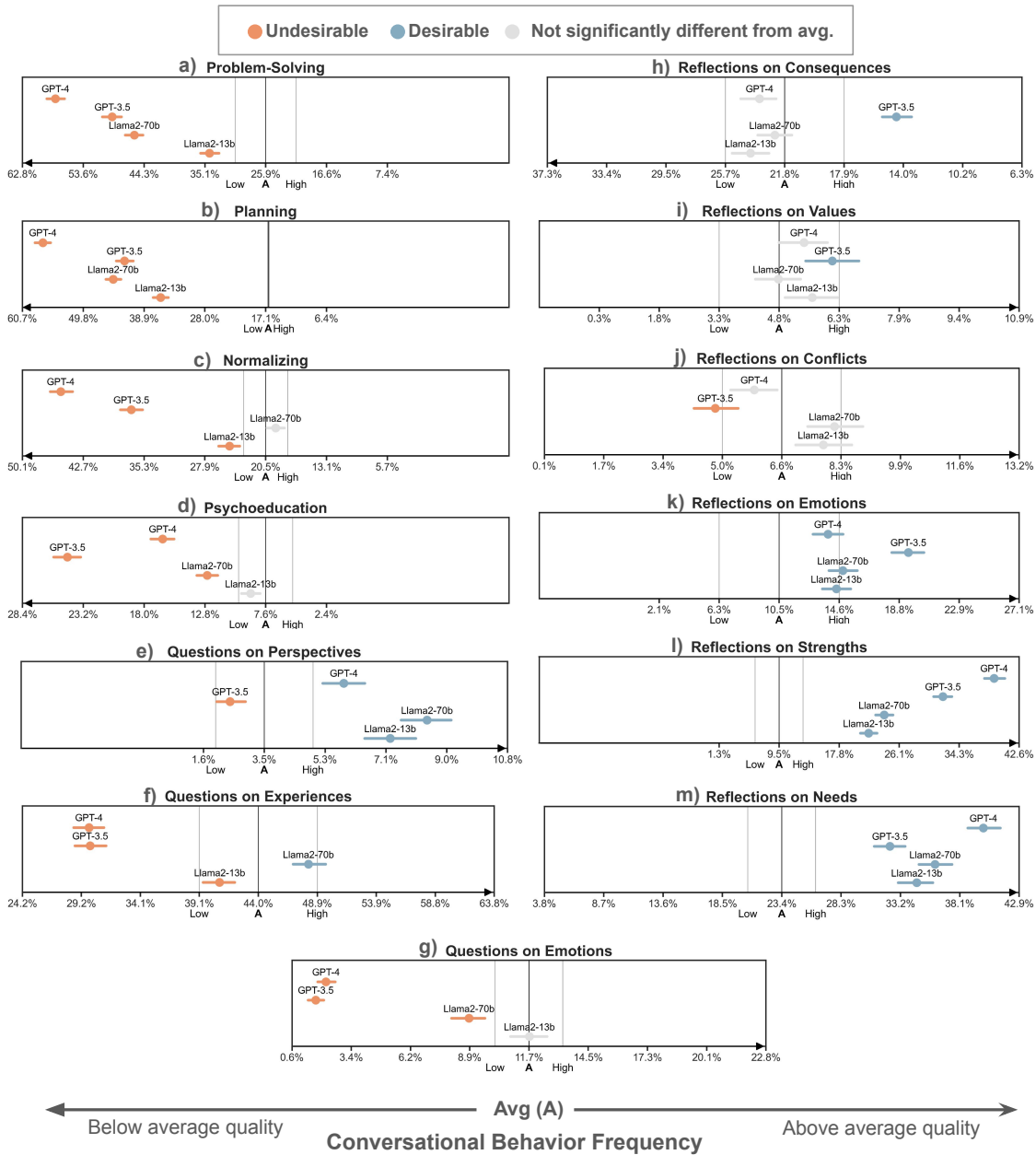
### **Behavior Frequency – How frequently do LLM Therapists exhibit specific conversational behaviors relative to human therapists?**

**Summary.** Based on behavior frequency, LLM therapists providing MI support often resemble therapists in *low-quality* MI sessions (Figure 4.5; Supplementary Table C.1). Both tend to offer more SOLUTIONS compared to average-quality human therapy. However, unlike low-quality human therapy, LLMs tend to emphasize REFLECTIONS.

**LLM therapists respond with significantly higher degrees of Solutions, similar to low-quality human therapy conversations.** Specifically, we find that GPT-4, GPT-3.5, Llama2-70b, and Llama2-13b are more frequent in exhibiting PROBLEM-SOLVING by 31.9%, 23.1%, 19.9%, and 8.5% respectively than average human therapists (Figure 4.5a). This more closely resembles the higher degrees of PROBLEM-SOLVING behavior observed in low-quality therapy, suggesting that such behavior may be undesirable. In fact, GPT-4, GPT-3.5, Llama2-70b, and Llama 2-13b even exceeds the frequency of low-quality therapy by 27.3%, 18.5%, 15.3%, and 3.9% respectively, raising concerns about their potentially overly aggressive approach to problem-solving. To corroborate this finding, we also evaluate the word usage of LLM therapists using LIWC – the gold-standard lexicon-based method for analyzing psychologically relevant constructs in text (Boyd et al., 2022) and find a similar trend significantly more Behavioral Activation (Kanter et al., 2010) words related to identifying goals, problem-solving, and planning (Supplementary Table C.9).

A likely, plausible reason for this problem-solving prominent behavior could be the Reinforcement Learning with Human Feedback (RLHF) alignment that these LLMs incorporate. A key focus of RLHF alignment in LLMs is helping users solve their tasks and provide advice (Bai et al., 2022; Ouyang et al., 2022). This so-called alignment may lead to LLMs overly focusing on offering solutions to client’s problems even when LLMs are used as therapists.

**LLM therapists respond with higher Normalizing and Psychoeducation, similar to low-quality human therapy.** Also, we we find that GPT-4, GPT-3.5, and Llama2-13b are more frequent in exhibiting NORMALIZING by 24.9%, 13.5%, and 4.4% respectively than average human therapists



**Figure 4.5:** Difference in the frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. A: average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in blue indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in orange indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in gray are not statistically significantly different from average-quality at  $p = 0.05$  at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test, following Bonferroni correction (m: number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. A key insight we find is that LLMs respond with significantly higher PROBLEM-SOLVING (subfigure (a)), similar to low-quality human therapy. On the other hand, LLMs respond with significantly higher REFLECTIONS ON STRENGTHS (subfigure (l)), similar to high-quality therapy, but with a frequency that significantly exceeds high-quality therapy.

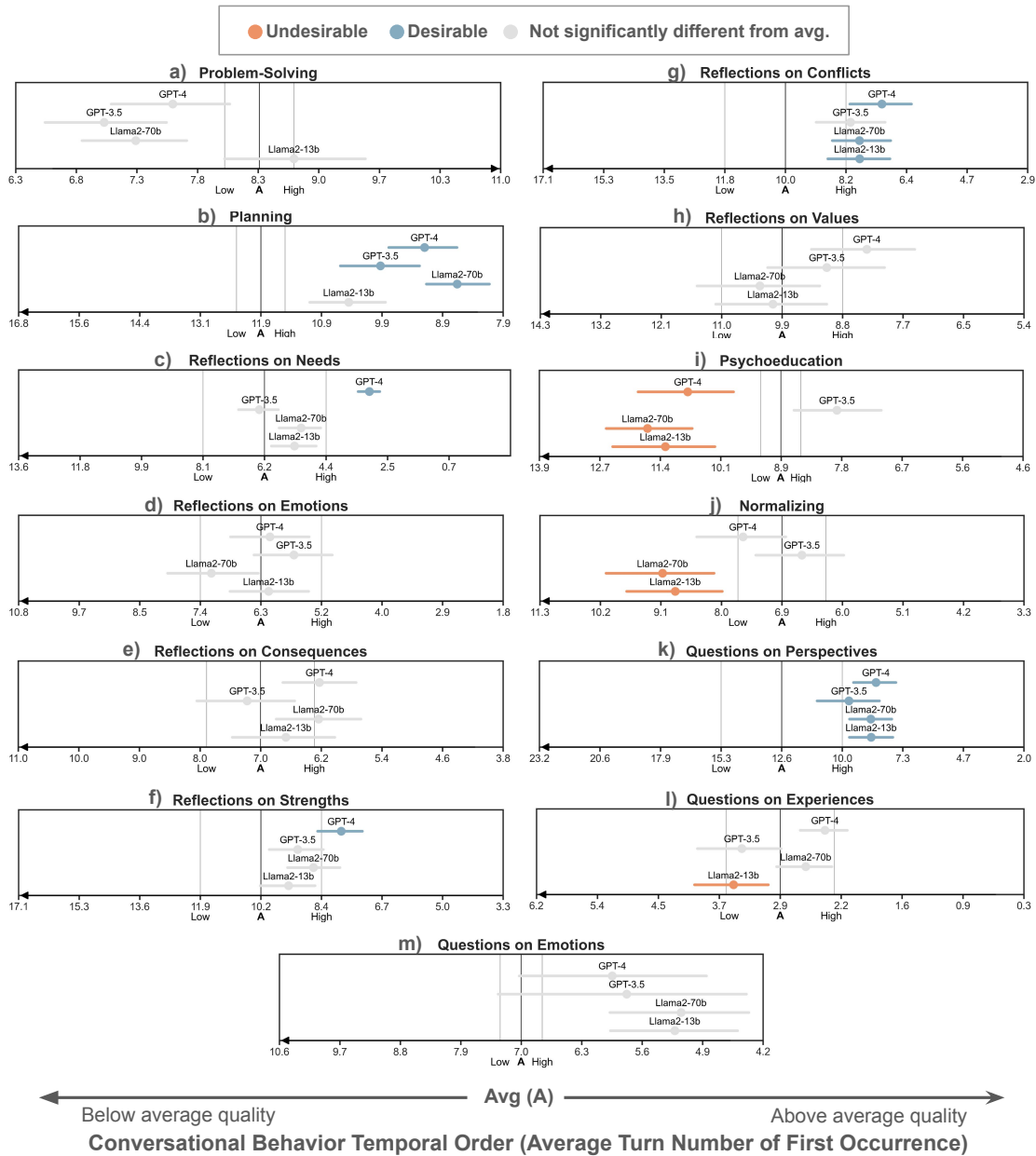
(Figure 4.5c). GPT-4, GPT-3.5, and Llama2-70b are also more frequent in exhibiting PSYCHOEDUCATION by 8.8%, 20.1%, and 5.0% respectively than average human therapists (Figure 4.5d). This more closely resembles the higher degrees of NORMALIZING and PSYCHOEDUCATION observed in low-quality therapy, suggesting that such behavior may be undesirable. Such behavior aligns with previous studies on RLHF data, which have demonstrated that “empathetic” and “informative” LLM generations are considered more helpful and more preferred based on human feedback (Sharma et al., 2023c).

**Most LLM therapists respond with significantly fewer Questions on Experiences and Questions on Emotions, similar to low-quality human therapy.** Among LLM therapists, we observe that GPT-4, GPT-3.5, and Llama2-13b are less frequent in exhibiting QUESTIONS ON EXPERIENCES by 14.2%, 12.9%, and 3.2% than average human therapists (Figure 4.5f). Furthermore, GPT-4, GPT-3.5, and Llama2-70b are less frequent in exhibiting QUESTIONS ON EMOTIONS by 9.5%, 9.8%, and 2.8% than average human therapists (Figure 4.5g). This more closely resembles the lower degrees of QUESTIONS ON EXPERIENCES and QUESTIONS ON EMOTIONS observed in low-quality therapy, suggesting that such behavior may be undesirable.

**LLM therapists respond with more Reflections, unlike low-quality human therapy.** Moreover, all LLM therapists respond with more REFLECTIONS than average-quality human therapy conversations in general (Figures 4.5h-m). We find that GPT-4, GPT-3.5, Llama2-70b, and Llama2-13b are more frequent in exhibiting REFLECTIONS ON EMOTIONS by 3.4%, 7.9%, 4.4%, and 4.0% respectively than average-quality human therapy. This resembles high-quality human therapy suggesting that such a behavior may be desirable. All LLM therapists are also above average-quality therapy in exhibiting REFLECTIONS ON STRENGTHS and REFLECTIONS ON NEEDS. However, their frequency significantly exceeds high-quality therapy as well, raising concerns about the desirability of such behavior.

### **Temporal Order of Behavior – How do LLM therapists structure their conversation compared to human therapists?**

**Summary.** LLM therapists providing MI support exhibit a temporal ordering of behaviors that often aligns with low-quality human MI therapy (Figure 4.6; Supplementary Table C.3). For



**Figure 4.6:** Difference in the *temporal order* of conversational behaviors, operationalized as the turn numbers in which behaviors are first exhibited in a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b), relative to average-, low-, and high-quality human therapy. A: average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the order is increasing (we flip the axis if low-quality is exhibited later than high-quality, such that low-quality is visualized below the average quality marker). Values colored in blue indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in orange indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in gray are not statistically significantly different from average-quality at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test, following Bonferroni correction (m: number of intents tested = 13). Error bars indicate 95% bootstrapped confidence intervals. Most LLM therapists start providing PLANNING (subfigure (b)) earlier in the conversations but provide NORMALIZING (subfigure (j)) later against common recommendations (Cochran and Cochran, 2015).

instance, both tend to offer SOLUTIONS before NORMALIZING.

**Most LLM therapists do not prioritize Normalizing before Solutions, similar to low-quality human therapy.** We compare the order between SOLUTIONS and NORMALIZING exhibited by LLM therapists (Figure 4.6a-b, Figure 4.6j). We find that GPT-4 and Llama2-70b provide PROBLEM-SOLVING 0.5 and 0.9 turns earlier and GPT-4, GPT-3.5, and Llama2-70b provide PLANNING 2.6, 1.8, and 3.2 turns earlier than average-quality human therapy (Figure 4.6a-b). In addition, GPT-4, Llama2-70b, and Llama2-13b exhibit NORMALIZING 1.0, 2.1, and 2.3 turns later than average-quality human therapy. This more closely resembles the earlier expression of SOLUTIONS and later expressions of NORMALIZING in low-quality therapy than high-quality therapy. Importantly, previous psychotherapy studies have shown that the earlier focus on providing SOLUTIONS rather than NORMALIZING may negatively impact the therapeutic relationship both in MI as well as general psychotherapy (Cochran and Cochran, 2015; Hill, 2009). This suggests that LLM therapists often exhibit an undesirable temporal ordering of SOLUTIONS and NORMALIZING.

**Most LLM therapists provide Psychoeducation later, similar to low-quality human therapy.** PSYCHOEDUCATION acts as an effective supplement approach to help clients understand therapy, and be better equipped with coping skills (Bhattacharjee et al., 2011). We find that GPT-4, Llama2-70b, and Llama2-13b first respond with PSYCHOEDUCATION 2.2, 2.6, and 3.0 turns later than therapists from average-quality human therapy respectively (Figure 4.6i). Interestingly, we do not find significant differences for GPT-3.5. This is more similar to the later expressions of PSYCHOEDUCATION in low-quality therapy than high-quality therapy suggesting that such a behavior may be undesirable.

Psychotherapists commonly begin sessions by imparting psychoeducation to set the objectives and to outline client expectations (e.g., *“In this session, we will focus on cognitive restructuring, which is a key technique to help with our negative thinking patterns...”*). However, our qualitative analysis suggests that LLM therapists frequently do not start sessions with such psychoeducation.

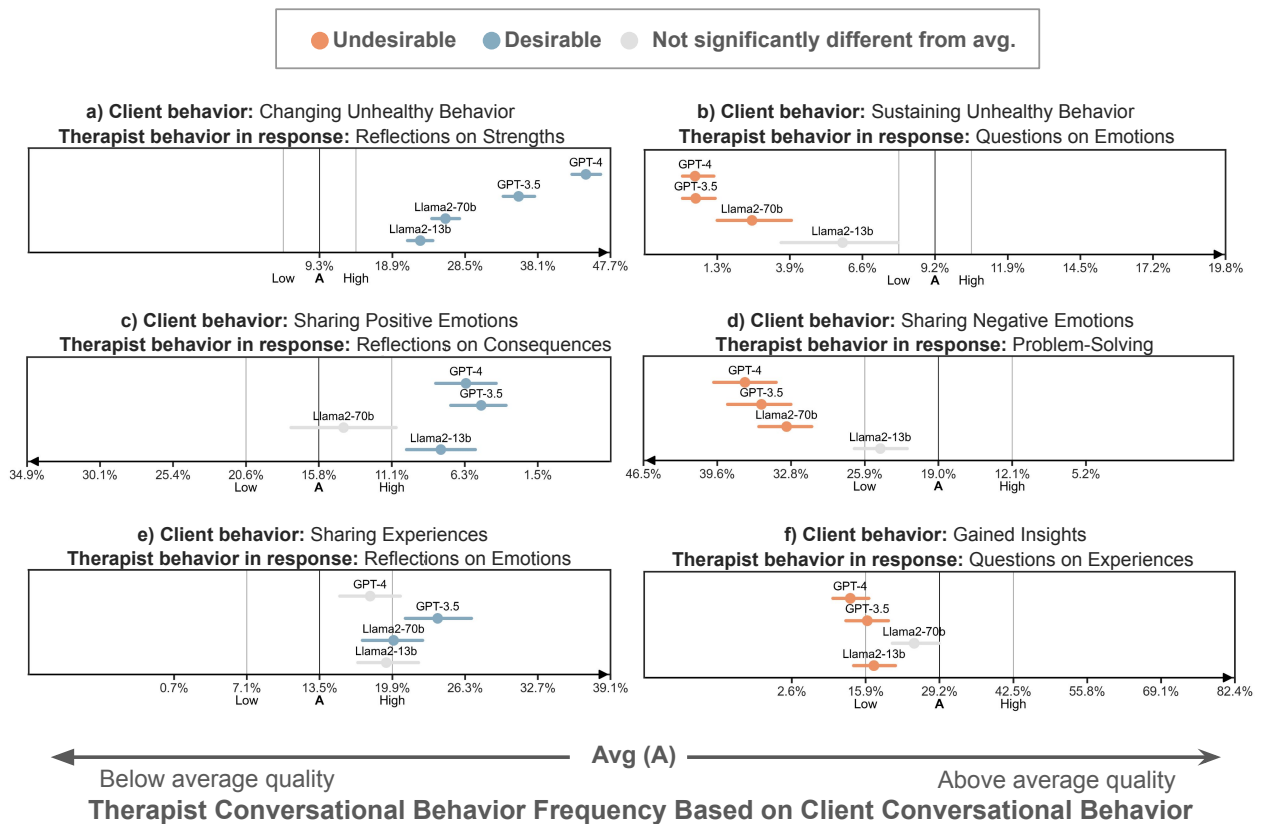
## **Adaptability to Client Behaviors – In what ways do LLM Therapists adapt to different client behaviors compared to human therapists?**

**Summary.** LLM therapists providing MI support respond to some client behaviors similar to low-quality human MI therapy (Figure 4.7; Supplementary Tables C.5-C.6). For example, both respond with more PROBLEM-SOLVING when clients SHARE NEGATIVE EMOTIONS. However, unlike low-quality human therapy, LLM therapists respond with more REFLECTIONS ON STRENGTHS when clients express CHANGING UNHEALTHY BEHAVIOR, which may be more desirable due to their empathetic and validating nature.

**If clients express Changing Unhealthy Behavior, LLM therapists respond with significantly higher Reflections on Strengths, unlike low-quality human therapy.** We observe that GPT-4, GPT-3.5, and Llama2-70b are more frequent in expressing REFLECTIONS ON STRENGTHS by 6.5%, 2.9%, and 2.4% respectively when clients express CHANGING UNHEALTHY BEHAVIOR, compared to average-quality human therapy (Figure 4.7a). This suggests that LLM therapists tend to reinforce the strengths of clients towards changing unhealthy behaviors when clients express a desire or take action to do so. This more closely resembles the higher degrees of REFLECTIONS ON STRENGTHS behavior observed in high-quality therapy than low-quality therapy, suggesting that such behavior may be desirable.

**If clients express Sustaining Unhealthy Behavior, LLM therapists respond with significantly fewer Questions on Emotions, similar to low-quality human therapy.** When clients express SUSTAINING UNHEALTHY BEHAVIOR, we find that GPT-4, GPT-3.5, and Llama2-70b are less frequent in exhibiting QUESTIONS ON EMOTIONS by 8.8%, 8.7%, and 6.7% respectively, compared to average-quality human therapy (Figure 4.7b). This more closely resembles low-quality therapy, suggesting that such behavior may be undesirable, potentially hindering the deeper understanding of the sustenance of unhealthy behavior.

**If clients Share Negative Emotions, LLM therapists respond with higher Problem-Solving, similar to low-quality human therapy.** We find that GPT-4, GPT-3.5, and Llama2-70b are more frequent in expressing PROBLEM-SOLVING by 18.1%, 16.6%, and 14.2% respectively when clients SHARE NEGATIVE EMOTIONS, compared to average-quality human therapy (Figure 4.7d). This



**Figure 4.7:** Difference between the frequency of conversational behaviors observed in LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) or low-quality human therapy *in response to specific client behaviors (Adaptability)*, relative to average-, low-, and high-quality human therapy. **A:** average-quality, Low: low-quality, and High: high-quality therapy. The direction of the arrow on the x-axis indicates the direction in which the frequency is increasing (we flip the axis if low-quality is more frequent than high-quality, such that low-quality is visualized below the average quality marker). Values colored in blue indicate desirable behaviors (significantly closer to high-quality than low-quality) whereas values colored in orange indicate undesirable behaviors (significantly closer to low-quality than high-quality). Values in gray are not statistically significantly different from average-quality at  $p = \frac{0.05}{m}$  using Two-sided Student's t-test, following Bonferroni correction ( $m$ : number of (client, therapist) intents tested =  $13 \times 6 = 78$ ). Error bars indicate 95% bootstrapped confidence intervals. Here, a key finding is that LLMs respond with significantly lower QUESTIONS ON EMOTIONS when clients express SUSTAINING UNHEALTHY BEHAVIOR (subfigure (b)), similar to low-quality human therapy.

more closely resembles low-quality therapy than high-quality suggesting that such a behavior may be undesirable.

**If clients share Gained Insights, all LLM therapists respond with fewer Questions on Experiences, similar to low-quality human therapy.** When clients express QUESTIONS ON EXPERIENCES, we find that GPT-4, GPT-3.5, and Llama2-13b express fewer QUESTIONS ON EXPERIENCES by 16.1%, 13.1%, and 11.8% respectively, compared to average-quality human therapy (Figure 4.7f). Asking fewer QUESTIONS ON EXPERIENCES could indicate lesser engagement with client insights. This more closely resembles low-quality therapy than high-quality therapy suggesting that such a behavior may be undesirable.

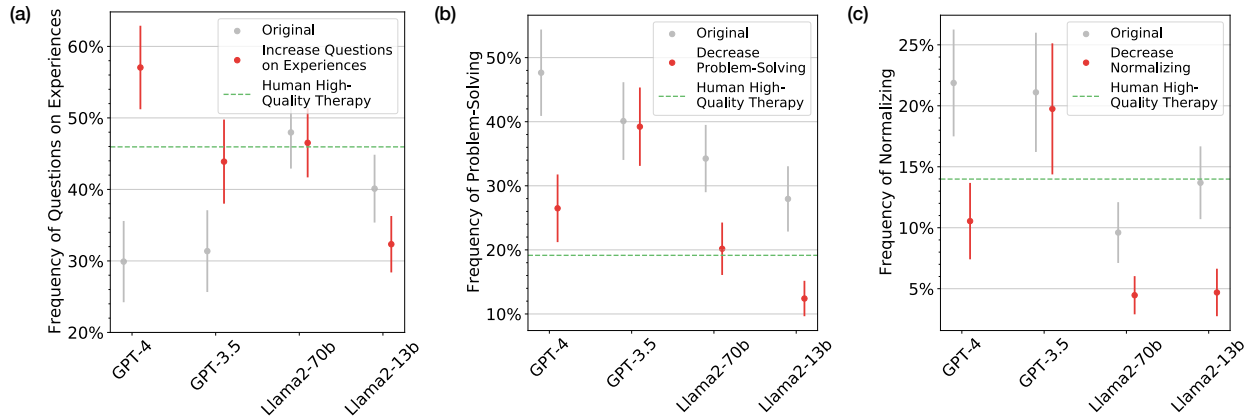
### **Generalizability of Bolt to other contexts**

BOLT is general and could be instantiated to assess different LLM therapists based on varied therapeutic frameworks. Our main analysis and findings are focused on the High-Low Quality dataset, which facilitates assessment within a specific therapeutic approach of Motivational Interviewing and enables the interpretation of findings within a specific therapeutic framework. To test the generalizability of our results, we apply BOLT on the HOPE dataset that includes more general therapy. We find highly similar results for all four of our analyses including the frequency (Supplementary Table C.11), temporal order (Supplementary Table C.13), and adaptability (Supplementary Tables C.15-C.16). This suggests that LLM therapists might be exhibiting similar behavior more broadly. Future work is needed to assess their behavior on more datasets and more types of therapeutic frameworks.

### **Can LLM Behavior be Modulated Towards High-Quality Psychotherapy?**

LLMs are being used to support mental health (Hamilton, 2023; Broderick, 2023; Stade et al., 2024; Youper, 2023; Sharma et al., 2023a,b; Al-Sibai, 2023; Reddit-1, 2023; Twitter, 2023), but the behavior of LLM therapists often differs significantly from high-quality human therapy. Therefore, it is crucial to develop methods that make them more consistent with high-quality care.

Here, we perform one specific experiment to study whether variations in prompts may help in modulating their behavior toward high-quality therapy. We noticed that LLM therapists respond



**Figure 4.8:** We incorporate simple prompts that aim to calibrate LLM therapists, specifically (a) increase QUESTIONS ON EXPERIENCES, (b) decrease PROBLEM-SOLVING, and (c) decrease NORMALIZING. Subfigures show changes in the frequency of conversational behaviors based on changes in prompts to different LLM therapists (GPT-4, GPT-3.5 turbo, Llama2-70b, Llama2-13b). Changes to individual LLMs are shown in pairs (left – Original prompt; right – modulated prompt). The corresponding high-quality human therapy behavior frequency is shown as green dashed lines. For instance, GPT-4 increases the frequency of QUESTIONS ON EXPERIENCES from 29.9% to 57.0% with the modulated prompt. In general, we find that only GPT-4 is able to modulate behavior frequency to a statistically and practically significant amount into the desired direction always, whereas the modulation is inconsistent for other models. Error bars indicate 95% bootstrapped confidence intervals.

less with QUESTIONS ON EXPERIENCES, more on PROBLEM-SOLVING, and more on NORMALIZING than human high-quality therapy (Figure 4.8, Results).

We noticed that LLM therapists respond less with QUESTIONS ON EXPERIENCES, more on PROBLEM-SOLVING, and more on NORMALIZING than human high-quality therapy (Figure 4.8, Results). Here, we vary our original system prompt stated in Methods to: (a) increase QUESTIONS ON EXPERIENCES, by adding an instruction that asks LLMs to “focus more on asking questions to allow client to express their experiences”, (b) decrease PROBLEM-SOLVING, by adding an instruction that asks LLMs to “focus less on offering possible solutions to client’s problem”, and (c) decrease NORMALIZING, by adding an instruction that asks LLMs to “focus less on validating client’s experiences or feelings as normal, on sympathizing with their challenges, and on providing reassurance”. Figure 4.8 shows the effects of this modulation on GPT-4, GPT-3.5 turbo, Llama2-70b, and Llama2-13b.

**Only GPT-4 is able to modulate these behaviors successfully.** We find that only GPT-4 is able to successfully modulate the three behaviors in the intended direction, whereas the behavior modulation of GPT-3.5, Llama2-70b, and Llama2-13b is inconsistent (Figure 4.8). Specifically, GPT-4 increases the occurrence of QUESTIONS ON EXPERIENCES from 29.9% to 57.0%, decreases

the PROBLEM-SOLVING behavior from 47.6% to 26.5%, and decreases NORMALIZING behavior from 21.9% to 10.5%.

**GPT-3.5 is able to increase behavior frequency successfully, but does not decrease behavior frequency.** Certain behaviors (e.g., QUESTIONS), might necessitate an increase in frequency, while others may call for less frequent occurrences (e.g., SOLUTIONS) depending on what is preferred in high-quality care. We observe that GPT-3.5 has better modulation performance when increasing QUESTIONS ON EXPERIENCES – it increases such questions from 31.4% to 43.9%. However, it attains unsatisfactory performance when decreasing PROBLEM-SOLVING (a reduction in frequency from 40.1% to 39.2%) and decreasing NORMALIZING (a reduction in frequency from 21.9% to 20.5%).

**Llama2 variants perform better with the modulation of decreasing a behavior rather than increasing.** Llama2-70b and Llama2-13b, in contrast to GPT-3.5, are able to decrease the frequency of behaviors but not increase them when required for modulation. Specifically, for PROBLEM-SOLVING, Llama2-70b decreases the behavior frequency from 34.3% to 20.2%, and Llama2-13b decreases it from 28.0% to 12.4%. For NORMALIZING, they decrease the frequency from 9.6% to 4.5%, and from 13.7% to 4.7% respectively. However, they fail to successfully increase the frequency of QUESTIONS ON EXPERIENCES. Llama2-70b slightly decreases the behavior frequency from 48.0% to 46.5%, and Llama2-13b decreases the behavior frequency from 40.1% to 32.3%.

Overall, our findings indicate that the ability to modulate the LLM behavior to a statistically and practically significant amount is inconsistent across LLMs. Currently, people using LLMs as therapists may struggle to modulate LLM behaviors to a desired direction always, potentially compromising the quality of care. Even if that was possible, most users do not have the expertise to design effective LLM prompts (Zamfirescu-Pereira et al., 2023), and may not possess a clear understanding of what constitutes quality psychotherapy. Future LLMs with improved prompt tuning may enable better control over behaviors.

### 4.2.3 Discussion

This section discusses our findings' implications for using LLMs for psychotherapy.

Model	Frequency			Order			Adaptability			Aggregate (Average)		
	✓	✗	?	✓	✗	?	✓	✗	?	✓	✗	?
GPT-4	30.8%	46.2%	23.0%	46.2%	7.7%	46.1%	24.3%	42.3%	33.4%	33.7%	32.1%	34.2%
GPT-3.5	38.5%	61.5%	0.0%	23.1%	15.4%	61.5%	24.4%	35.9%	39.7%	28.7%	37.6%	33.7%
Llama2-70b	38.5%	30.8%	30.7%	46.2%	15.4%	38.4%	30.8%	28.2%	41.0%	38.4%	24.8%	36.7%
Llama2-13b	30.8%	30.8%	38.4%	30.8%	23.1%	46.1%	23.1%	20.5%	56.4%	28.2%	24.8%	47.0%

**Table 4.2:** Percentage of analysis in which the models were found to be desirable (✓), undesirable (✗), or uncertain (?). On one hand, in some cases, LLMs demonstrate desirable behaviors resembling high-quality therapy, which could be promising. However, the prevalence of undesirable behaviors resembling low-quality therapy is concerning. These behaviors deviate from established therapeutic guidelines and could affect the overall quality of care.

### More research is needed to ensure quality care

LLM technologies hold great promise to improve the access to mental health support through scalable interventions that can reach a large population (De Choudhury et al., 2023). To illustrate this potential, developers and end users have shared anecdotal examples on social media and other platforms that demonstrate striking similarities between LLMs like ChatGPT and human therapists (e.g., being able to problem-solve (Twitter, 2023) or being heard (Reddit-1, 2023; Reddit-2, 2023)). This is prompting developers and users alike to substitute human therapists and established, evidence-based techniques with them. Many people are now choosing LLM therapists to work through their mental health issues (Al-Sibai, 2023; Reddit-1, 2023; Twitter, 2023).

However, our research suggests that LLM therapists currently often resemble the behavior of therapists in low-quality therapy rather than high-quality therapy sessions. Table 4.2 summarizes the number of instances in which the LLM therapist behavior was found to be desirable (✓), undesirable (✗), or uncertain (?). We can interpret these results from multiple perspectives. On one hand, some behaviors show promise aligning more closely high-quality therapy than low-quality therapy. Conversely, the prevalence of undesirable behaviors closer to low-quality therapy could be a cause for concern. These behaviors deviate from established treatment methods and are discouraged according to psychotherapy recommendations. This may result in inferior and possibly harmful interactions, or unhealthy relationships between clients and LLMs, affecting overall therapeutic alliance and quality of care (Horvath, 2001; Fairburn and Cooper, 2011). Our results suggest that more research is needed to ensure high-quality care, consistent with concerns raised

by experts about the lack of evidence supporting the efficacy of LLMs and the risks associated with their use for mental health support (Stade et al., 2024; De Choudhury et al., 2023; Li et al., 2020; Tate et al., 2023).

### **LLM training prioritizes certain behaviors, but psychotherapy applications exhibit conflicting needs**

Our research identifies behaviors that LLMs currently overemphasize (e.g., PROBLEM-SOLVING) or underemphasize (e.g., asking QUESTIONS), which are actually more closely associated with low-quality therapy sessions. The overemphasized/underemphasized behaviors could have been learned during RLHF which LLM-training organizations use to align LLM behaviors with human preferences and improve model performance (Bai et al., 2022; Ouyang et al., 2022). However, such behaviors may not be preferable across all settings, including not in mental health. Current RLHF alignment focuses on short-term outcomes like the helpfulness of the immediate next response (Bai et al., 2022; Ouyang et al., 2022). However, mental health prioritizes longer-term outcomes like assessing behavioral change across multiple sessions (Lambert, 2013). Such a notion may affect which behaviors are emphasized at different stages of a conversation.

Recent studies advocate for a pluralistic alignment of LLMs to accommodate diverse values and perspectives (Sorensen et al., 2024). In particular, RLHF optimizing for a reward from a single user is insufficient in representing diverse human preferences (Chakraborty et al., 2024) and is insufficient for tailoring LLMs to specific contexts like therapy. This reward is currently based on a “third person without relevant expertise,” which is unlikely to train LLMs to exhibit the behaviors in therapy. This highlights the need for a human-centered approach in building these models that considers the unique requirements of different applications, such as the quality of care in healthcare.

### **Need for public, transparent evaluation methods to inform development, deployment, and adoption of LLM-supported mental health applications**

As LLMs are integrated with new and existing mental health interventions in both commercial (Hamilton, 2023; Broderick, 2023; Youper, 2023) and academic settings (Stade et al., 2024; Sharma

et al., 2023a,b; Hsu et al., 2023; Xu et al., 2023), it is important to have public, transparent, trusted means of evaluating LLMs. Our framework serves as an initial step towards this goal. More work is needed to comprehensively assess LLM use in mental health, particularly in terms of quality, safety, and ethical considerations (De Choudhury et al., 2023; Pendse et al., 2022; Huang et al., 2023; Duffourc and Gerke, 2023; De Freitas et al., 2022; Weidinger et al., 2021). This important line of research could build upon the qualitative perspectives offered by mental health experts and researchers (De Choudhury et al., 2023; Chung et al., 2023; Stade et al., 2024) and quality, safety, and ethical considerations set up in psychotherapy research (Miller and Rollnick, 2012; Lambert, 2013; Fairburn and Cooper, 2011; Barlow et al., 1999; Waltz et al., 1993).

### 4.3 Related Work

Recent research has offered qualitative insights into the opportunities, challenges, and risks associated with using AI systems for mental health (De Choudhury et al., 2023; Chung et al., 2023; Stade et al., 2024; Dash et al., 2023). Our work develops evaluation methods for systematic quantitative assessment of the behavior and quality of AI systems used for mental health support. Researchers have also developed LLM-based systems to assist mental health applications (Hsu et al., 2023; Xu et al., 2023; Perez-Rosas et al., 2023; Pérez-Rosas et al., 2022). Our work serves as an initial step towards how such systems can be more thoroughly evaluated through clinical trials as well as automated assessment.

More broadly, research has assessed the use of LLMs in psychology (Demszky et al., 2023), computational social science (Ziems et al., 2023), and other related applications (Bail, 2023; Clusmann et al., 2023). Also, various works have developed methods to evaluate LLM behavior including their alignment with US demographic groups (Santurkar et al., 2023), investigating the degree to which model responses correspond with user beliefs (Sharma et al., 2023c), evaluating their ability to comprehend psychological emotional stimuli (Li et al., 2023), and studying whether LLM generations are grounded in dialogue acts (Shaikh et al., 2023). Our study aims to further this body of work by evaluating the behavior of LLMs in responding to clients who are seeking mental health support in response to clients seeking mental health support.

## 4.4 Summary of Contribution to Thesis

This chapter delves into the evaluation of efficacy and safety of -AI collaboration systems for mental health. Through a clinical trial with human participants, we showed that AI-based self-guided cognitive restructuring intervention demonstrates superior short-term efficacy and long-term engagement as compared to traditional approaches, while maintaining similar levels of long-term efficacy. We also develop a guiding framework for ethical and safety considerations. Moreover, we proposed BOLT, a novel computational framework to systematically assess the conversational behavior of LLMs when employed as therapists. Our analysis revealed that LLMs often resemble behaviors more commonly exhibited in low-quality therapy rather than high-quality therapy.

# Chapter 5

## Conclusion

### 5.1 Summary of Contributions

Approximately 20% of people worldwide are experiencing mental health problems, but less than half receive any treatment (WHO, 2022; Olfson, 2016). Due to widespread clinician shortages, lengthy waiting lists, and lack of insurance coverage, many vulnerable individuals have limited access to therapy and counseling. In addition, mental health issues are heavily stigmatized, which frequently prevents individuals from seeking appropriate care (Sickel et al., 2014).

My dissertation research aimed to bridge this critical gap between need and access to mental health support. I built and evaluated human-AI collaboration systems to support people who seek mental health support and people who provide mental health support by developing human-centered NLP methods and designing and deploying human-AI interaction.

In Chapter 2, I developed a human-AI collaboration approach for helping untrained peer supporters write more empathic responses. I introduced a new task of *empathic rewriting* to transform low-empathy conversational posts to higher empathy. I proposed PARTNER, a deep reinforcement learning agent that learns to make sentence-level edits to posts to increase the expressed level of empathy while maintaining conversation quality. Through a randomized controlled trial with real-world peer supporters on TalkLife, a large online peer-to-peer support platform, I showed that our human-AI collaboration approach leads to a 19.6% increase in expressed empathy overall and a 39% increase for those who report challenges in providing empathic support.

Fully replacing humans with AI for empathic care has previously drawn skepticism from

psychotherapists (Blease et al., 2020; Doraiswamy et al., 2020). My findings suggest that it is feasible to empower untrained peer supporters with appropriate AI-assisted technologies in relatively lower-risk settings, such as peer-to-peer support (Lee et al., 2021; Vaidyam et al., 2021; Miner et al., 2019; Imel et al., 2015; Kazdin and Blase, 2011). Moreover, empathy is complex and nuanced (Davis et al., 1980; Riess, 2017; Blease et al., 2020; Doraiswamy et al., 2020) and thus presents a greater challenge for AI than many other human–AI collaboration tasks, such as checking grammar in text or scheduling therapy appointments. My thesis develops novel computational techniques to effectively model empathy.

Next, in Chapter 3, I investigated how human-AI collaboration can support people who seek support through self-guided mental health interventions. I took cognitive restructuring, an evidence-based therapeutic technique to overcome negative thinking (Beck, 1976), as a case study. In a large-scale randomized field study on a large mental health website with 15,531 participants, I designed and evaluated a system that uses language models to support people through various steps of cognitive restructuring. My findings revealed that our system positively impacts emotional intensity for 67% of participants and helps 65% overcome negative thoughts. Although adolescents reported relatively worse outcomes, we found that tailored interventions that simplify language model generations improve overall effectiveness and equity.

My findings open up opportunities for improved learning and practicing of key mental health strategies and coping skills through human-AI collaboration interventions. Such interventions could complement traditional treatment options, e.g., by being accessible to users when they have difficulties finding a therapist, or in between sessions. Also, my thesis develops novel computational techniques to effectively model the cognitively difficult process of effectively reframing people’s negative thoughts.

Finally, in Chapter 4, I studied how we can evaluate the efficacy and safety of human-AI collaboration systems for mental health. Through a clinical trial with human participants, I showed that AI-based self-guided cognitive restructuring intervention demonstrates superior short-term efficacy and long-term engagement as compared to traditional approaches, while maintaining similar levels of long-term efficacy. Moreover, I proposed BOLT, a novel computational framework to systematically assess the conversational behavior of LLM therapists. My analysis revealed that LLMs often resemble behaviors more commonly exhibited in low-quality therapy rather than

high-quality therapy.

My thesis contributed two human-AI collaboration systems to support mental health and well-being. The first system helped untrained peer supporters on online peer-to-peer support platforms express empathy more effectively. The second system assisted individuals in self-guided cognitive restructuring of negative thoughts. I demonstrated the effectiveness of these systems through randomized controlled trials with real-world participants in ecologically valid settings. My cognitive restructuring system has been deployed in the real-world by Mental Health America.

Broadly, my thesis makes significant technical contributions to the fields of artificial intelligence, natural language processing, human-computer interaction, psychology, and mental health.

## **5.2 Future Work**

This dissertation made significant progress in developing human-AI collaborative systems to support mental health and well-being. However, several directions remain to be explored to maximize the impact of these systems and facilitate their widespread adoption.

### **5.2.1 Long-term Assistance**

Current human-AI collaboration systems excel as short-term, in-the-moment assistants. As shown in my dissertation work, they can assist individuals in short-term tasks such as rewriting messages to make them more empathetic or suggesting different ways of thinking. However, for mental health and well-being, long-term outcomes are critical. While an intervention might initially regulate emotions and reduce the symptoms of a mental illness, it is the sustained impact over time that truly reveals its effectiveness. Without sustained effects, the symptoms are likely to resurface. Moreover, when used long-term, the intervention needs to incorporate the changing user skill levels over time. This raises the critical question of how to design human-AI collaboration systems that adapt to evolving user contexts over extended periods. Future research could explore how to effectively model long-term user behavior, intentions, and goals. This would enable human-AI collaboration systems that offer a more personalized and adaptive experience by understanding long-term behavior dynamics, ultimately leading to improved mental health

outcomes.

### 5.2.2 Robust Safety Evaluation

Another significant challenge that I have identified during my dissertation work is designing frameworks for more robust safety assessments of human-AI collaboration systems designed for the high-risk setting of mental health. Current AI safety benchmarks primarily focus on broader societal harms like hate speech, discrimination, and misinformation (Ganguli et al., 2022; Gehman et al., 2020). While these are important, they often fail to address the nuanced and individual-level risks that can disproportionately affect people with psychological vulnerabilities. Due to the vulnerability of individuals seeking mental health support, LLM responses, even if detrimental, may be perceived as helpful, leading to “pathological helpfulness” (Oakley et al., 2011) by LLMs that could exacerbate existing mental health conditions or even trigger new ones. Consider a scenario where a person struggling with an eating disorder is interacting with an LLM. If the LLM, perhaps in response to a query about weight loss, provides information or advice that encourages further efforts into weight loss, it could potentially exacerbate their disorder. The person may interpret the LLM’s responses as validation or reinforcement of their harmful beliefs, leading to increased distress and potential escalation of their psychological behaviors. However, similar weight loss advice given to someone without an eating disorder might not have the same negative impact. They may be able to interpret the information more objectively and make informed decisions about their health. Future work could focus on developing more fine-grained safety benchmarks that account for individual vulnerabilities, incorporating diverse user feedback and needs, and establishing ethical guidelines for responsible LLM deployment in sensitive contexts. Such an objective is aligned with growing calls for more comprehensive and nuanced safety evaluations in mental health settings (De Choudhury et al., 2023; Stade et al., 2024).

### 5.2.3 Adapting to Other Applications and Contexts

My thesis modeled key human-AI collaboration behavior for mental health and well-being. Future work could adapt this research to other applications such as providing feedback on writing

and interpersonal communication (Lin et al., 2024; Fu et al., 2024), facilitating skill learning for students (Hofman et al., 2023), correcting misinformation (Zhou et al., 2024), improving workplace practices and productivity (Butler et al., 2023), and fostering creativity (Chakrabarty et al., 2024). Moreover, while my thesis focused on text-based interactions, the underlying principles could be adapted to in-person contexts. For example, our human-AI collaboration system to help people express empathy more effectively could be extended to improve broader in-person interpersonal communication. Similarly, through human-AI collaboration, we can support students in learning different skills, potentially providing personalized feedback and teaching.



# Bibliography

Carol E Adair, Gerald M McDougall, Craig R Mitton, Anthony S Joyce, T Cameron Wild, Alan Gordon, Norman Costigan, Laura Kowalsky, Gloria Pasmeny, and Anora Beckie. 2005. Continuity of care and health outcomes among persons with severe mental illness. *Psychiatric Services*.

Noor Al-Sibai. 2023. *OPENAI EMPLOYEE SAYS SHE'S NEVER TRIED THERAPY BUT CHAT-GPT IS PRETTY MUCH A REPLACEMENT FOR IT*, accessed 2023.

Ashley Batts Allen and Mark R Leary. 2010. Self-compassion, stress, and coping. *Social and personality psychology compass*.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *CHI*.

Joanna J Arch, Georg H Eifert, Carolyn Davies, Jennifer C Plumb Vilaradaga, Raphael D Rose, and Michelle G Craske. 2012. Randomized clinical trial of cognitive behavioral therapy (cbt) versus acceptance and commitment therapy (act) for mixed anxiety disorders. *Journal of consulting and clinical psychology*.

Kyle Arnold. 2014. Behind the mirror: Reflective listening and its tain in the work of carl rogers. *The Humanistic Psychologist*.

- Tal August, Katharina Reinecke, and Noah A Smith. 2022. Generating scientific definitions with controllable complexity. In *ACL*.
- Shelli Avenevoli, Joel Swendsen, Jian-Ping He, Marcy Burstein, and Kathleen Ries Merikangas. 2015. Major depression in the national comorbidity survey–adolescent supplement: Prevalence, correlates, and treatment. *Journal of the American Academy of Child & Adolescent Psychiatry*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Christopher Bail. 2023. Can generative artificial intelligence improve social science?
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *CHI*.
- Federica Barbieri. 2008. Patterns of age-based linguistic variation in american english. *Journal of sociolinguistics*.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *EMNLP Findings*.
- David H Barlow, Jill T Levitt, and Lynn F Bufka. 1999. The dissemination of empirically supported treatments: a view to the future. *Behaviour Research and Therapy*.
- Godfrey T Barrett-Lennard. 1981. The empathy cycle: Refinement of a nuclear concept. *Journal of counseling psychology*.
- C Daniel Batson. 2009. These things called empathy: eight related but distinct phenomena.
- Amit Baumel, Frederick Muench, Stav Edan, and John M Kane. 2019. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. *JMIR*.

- Tom L Beauchamp and James F Childress. 2001. *Principles of biomedical ethics*. Oxford University Press, USA.
- Aaron T Beck. 1976. *Cognitive therapy and the emotional disorders*. International Universities Press.
- Aaron T Beck, Denise D Davis, and Arthur Freeman. 2015. *Cognitive therapy of personality disorders*. Guilford Publications.
- Aaron T Beck, Maria Kovacs, and Arlene Weissman. 1975. Hopelessness and suicidal behavior: An overview. *JAMA*.
- Judith S Beck. 2005. *Cognitive therapy for challenging problems: What to do when the basics don't work*. Guilford Press.
- Christopher G Beevers, Tony T Wells, and Ivan W Miller. 2007. Predicting response to depression treatment: The role of negative cognition. *Journal of Consulting and Clinical Psychology*.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.
- Dipanjana Bhattacharjee, Altul Kumar Rai, Narendra Kumar Singh, Pradeep Kumar, Sanjay Kumar Munda, and Basudeb Das. 2011. Psychoeducation: A measure to strengthen psychiatric treatment. *Delhi Psychiatry Journal*.
- Robert A Bjork. 2017. Creating desirable difficulties to enhance learning. *Carmarthen: Crown House Publishing*.
- Robert A Bjork and Elizabeth L Bjork. 2020. Desirable difficulties in theory and practice. *Journal of Applied research in Memory and Cognition*.
- Ann Blandford, Jo Gibbs, Nikki Newhouse, Olga Perski, Aneesha Singh, and Elizabeth Murray. 2018. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digital health*.
- C Blease, C Locher, M Leon-Carlyle, and M Doraiswamy. 2020. Artificial intelligence and the future of psychiatry: Qualitative findings from a global physician survey. *Digit Health*.

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *ACL*.
- Arthur C Bohart, Robert Elliott, Leslie S Greenberg, and Jeanne C Watson. 2002. Empathy. In John C Norcross, editor, *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients* , (pp. Oxford University Press, xii, New York, NY, US.
- Arthur C Bohart and Leslie S Greenberg. 1997. *Empathy reconsidered: New directions in psychotherapy*. American Psychological Association.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*.
- Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*.
- Judith Borghouts, Elizabeth Eikey, Gloria Mark, Cinthia De Leon, Stephen M Schueller, Margaret Schneider, Nicole Stadnick, Kai Zheng, Dana Mukamel, and Dara H Sorkin. 2021. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *JMIR*.
- Eliane M Boucher, Emily C McNaughton, Nicole Harake, Julia L Stafford, and Acacia C Parks. 2021. The impact of a digital intervention (happify) on loneliness during covid-19: qualitative focus group. *JMIR Mental Health*.
- Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*.
- Alain Brailon and Françoise Taiebi. 2020. Practicing “reflective listening” is a mandatory prerequisite for empathy. *Patient education and counseling*.
- Ryan Broderick. 2023. *People are using AI for therapy, whether the tech is ready for it or not, accessed 2023*.

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *EMNLP*.
- Franziska Burger, Mark A Neerinx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: Extracting schemas from thought records. *PloS one*.
- Brian L Burke, Hal Arkowitz, and Marisa Menchola. 2003. The efficacy of motivational interviewing: a meta-analysis of controlled clinical trials. *Journal of consulting and clinical psychology*.
- Hannah A Burkhardt, George S Alexopoulos, Michael D Pullmann, Thomas D Hull, Patricia A Areán, and Trevor Cohen. 2021. Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. *JMIR*.
- David D Burns. 1980. *Feeling Good: The New Mood Therapy*. Avon New York.
- Andrew C Butler, Jason E Chapman, Evan M Forman, and Aaron T Beck. 2006. The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical psychology review*.
- Jenna Butler, Sonia Jaffe, Nancy Baym, Mary Czerwinski, Shamsi Iqbal, Kate Nowak, Sean Rintel, Abigail Sellen, Mihaela Vorvoreanu, Najeeb G. Abdulhamid, Judith Amores, Reid Andersen, Kagonya Awori, Maxamed Axmed, danah boyd, James Brand, Georg Buscher, Dean Carignan, Martin Chan, Adam Coleman, Scott Counts, Madeleine Daepp, Adam Fourney, Daniel G. Goldstein, Andy Gordon, Aaron L Halfaker, Javier Hernandez, Jake Hofman, Jenny Lay-Flurrie, Vera Liao, Siân Lindley, Sathish Manivannan, Charlton Mcilwain, Subigya Nepal, Jennifer Neville, Stephanie Nyairo, Jacki O'Neill, Victor Poznanski, Gonzalo Ramos, Nagu Rangan, Lacey Rosedale, David Rothschild, Tara Safavi, Advait Sarkar, Ava Scott, Chirag Shah, Neha Parikh Shah, Teny Shapiro, Ryland Shaw, Auste Simkute, Jina Suh, Siddharth Suri, Ioana

- Tanase, Lev Tankelevitch, Adam Troy, Mengting Wan, Ryen W. White, Longqi Yang, Brent Hecht, and Jaime Teevan. 2023. [Microsoft new future of work report 2023](#). Technical Report MSR-TR-2023-34, Microsoft.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*.
- Jie Cao, Michael Tanana, Zac E Imel, Eric Poitras, David C Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *arXiv preprint arXiv:1907.00326*.
- Timothy A Carey, Margaret Carey, Kirsten Stalker, Richard J Mullan, Lindsey K Murray, and Margaret B Spratt. 2007. Psychological change from the inside looking out: A qualitative investigation. *Counselling and Psychotherapy Research*.
- Louis G Castonguay and Clara E Hill. 2017. *How and why are some therapists better than others?: Understanding therapist effects*. American Psychological Association.
- Ana Mari Cauce, Melanie Domenech-Rodríguez, Matthew Paradise, Bryan N Cochran, Jennifer Munyi Shea, Debra Srebnik, and Nazli Baydar. 2002. Cultural and contextual influences in mental health help seeking: a focus on ethnic minority youth. *Journal of consulting and clinical psychology*.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *CHI*.
- Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, A. S. Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *ArXiv*, abs/2402.08925.
- Jonathan H Chen and Steven M Asch. 2017. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N. Engl. J. Med.*
- Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *SIGKDD*.

- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.
- Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857*.
- David A Clark. 2013. Cognitive restructuring. *The Wiley handbook of cognitive behavioral therapy*.
- Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *III*.
- Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Nature Communications Medicine*.
- Jeff L Cochran and Nancy H Cochran. 2015. *The heart of counseling: Counseling skills through therapeutic relationships*. Routledge.
- Simon Coghlan, Kobi Leins, Susie Sheldrick, Marc Cheong, Piers Gooding, and Simon D'Alfonso. 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. *DIGITAL HEALTH*.
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*.
- Sunny Collings and Thomas Niederkrotenthaler. 2012a. Suicide prevention and emergent media: surfing the opportunity.
- Sunny Collings and Thomas Niederkrotenthaler. 2012b. Suicide prevention and emergent media: surfing the opportunity. *Crisis*.
- Pamela Y Collins, Vikram Patel, Sarah S Joestl, Dana March, Thomas R Insel, Abdallah S Daar, Scientific Advisory Board and the Executive Committee of the Grand Challenges on Global Mental Health, Warwick Anderson, Muhammad A Dhansay, Anthony Phillips, Susan Shurin, Mark Walport, Wendy Ewart, Sir John Savill, Isabel A Bordin, E Jane Costello, Maureen Durkin,

- Christopher Fairburn, Roger I Glass, Wayne Hall, Yueqin Huang, Steven E Hyman, Kay Jami-son, Sylvia Kaaya, Shitij Kapur, Arthur Kleinman, Adesola Ogunniyi, Angel Otero-Ojeda, Mu-Ming Poo, Vijayalakshmi Ravindranath, Barbara J Sahakian, Shekhar Saxena, Peter A Singer, and Dan J Stein. 2011. Grand challenges in global mental health. *Nature*.
- Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. ACL.
- Peter Damielson, Robert Audi, Cristina Bicchieri, et al. 2004. *The Oxford handbook of rationality*. Oxford University Press, USA.
- Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, et al. 2023. Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery. *arXiv preprint arXiv:2304.13714*.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.
- Daniel David, Steven Jay Lynn, and Albert Ellis. 2009. *Rational and irrational beliefs: Research, theory, and clinical practice*. Oxford University Press.
- Larry Davidson, Matthew Chinman, Bret Kloos, Richard Weingarten, David Stayner, and Jacob Kraemer Tebes. 1999. Peer support among individuals with severe mental illness: A review of the evidence. *Clinical psychology: Science and practice*.
- Mark H A Davis, Miles P Davis, M Davis, Matthew Davis, Mark Davis, Mm Davis, M Davis, F Davis, H Davis, and I W Davis. 1980. A multidimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology*.
- Ryan Daws. 2020. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. <https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/>. Accessed: 2022-1-25.

- Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.
- Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *CSCW*.
- Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2022. Chatbots and mental health: Insights into the safety of generative ai. *Journal of Consumer Psychology*.
- William N Dember and Larry Penwell. 1980. Happiness, depression, and the pollyanna principle. *Bulletin of the Psychonomic Society*.
- Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margaret Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*.
- Sona Dimidjian, Manuel Barrera Jr, Christopher Martell, Ricardo F Muñoz, and Peter M Lewinsohn. 2011. The origins and current status of behavioral activation treatments for depression. *Annual review of clinical psychology*.
- Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*.
- P Murali Doraiswamy, Charlotte Blease, and Kaylee Bodner. 2020. Artificial intelligence and the future of psychiatry: Insights from a global physician survey. *Artif. Intell. Med.*

- Changming Duan and Clara E Hill. 1996. The current state of empathy research. *Journal of counseling psychology*.
- Mindy Duffourc and Sara Gerke. 2023. Generative ai in health care and liability risks for physicians and safety concerns for patients. *Jama*.
- Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. 2011. Empathy. *Psychotherapy*.
- Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*.
- Christopher G Fairburn and Zafra Cooper. 2011. Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy*.
- Stephen B Fawcett and Leslie Borck-Jameson. 2014. *Learning counseling and problem-solving skills*. Routledge.
- Theresa Fleming, Lynda Bavin, Mathijs Lucassen, Karolina Stasiak, Sarah Hopkins, and Sally Merry. 2018. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. *Journal of medical Internet research*.
- Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuvveer Peri, Derek D Caperton, James Gibson, Michael J Tanana, Panayiotis Georgiou, et al. 2022. Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*.
- Luciano Floridi and Josh Cowls. 2022. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*.
- Liye Fu, Susan R. Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *EMNLP*.
- Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. From text to self: Users' perception of aimc tools on interpersonal communication and self. In *CHI*.

- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askill, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
- Sandra Garrido, Chris Millington, Daniel Cheers, Katherine Boydell, Emery Schubert, Tanya Meade, and Quang Vinh Nguyen. 2019. What works and what doesn't work? a systematic review of digital mental health interventions for depression and anxiety in young people. *Frontiers in psychiatry*.
- Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *WWW*.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realex-icityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP Findings*.
- James Gibson, Doğan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Interspeech*.
- Simon B Goldberg, Tony Rousmaniere, Scott D Miller, Jason Whipple, Stevan Lars Nielsen, William T Hoyt, and Bruce E Wampold. 2016. Do psychotherapists improve with time and experience? a longitudinal analysis of outcomes in a clinical setting. *Journal of counseling psychology*.
- Lizabeth A Goldstein, Abby D Adler Mandel, Robert J DeRubeis, and Daniel R Strunk. 2020. Outcomes, skill acquisition, and the alliance: Similarities and differences between clinical trial and student therapists. *Behaviour research and therapy*.
- Mark A Guadagnoli and Timothy D Lee. 2004. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of motor behavior*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,

- and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.
- Jay Haley. 1992. *Problem-solving therapy*. John Wiley & Sons.
- Jose Hamilton. 2023. *Why Generative AI (LLM) Is Ready for Mental Healthcare, accessed 2023*.
- Louise C Hawkley and John T Cacioppo. 2010. Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of behavioral medicine*.
- Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A probabilistic formulation of unsupervised text style transfer. In *ICLR*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- Liesl M Heinrich and Eleonora Gullone. 2006. The clinical significance of loneliness: A literature review. *Clinical psychology review*.
- Tina Hernandez-Boussard, Selen Bozkurt, John PA Ioannidis, and Nigam H Shah. 2020. Minimar (minimum information for medical ai reporting): developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association*.
- Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.
- Jake M Hofman, Daniel G Goldstein, and David M Rothschild. 2023. Steroids, sneakers, coach: The spectrum of human-ai relationships. *Available at SSRN 4578180*.
- Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*.
- Mohammadreza Hojat, Michael J Vergare, Kaye Maxwell, George Brainard, Steven K Herrine, Gerald A Isenberg, Jon Veloski, and Joseph S Gonnella. 2009. The devil is in the third year: a longitudinal study of erosion of empathy in medical school. *Acad. Med.*

- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.
- Adam O Horvath. 2001. The alliance. *Psychotherapy: Theory, research, practice, training*.
- Adam O Horvath and Leslie S Greenberg. 1989. Development and validation of the working alliance inventory. *Journal of counseling psychology*.
- Ahmed Hosny and Hugo JWL Aerts. 2019. Artificial intelligence for global health. *Science*.
- Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *CHI*.
- Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Julie S Hui, Darren Gergle, and Elizabeth M Gerber. 2018. Introassist: A tool to support writing introductory help requests. In *CHI*.
- Thomas D Hull and Kush Mahan. 2017. A study of asynchronous mobile-enabled sms text psychotherapy. *Telemedicine and e-Health*.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *CoRR abs/1905.01969*. *External Links: Link Cited by*.

- Natalie E Hundt, Joseph Mignogna, Cathy Underhill, and Jeffrey A Cully. 2013. The relationship between use of cbt skills and depression treatment outcome: A theoretical and methodological review of the literature. *Behavior therapy*.
- Zac E Imel, Mark Steyvers, and David C Atkins. 2015. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*.
- Kokil Jaidka, Niyati Chhaya, Saran Mumick, Matthew Killingsworth, Alon Halevy, and Lyle Ungar. 2020. Beyond positive emotion: Deconstructing happy moments based on writing prompts. In *ICWSM*.
- Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *EMNLP*.
- Jonathan W Kanter, Rachel C Manos, William M Bowe, David E Baruch, Andrew M Busch, and Laura C Rusch. 2010. What is behavioral activation?: A review of the empirical literature. *Clinical psychology review*.
- Berton H Kaplan, John C Cassel, and Susan Gore. 1977. Social support and health. *Medical care*.
- Carole A Kaplan, Anne E Thompson, and Sheila M Searson. 1995. Cognitive behaviour therapy in children and adolescents. *Archives of disease in childhood*.
- Eirini Karyotaki, Annet Kleiboer, Filip Smit, David Trevor Turner, A Mira Pastor, Gerhard Andersson, Thomas Berger, Cristina Botella, JM Breton, Per Carlbring, et al. 2015. Predictors of treatment dropout in self-guided web-based interventions for depression: an ‘individual patient data’ meta-analysis. *Psychological medicine*.
- Alan E Kazdin and Stacey L Blase. 2011. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspect. Psychol. Sci*.
- Alan E Kazdin and Sarah M Rabbitt. 2013. Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science*.

- Ryan Kelly, Daniel Gooch, and Leon Watts. 2018. 'it's more like a letter': An exploration of mediated conversational effort in message builder. *Proc. ACM Hum.-Comput. Interact.*
- Vivien Kemp and Anthony R Henderson. 2012. Challenges faced by mental health peer support workers: peer support from the peer supporter's point of view. *Psychiatric rehabilitation journal.*
- Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *IJCNLP*.
- Rachel Kornfield, Jonah Meyerhoff, Hannah Studd, Ananya Bhattacharjee, Joseph Jay Williams, Madhu Reddy, and David C Mohr. 2022. Meeting users where they are: User-centered design of an automated text messaging tool to support the mental health of young adults. In *CHI*.
- Rachel Kornfield, Caitlin A Stamatis, Ananya Bhattacharjee, Bei Pang, Theresa Nguyen, Joseph J Williams, Harsh Kumar, Sarah Popowski, Miranda Beltzer, Christopher J Karr, et al. 2023. A text messaging intervention to support the mental health of young adults: User engagement and feedback from a field trial of an intervention prototype. *Internet Interventions*.
- Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *ACL*.
- Kaylee P Kruzan, Ellen E Fitzsimmons-Craft, Mallory Dobias, Jessica L Schleider, and Abhishek Pratap. 2022a. Developing, deploying, and evaluating digital mental health interventions in spaces of online help-and information-seeking. *Procedia computer science*.
- Kaylee Payne Kruzan, Jonah Meyerhoff, Theresa Nguyen, Madhu Reddy, David C Mohr, and Rachel Kornfield. 2022b. "i wanted to see how bad it was": Online self-screening as a critical transition point among young adults with common mental health conditions. In *CHI*.
- Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman AS Farb, and Joseph Jay Williams. 2023. Exploring the use of large language models for improving the awareness of mindfulness. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*.
- Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *ACL-IJCNLP Findings*.

- Michael J Lambert. 2013. *Bergin and Garfield's handbook of psychotherapy and behavior change*. John Wiley & Sons.
- Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging*.
- Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathleen McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.
- Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xingxu Xie. 2023. [Large language models understand and can be enhanced by emotional stimuli](#).
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *EMNLP*.
- Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *ACL*.
- Ron C Li, Steven M Asch, and Nigam H Shah. 2020. Developing a delivery science for artificial intelligence in healthcare. *NPJ digital medicine*.

- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.
- Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered mental health stigma in masked language models. In *EMNLP*.
- Inna Wanyin Lin, Ashish Sharma, Christopher Michael Rytting, Adam S Miner, Jina Suh, and Tim Althoff. 2024. Imbue: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556*.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP*.
- Marsha Linehan. 2014. *DBT? Skills training manual*. Guilford Publications.
- Marsha M Linehan, Kathryn E Korlund, Melanie S Harned, Robert J Gallop, Anita Lungu, Andrada D Neacsiu, Joshua McDavid, Katherine Anne Comtois, and Angela M Murray-Gregory. 2015. Dialectical behavior therapy for high suicide risk in individuals with borderline personality disorder: a randomized clinical trial and component analysis. *JAMA psychiatry*.
- Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022a. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.
- Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike

- Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*.
- Lars-Gunnar Lundh. 2019. Three modes of psychotherapy and their requisite core skills. *Counseling and Psychotherapy Research*.
- Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *IJCAI*.
- David D Luxton, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: a public health perspective. *American journal of public health*.
- Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. 2017. A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods. *Internet interventions*.
- Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. *EMNLP*.
- Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczós, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *ACL*.
- Candelaria I Mahlke, Ute M Krämer, Thomas Becker, and Thomas Bock. 2014. Peer support in mental health services. *Current opinion in psychiatry*.
- Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A Smith, and James Henderson. 2020. Plug and play autoencoders for conditional text generation. In *EMNLP*.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*.

- Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *WSDM*.
- Nicole Martinez-Martin, Karola Kreitmair, et al. 2018. Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. *JMIR mental health*.
- Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *CHI*.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*.
- William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.
- William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.
- Adam S Miner, Albert Haque, Jason A Fries, Scott L Fleming, Denise E Wilfley, G Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A Arnow, W Stewart Agras, et al. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digital Medicine*.
- Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in psychiatry*.
- Christine Moberg, Andrea Niles, and Dale Beermann. 2019. Guided self-help works: randomized waitlist controlled trial of pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. *JMIR*.

- David C Mohr, Aaron R Lyon, Emily G Lattie, Madhu Reddy, and Stephen M Schueller. 2017. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. *JMIR*.
- Robert R Morris and Rosalind Picard. 2012. Crowdsourcing collective emotional intelligence. *arXiv preprint arXiv:1204.3481*.
- Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *JMIR*.
- Theresa B Moyers, Tim Martin, Jennifer K Manuel, Stacey ML Hendrickson, and William R Miller. 2005. Assessing competence in the use of motivational interviewing. *Journal of substance abuse treatment*.
- Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *WWW*.
- J A Naslund, K A Aschbrenner, L A Marsch, and S J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiol. Psychiatr. Sci.*
- Richard Nelson-Jones. 2013. Practical counselling and helping skills: text and activities for the lifeskills counselling model. *Practical Counselling and Helping Skills*.
- Paula Nunes, Stella Williams, Bidyadhar Sa, and Keith Stevenson. 2011. A study of empathy decline in students from five health disciplines during their first year of training. *J. Int. Assoc. Med. Sci. Educ.*
- Barbara Oakley, Ariel Knafo, Guruprasad Madhavan, and David Sloan Wilson. 2011. *Pathological altruism*. Oxford University Press.
- Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*.
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.

- Lars-Göran Öst, Anna Karlstedt, and Sara Widén. 2012. The effects of cognitive behavior therapy delivered by students in a psychologist training program: An effectiveness study. *Behavior Therapy*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Shireen Patel, Athfah Akhtar, Sam Malins, Nicola Wright, Emma Rowley, Emma Young, Stephanie Sampson, and Richard Morriss. 2020. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. *JMIR*.
- Sundip Patel, Alexis Pelletier-Bui, Stephanie Smith, Michael B Roberts, Hope Kilgannon, Stephen Trzeciak, and Brian W Roberts. 2019. Curricula for empathy and compassion training in medical education: A systematic review. *PloS one*.
- Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*. Harvard University Press.
- Sachin R Pendse, Kate Niederhoffer, and Amit Sharma. 2020. Cross-cultural differences in the use of online mental health support forums. *CSCW*, 14.
- Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From treatment to healing: envisioning a decolonial digital mental health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
- Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the effects of technological writing assistance for support providers in online mental health community. In *CHI*.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *ACL*.

- Veronica Perez-Rosas, Ken Resnicow, Rada Mihalcea, et al. 2023. Verve: Template-based reflective rewriting for motivational interviewing. In *EMNLP*.
- Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, et al. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *EMNLP*.
- Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *ACL*.
- Erika S Poole. 2013. Hci and mobile health interventions: how human–computer interaction can contribute to successful mobile health interventions. *Translational behavioral medicine*.
- Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *CHI*.
- Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *AAAI*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. [https://d4mucfpksywv.cloudfront.net/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf). Accessed: 2022-1-25.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*.
- Shanaya Rathod, Narsimha Pinninti, Muhammed Irfan, Paul Gorczynski, Pranay Rathod, Lina

- Gega, and Farooq Naeem. 2017. Mental health service provision in low- and Middle-Income countries. *Health Serv Insights*.
- Erik Rautalinko, Hans-Olof Lisper, and Bo Ekehammar. 2007. Reflective listening in counseling: effects of training time and evaluator social skills. *American journal of psychotherapy*.
- Reddit-1. 2023. *Using ChatGPT as a therapist?, accessed 2023*.
- Reddit-2. 2023. *ChatGPT is better than my therapist, holy shit, accessed 2023*.
- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *ACL*.
- Stefan Rennick-Egglestone, Sarah Knowles, Gill Toms, Penny Bee, Karina Lovell, and Peter Bower. 2016. Health technologies' in the wild' experiences of engagement with computerised cbt. In *CHI*.
- Jordan P Richardson, Cambray Smith, Susan Curtis, Sara Watson, Xuan Zhu, Barbara Barry, and Richard R Sharp. 2021. Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digit Med*.
- Helen Riess. 2017. The science of empathy. *J Patient Exp*.
- Elliot Robert, Arthur C Bohart, JC Watson, and LS Greenberg. 2011. Empathy. *Psychotherapy*.
- Koustuv Saha and Amit Sharma. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. In *ICWSM*.
- Luzette A Samargia, Elizabeth M Saewyc, and Barbara A Elliott. 2006. Foregone mental health care and self-reported access barriers among adolescents. *The Journal of School Nursing*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. *Whose opinions do language models reflect? ArXiv, abs/2303.17548*.
- Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. Inmt: Interactive neural machine translation prediction. In *EMNLP (System Demonstrations)*.

- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.
- David Satcher. 2001. *Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the surgeon general*. US Department of Health and Human Services.
- Jessica L Schleider, Mallory L Dobias, Jenna Y Sung, and Michael C Mullarkey. 2020. Future directions in single-session youth mental health interventions. *Journal of Clinical Child & Adolescent Psychology*.
- Jessica L Schleider, Michael C Mullarkey, Kathryn R Fox, Mallory L Dobias, Akash Shroff, Erica A Hart, and Chantelle A Roulston. 2022. A randomized trial of online single-session interventions for adolescent depression during covid-19. *Nature Human Behaviour*.
- Jessica Schroeder, Jina Suh, Chelsey Wilks, Mary Czerwinski, Sean A Munson, James Fogarty, and Tim Althoff. 2020. Data-driven implications for translating evidence-based psychotherapies into technology-delivered interventions. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*.
- Jessica Schroeder, Chelsey Wilkes, Kael Rowan, Arturo Toledo, Ann Paradiso, Mary Czerwinski, Gloria Mark, and Marsha M Linehan. 2018. Pocket skills: A conversational mobile web app to support dialectical behavioral therapy. In *CHI*, pages 1–15.
- Stephen M Schueller, Martha Neary, Jocelyn Lai, and Daniel A Epstein. 2021. Understanding people’s use of and perspectives on mood-tracking apps: interview study. *JMIR*.
- Craig S Schwalbe, Hans Y Oh, and Allen Zweben. 2014. Sustaining motivational interviewing: a meta-analysis of training studies. *Addiction*.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL-HLT*.
- Robert L Selman. 1980. *Growth of interpersonal understanding*. Academic Press.
- Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E

- Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *CSCW*.
- Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023. Grounding or guesswork? large language models are presumptive grounders. *arXiv preprint arXiv:2311.09144*.
- Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *ICWSM*.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW/TheWebConf*.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023a. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*.
- Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP*.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023b. Cognitive reframing of negative thoughts through human-language model interaction. In *ACL*.
- Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *CHI*.
- Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *CHI*.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023c. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

- Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *ACL*, pages 3096–3107.
- Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *SIGDIAL*.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NeurIPS*.
- Jane Shkel, Gavin Green, Stacey Le, Benjamin Kaveladze, Veronique Marcotte, Kevin Rushton, Theresa Nguyen, and Stephen M Schueller. 2023. Understanding users’ experiences of a novel web-based cognitive behavioral therapy platform for depression and anxiety: Qualitative interviews from pilot trial participants. *JMIR Formative Research*.
- Amy E Sickel, Jason D Seacat, and Nina A Nabors. 2014. Mental health stigma update: A review of consequences. *Advances in Mental Health*.
- Wendy K Silverman, William M Kurtines, Golda S Ginsburg, Carl F Weems, Peyton White Lumpkin, and Delight Hicks Carmichael. 1999. Treating anxiety disorders in children with group cognitive-behavioral therapy: A randomized clinical trial. *Journal of consulting and clinical psychology*.
- C Estelle Smith, William Lane, Hannah Miller Hillberg, Daniel Kluver, Loren Terveen, and Svetlana Yarosh. 2021. Effective strategies for crowd-powered cognitive reappraisal systems: A field deployment of the flip\* doubt web application for mental health. *CSCW*.
- Diana MY Smith, Sarah M Lipson, Shirley B Wang, and Kathryn R Fox. 2022a. Online methods in adolescent self-injury research: Challenges and recommendations. *Journal of Clinical Child & Adolescent Psychology*.
- Wally Smith, Greg Wadley, Sarah Webber, Benjamin Tag, Vassilis Kostakos, Peter Koval, and James J Gross. 2022b. Digital emotion regulation in everyday life. In *CHI*.

- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*. Cambridge, MA.
- Nili Solomonov, Kevin S McCarthy, Bernard S Gorman, and Jacques P Barber. 2019. The multi-theoretical list of therapeutic interventions–30 items (multi-30). *Psychotherapy Research*.
- Jiyoung Song, Boris Litvin, Ryan Allred, Shiyu Chen, Thomas D Hull, and Patricia A Areán. 2023. Comparing message-based psychotherapy to once-weekly, video-based psychotherapy for moderate depression: Randomized controlled trial. *JMIR*.
- Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.
- Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*.
- Mark A Stebnicki. 2007. Empathy fatigue: Healing the mind, body, and spirit of professional counselors. *Am. J. Psychiatr. Rehabil.*
- Ian Stewart, Charles Welch, Lawrence An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2023. Expressive interviewing agents to support health-related behavior change: A study of covid-19 behaviors. *JMIR formative research*.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *Computational Linguistics*, pages 339–374.

- Daniel R Strunk, Shannon N Hollars, Abby D Adler, Lizabeth A Goldstein, and Justin D Braun. 2014. Assessing patients' cognitive therapy skills: Initial evaluation of the competencies of cognitive therapy scale. *Cognitive therapy and research*.
- Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *J. Subst. Abuse Treat.*
- Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *JMIR*.
- Steven Tate, Sajjad Fouladvand, Jonathan H Chen, and Chwen-Yuen Angie Chen. 2023. The chatgpt therapist will see you now: Navigating generative artificial intelligence's potential in addiction medicine research and patient care.
- John Torous, Jessica Lipschitz, Michelle Ng, and Joseph Firth. 2020. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. *Journal of affective disorders*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- CB Truax and RR Carkhuff. 1967. Modern applications in psychology. *Toward effective counseling and psychotherapy: Training and practice*. Hawthorne, NY, US: Aldine Publishing Co.
- Kadi Tulver, Karl Kristjan Kaup, Ruben Laukkonen, and Jaan Aru. 2023. Restructuring insight: An integrative review of insight in problem-solving, meditation, psychotherapy, delusions and psychedelics. *Consciousness and cognition*.
- Twitter. 2023. *GPT is a better therapist than any therapist I've ever tried, accessed 2023*.

- Aditya Nrusimha Vaidyam, Danny Linggonegoro, and John Torous. 2021. Changes to the psychiatric chatbot landscape: A systematic review of conversational agents in serious mental illness: Changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse. *Can. J. Psychiatry*.
- T Michael Vallis, Brian F Shaw, and Keith S Dobson. 1986. The cognitive therapy scale: psychometric properties. *Journal of consulting and clinical psychology*.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- Dave Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *ICWSM*.
- Jennifer Waltz, Michael E Addis, Kelly Koerner, and Neil S Jacobson. 1993. Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of consulting and clinical psychology*.
- Zijian Wang and David Jurgens. 2018. It’s going to be okay: Measuring access to support in online communities. In *EMNLP*.
- Jeanne C Watson. 2007. Facilitating empathy. *European Psychotherapy*.
- Jeanne C Watson, Rhonda N Goldman, and Margaret S Warner. 2002. *Client-centered and experiential psychotherapy in the 21st century: Advances in theory, research, and practice*. PCCS Books.
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*.
- Charles Welch, Allison Lahnala, Verónica Pérez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.
- Robert West and Eric Horvitz. 2019. Reverse-engineering satire, or “paper on computational humor accepted despite making serious advances”. In *AAAI*.
- WHO. Mental disorders. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>. Accessed: 2022-1-25.
- WHO. 2022. Mental health and covid-19: early evidence of the pandemic’s impact: scientific brief, 2 march 2022. Technical report, World Health Organization.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.
- Coralie J Wilson and Frank P Deane. 2012. Brief report: Need for autonomy and other perceived barriers relating to adolescents’ intentions to seek professional mental health care. *Journal of adolescence*.
- M J Wolf, K Miller, and F S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft’s tay “experiment,” and wider implications. *ACM SIGCAS Comput. Soc.*
- Jesse H Wright, Jesse J Owen, Derek Richards, Tracy D Eells, Thomas Richardson, Gregory K Brown, Marna Barrett, Mary Ann Rasku, Geneva Polser, and Michael E Thase. 2019. Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis. *The Journal of clinical psychiatry*.
- Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. *ACL*.

- Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.
- Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *CHI*.
- Youper. 2023. *Mental Health GPTs, accessed 2023*.
- JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *CHI*.
- Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *ACL*.
- Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019a. Finding your voice: The linguistic development of mental health counselors. In *ACL*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.
- Xinyi Zhou, Ashish Sharma, Amy X Zhang, and Tim Althoff. 2024. Correcting misinformation on social media with a large language model. *arXiv preprint arXiv:2403.11169*.
- Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *SIGdial*.
- Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *ACL*.
- Caleb Ziems, Omar Shaikh, Zhehao Zhang, William Held, Jiaao Chen, and Diyi Yang. 2023. Can large language models transform computational social science? *Computational Linguistics*.



# Chapter A

## Supplementary Materials for Human-AI Collaboration to Empower People Providing Support

### A.1 Data Collection Details

#### Annotation Instructions

For each (seeker post, response post) pair, the annotators were asked the following four questions:

1. **(Mental Health Related)** Is the seeker talking about a mental health related issue or situation in his/her post?<sup>1</sup>
  - Yes
  - No
2. **(Emotional Reactions)** Does the response express or allude to warmth, compassion, concern or similar feelings of the responder towards the seeker?
  - No
  - Yes, the response alludes to these feelings but the feelings are not explicitly expressed

---

<sup>1</sup>We use this question for filtering non-mental related posts from the data collection process

- Yes, the response has an explicit mention of these feelings

3. **(Interpretations)** Does the response communicate an understanding of the seeker's experiences and feelings? In what manner?

- No
- Yes, the response communicates an understanding of the seeker's experiences and/or feelings

If the answer to the above question was "Yes", the annotators were further asked to annotate one or more of the following:

- The response contains conjectures or speculations about the seeker's experiences and/or feelings
- The responder has reflected back on similar experiences of their own or others
- The responder has also described similar experiences of their own or others
- The response contains paraphrases of the seeker's experiences and/or feelings

4. **(Explorations)** Does the response make an attempt to explore the seeker's experiences and feelings?

- No
- Yes, but the exploration is generic
- Yes, and the exploration is specific

The detailed instructions can be found at <https://mhannotate-test.cs.washington.edu/annotate/readme.html>.

## Interactive Training of Crowdworkers

The crowdworkers on Upwork were initially provided with our entire annotation instructions and an interactive training system<sup>2</sup> containing ten examples. After this initially automated training, we scheduled a 1hour long phone call with them to discuss our annotation instructions

---

<sup>2</sup>This system contained prompts of manually written feedback for both correct and incorrect annotations.

and annotation interface. During the phone call, crowdworkers also asked questions on the annotation guidelines which greatly helped in addressing potential ambiguities. After the phone call, we assigned them 20 tasks each (randomly chosen; different for each crowdworker). We manually evaluated the annotations on those 20 tasks. Based on the evaluation, we either decided to discontinue with the crowdworker (there were two such crowdworkers) or we provided them further manual feedback. Throughout the process, crowdworkers actively asked questions through the chat feature on Upwork. After the initial training phase, we also did spot checks on quality (at least two times for each crowdworker;  $\geq 20$  posts each) to provide them further feedback.<sup>3</sup>

## A.2 Reproducibility

### Implementation Details

**Code.** Our codes are based on the huggingface library (<https://huggingface.co/>). We make them publicly available at <https://github.com/behavioral-data/Empathy-Mental-Health>.

**Seed Value.** For all our experiments, we used the seed value of 12.

### Hyperparameter Fine-tuning

We searched through the following space of hyperparameters for fine-tuning our model:

- learning rate = 1e-5, 2e-5, 5e-5, 1e-4, 5e-4

<sup>3</sup>Crowdworkers only needed minor feedback on these posts.

	Data Source	Train			Dev			Test		
		No	Weak	Strong	No	Weak	Strong	No	Weak	Strong
Emotional Reactions	TalkLife	52.02%	41.55%	6.43%	49.44%	44.66%	5.90%	52.28%	41.27%	6.45%
	Reddit	65.80%	29.52%	4.68%	66.87%	26.88%	6.25%	66.98%	27.39%	5.63%
Interpretations	TalkLife	78.39%	3.33%	18.28%	77.20%	4.00%	18.80%	79.26%	2.69%	18.04%
	Reddit	54.59%	3.63%	41.77%	48.12%	4.37%	47.5%	48.83%	3.91%	47.26%
Explorations	TalkLife	72.87%	10.56%	16.57%	73.88%	10.11%	16.01%	73.40%	11.09%	15.51%
	Reddit	83.41%	3.80%	12.79%	89.94%	62.89%	9.44%	85.60%	3.13%	11.27%

**Table A.1:** Train/Dev/Test Splits.

- $\lambda_{EI} = 1$
- $\lambda_{RE} = 0.1, 0.2, 0.5, 1$

## Runtime Analysis

**Domain-Adaptive Pre-training Time.** We conducted domain-adaptive pre-training on four RTX 2080 Ti GPUs. Pre-training S-Encoder took around 22 hours. Pre-training R-Encoder took around 38 hours. Both are pre-trained for three epochs.

**Model Training Time.** We trained our model on one RTX 2080 Ti GPU. The training approximately takes five minutes. Our model is trained for four epochs.

## Train, Dev, Test Splits

We split both the datasets into train, dev, and test sets in the ratio of 75:5:20. Table A.1 contains the statistics of the train, dev, and test splits.

## Number of Parameters

The total number of parameters of our model = 2 \* number of parameters of RoBERTa<sub>BASE</sub> + parameters in the linear layers  $\approx 2*125M + 2 * .5M = 251M$

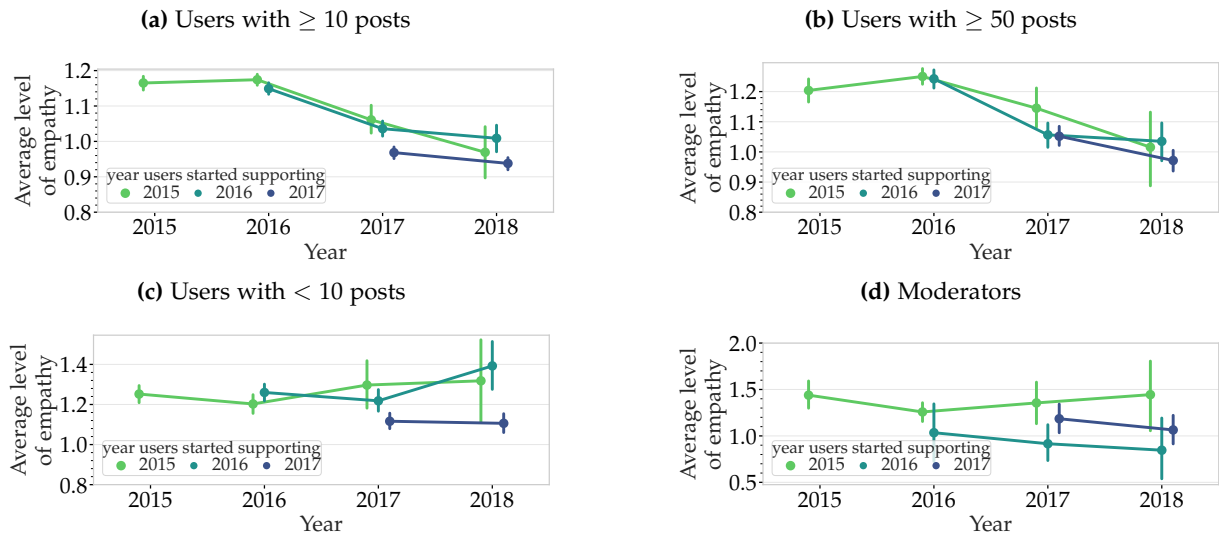
## Reddit dataset

The entire Reddit dataset can be accessed through its archive on Google BigQuery at [https://bigquery.cloud.google.com/table/fh-bigquery:reddit\\_comments.2015\\_05?pli=1](https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2015_05?pli=1)

## A.3 Potential Confounding Factors in Analysis of Variation of Empathy over Time

We note that such an analysis can be affected by several confounding factors such as old vs. new users, user dropout, and low activity of several users. To account for these factors, we stratify users by the year in which they started supporting on the platform (2015, 2016, 2017) and analyze

**Figure A.1:** Empathy over time analysis of various user groups. We find similar trends across multiple groups.



the average levels of empathy during subsequent years in each stratum. We further filter users with  $< 10$  posts and only consider users who stay on the platform for at least a year.

In addition, we analyze various user groups but observe similar trends (Fig. A.1).

## A.4 Randomized Controlled Trial Details

**Table A.2:** Description of our randomized controlled trial (RCT) study population, setting and model, following reporting standards for artificial intelligence in health care from [Hernandez-Boussard et al. \(2020\)](#)

Feature	Description
<b>Study population and setting</b>	
- Population:	300 TalkLife users; 161 in Human Only (control); 139 in Human + AI (treatment).
- Study setting:	Non-clinical, online platform outside of TalkLife to ensure platform users' safety, through an interface similar to TalkLife's chat feature.
- Data collected in RCT:	Participants responded to 10*300=3000 seeker posts (1500 unique seeker posts duplicated across control and treatment), generating 3000 responses (1610 in control, 1390 in treatment). An independent set of 50 participants rated 1390 pairs of control and treatment responses on empathy preference.
- Cohort selection:	Participants were sent a recruitment request after they submitted a response on the TalkLife platform, with an aim of targeting active peer supporters. Participants were excluded if they dropped out of the study before completion.
- Registration:	We did not pre-register on <a href="#">ClinicalTrials.gov</a> because our study was conducted in a non-clinical setting.
<b>Participant demographic characteristics</b>	
- Age:	Mean=26.3 years; Std=9.5 years
- Gender:	Female: 54.3%; Male: 36.7%; Non-binary: 7.3%; Preferred not to say: 1.7%
- Race/Ethnicity:	White: 45.7%; Asian: 20.3%; Hispanic or Latino: 10.7%; Black or African American: 10.3%; Pacific Islander or Hawaiian: 0.7%; American Indian or Alaska Native: 0.3%; Preferred not to say: 12.0%
<b>Hailey's modeling components</b>	
- Model output:	Empathic rewriting of the response post
- Target user:	Peer supporter (users who provide peer-to-peer support to the support seeker)
- Data splitting:	Training: 3.2M; Test: 0.1M; Validation: 0.1M (seeker post, response post) pairs
- Gold standard:	180 empathic rewritings from human therapy experts used for evaluation of the original PARTNER model
- Model task:	Text generation
- Model architecture:	Deep reinforcement learning with a transformer based language model as its policy.
- Optimization:	Based on reward functions to increase empathy in posts and maintain text fluency, sentence coherence, context specificity, and diversity.
- Internal validation:	Automatic and human evaluation on hold-out test set.
- External validation:	The empathy scale used in HAILEY and PARTNER has been shown to correlate with "likes" from the support seeker and the forming of relationships b/w support seekers and peer supporters, consistent with empathy theory ( <a href="#">Bohart et al., 2002</a> ; <a href="#">Elliott et al., 2011</a> ; <a href="#">Watson et al., 2002</a> ). Our present randomized controlled trial represents an external evaluation of the rewriting modeling components of HAILEY and PARTNER.
- Transparency:	Data is available from TalkLife through a Data License Agreement; code is available via GitHub ( <a href="https://github.com/behavioral-data/PARTNER">github.com/behavioral-data/PARTNER</a> ).

**Figure A.2:** Empathy training used in our study. Participants in both the Human + AI (treatment) and Human Only (control) groups received the same training. The training included the empathy definition, a framework of common ways of expressing empathy in responses, and examples of empathic responses. This ensures that participants were working under similar understandings of empathy. In practice, such training is very rare and the effect of training typically diminishes over time. The identified difference in empathy between treatment and control groups in our study therefore conservatively estimates the impact of our AI-in-the-loop feedback system, and not baseline differences in empathy definitions. The effect in practice may be larger than the intentionally conservative estimates produced here, as such training is uncommon on current mental health platforms.

### Expressing empathy in responses

A key component of your responses should be **empathy** -- You should try and express empathy towards the seeker in your responses.

#### Empathy

Empathy is the ability to **understand** or **feel** the emotions and experiences of others and express that understanding in responses.

We adopt the widely-popular Roger's (1980) definition of empathy which highlights both **perspective-taking processes** and the **bodily-based emotional simulation processes** of empathy:

- "[Empathy is] the **therapist's sensitive ability and willingness to understand the client's thoughts, feelings and struggles from the client's point of view** (p. 85).... "It means entering the private perceptual world of the other...**being sensitive, moment by moment**, to the changing felt meanings which flow in this other person... It means **sensing meanings of which he or she is scarcely aware.**" ([Elliot et al.](#))

#### Empathic Responses

Since the focus here is on writing empathic responses, **expressing empathy in responses is key**. Empathic responses typically involve:

- Reacting with emotions felt after reading a post (e.g., *I feel sorry for you*)
- Communicating an understanding of feelings and experiences (e.g., *This must be terrifying*)
- Improving understanding by exploring feelings and experiences (e.g., *Are you feeling alone right now?*)

#### Examples of empathic responses

- **Seeker Post:** My whole family hates me.
- **Response Post:** I'm sorry to hear about your situation. If that happened to me, I would feel really isolated.

- **Seeker Post:** I feel like nobody cares about my existence.
- **Response Post:** It's hard to find others who can relate. I feel the same.

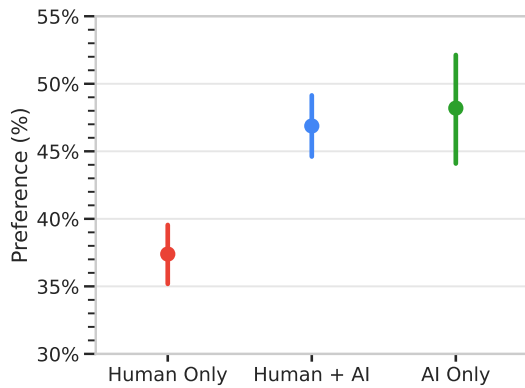
- **Seeker Post:** I can't deal with this part of my bipolar. I need help.
- **Response Post:** Being manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?

We will now start the study!

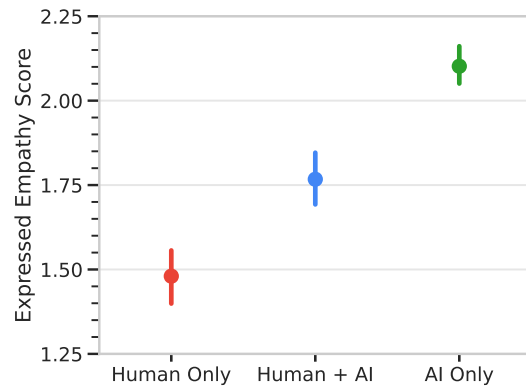
Start

**Figure A.3:** Comparison of Human Only (control) and Human + AI (treatment) responses with AI Only responses (generated directly from PARTNER(Sharma et al., 2021), the deep reinforcement learning model for empathic rewriting, used as a foundation for HAILEY (Methods)). **(a)** Through human evaluation from an independent set of TalkLife users, we found that AI Only responses have a similar preference as the Human + AI responses (48.2% vs. 46.9%; N=161;  $p=0.23$ ; Two-sided Student's t-test) but a higher preference than the Human Only responses (48.2% vs. 37.4%; N=139;  $p=3.3 \times 10^{-5}$ ; Two-sided Student's t-test). **(b)** Automatic estimation of empathy, on the contrary, suggested that AI Only responses have a higher expressed empathy score compared to Human + AI responses (2.10 vs. 1.77; N=139; Cohen's  $d=0.28$ ;  $p=5.3 \times 10^{-13}$ ; Two-sided Student's t-test). Importantly however, note that the AI Only responses were optimized on the same scoring function that we use to automatically estimate empathy, which likely explains the high scores of the AI Only approach. **(c)** However, while the authenticity of Human Only and Human + AI responses was comparable (69.6% vs. 65.4%; N=139;  $p=0.01$ ; Two-sided Student's t-test), the authenticity of AI Only responses was significantly lower (36.5% vs. 65.4%; N=161;  $p=3.7 \times 10^{-8}$ ; Two-sided Student's t-test). This highlights the key issue of authenticity with using AI Only, alongside safety, privacy, bias and other unintended consequences in the high-risk setting of mental health. To summarize, we find that Human + AI is the only approach that leads to both high empathy and high authenticity. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

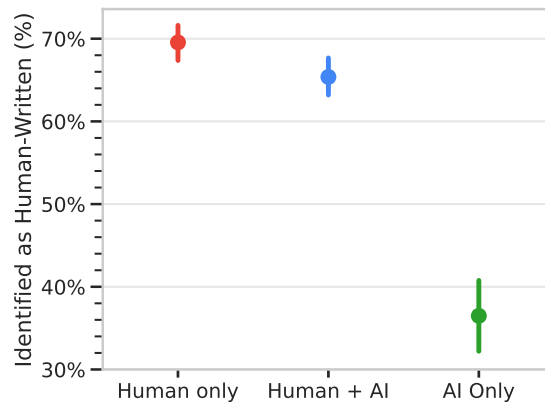
**(a) Human Evaluation:** Which response is more empathic?



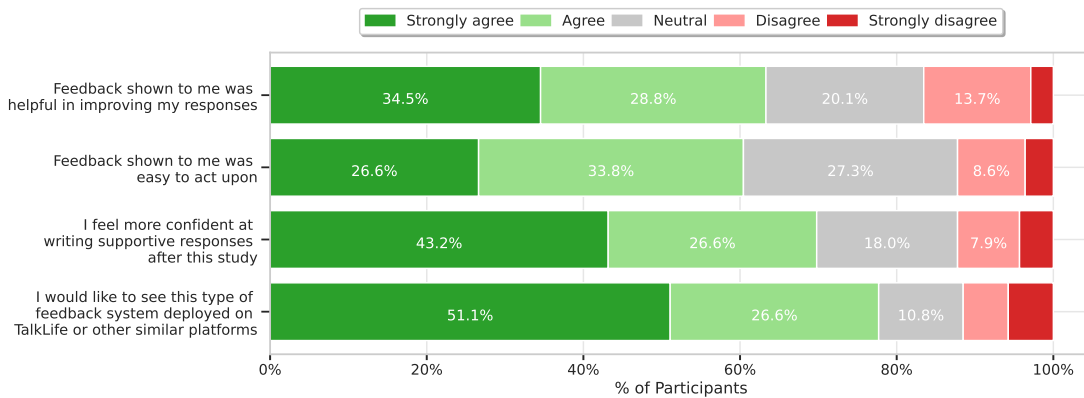
**(b) Automatic/AI-based Evaluation:** Expressed empathy score



**(c) Authenticity:** Is the response human-written or computer-generated?

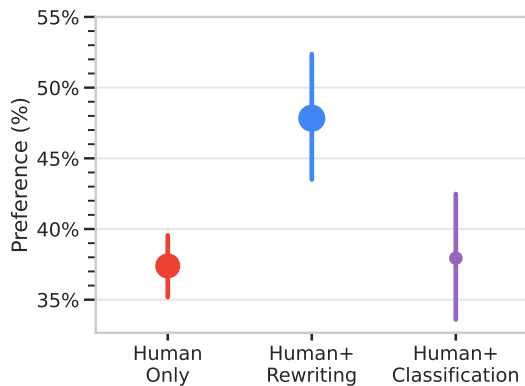


**Figure A.4:** Perceptions of Human + AI (treatment) group participants as reported in phase IV (post-intervention survey). We observed that more than 63.3% of participants found the current feedback helpful, 60.4% found it actionable and 69.8% of participants self-reported feeling more confident at providing support after our study. Also, 77.7% of participants wanted this type of feedback system to be deployed on TalkLife or other similar peer-to-peer support platforms, indicating potential opportunities for deployment in real-world.

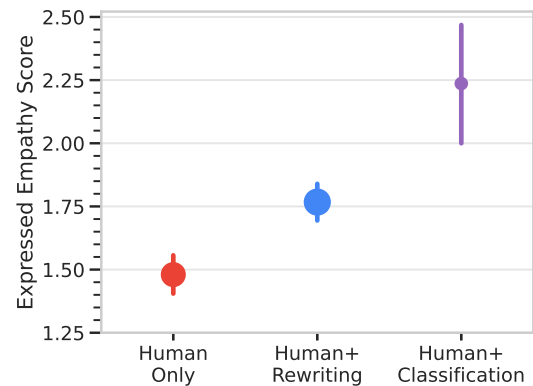


**Figure A.5:** Comparison of our rewriting-based AI treatment with a secondary classification-based AI treatment. A classification-based AI treatment provided participants with an option to request empathy classification scores for their responses, as opposed to the more granular feedback consisting of concrete suggestions to edit responses in our primary rewriting-based approach (Supplementary Figure A.6). Our hypothesis was that such a treatment should be less actionable and is likely to lead to less empathic responses than the rewriting-based treatment. In our study, we assigned a secondary classification-based treatment to 10% of the incoming participants at random (N=30). **(a)** Through human evaluation from an independent set of TalkLife users, we found that the Human + Classification responses have a significantly lower preference than the Human + Rewriting responses (37.9% vs. 47.8%; N=30; p=0.002; Two-sided Student’s t-test). **(b)** Automatic estimation of empathy, on the contrary, suggested that the Human + Classification responses have a higher expressed empathy score compared to Human + Rewriting responses (2.24 vs. 1.77; N=30; Cohen’s d=0.37; p=4.7 \* 10<sup>-6</sup>; Two-sided Student’s t-test). As the same score is also exposed to participants just-in-time in the classification-based treatment, it may have led participants to be put particular emphasis on a high expressed empathy score, which participants in the rewriting-based treatment feedback didn’t have direct access to. **(c)** We found that less participants in the classification-based treatment group agree on deploying the system on TalkLife than the rewriting-based treatment (63.3% vs. 77.7%; N=30; p=0.0998; Two-sided Student’s t-test; Supplementary Figure A.4). Also, we observed that more participants in the classification-based treatment disagree on its actionability than participants in the rewriting-based treatment, but the difference may not be statistically significant due to the limited power (23.3% vs. 12.2%; N=30; p=0.1154; Two-sided Student’s t-test). The area of the points in the plots is proportional to the number of participants in the respective control/treatment conditions. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

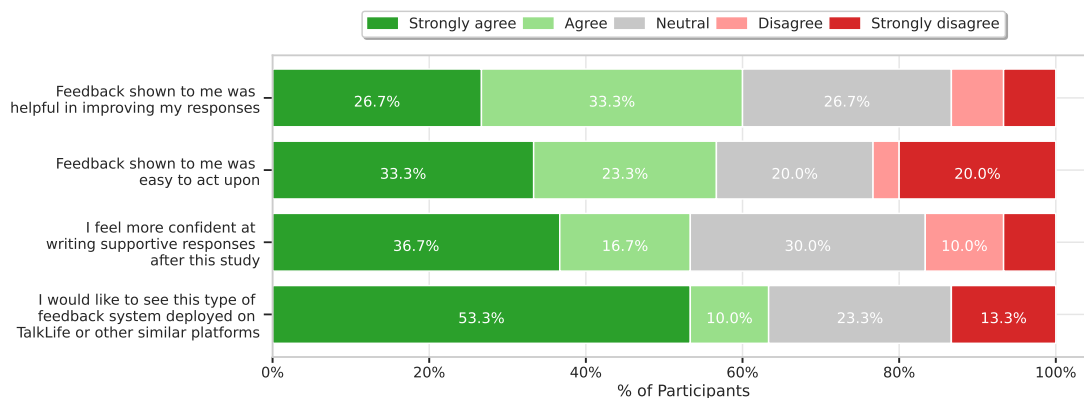
**(a) Human Evaluation:** Which response is more empathic?



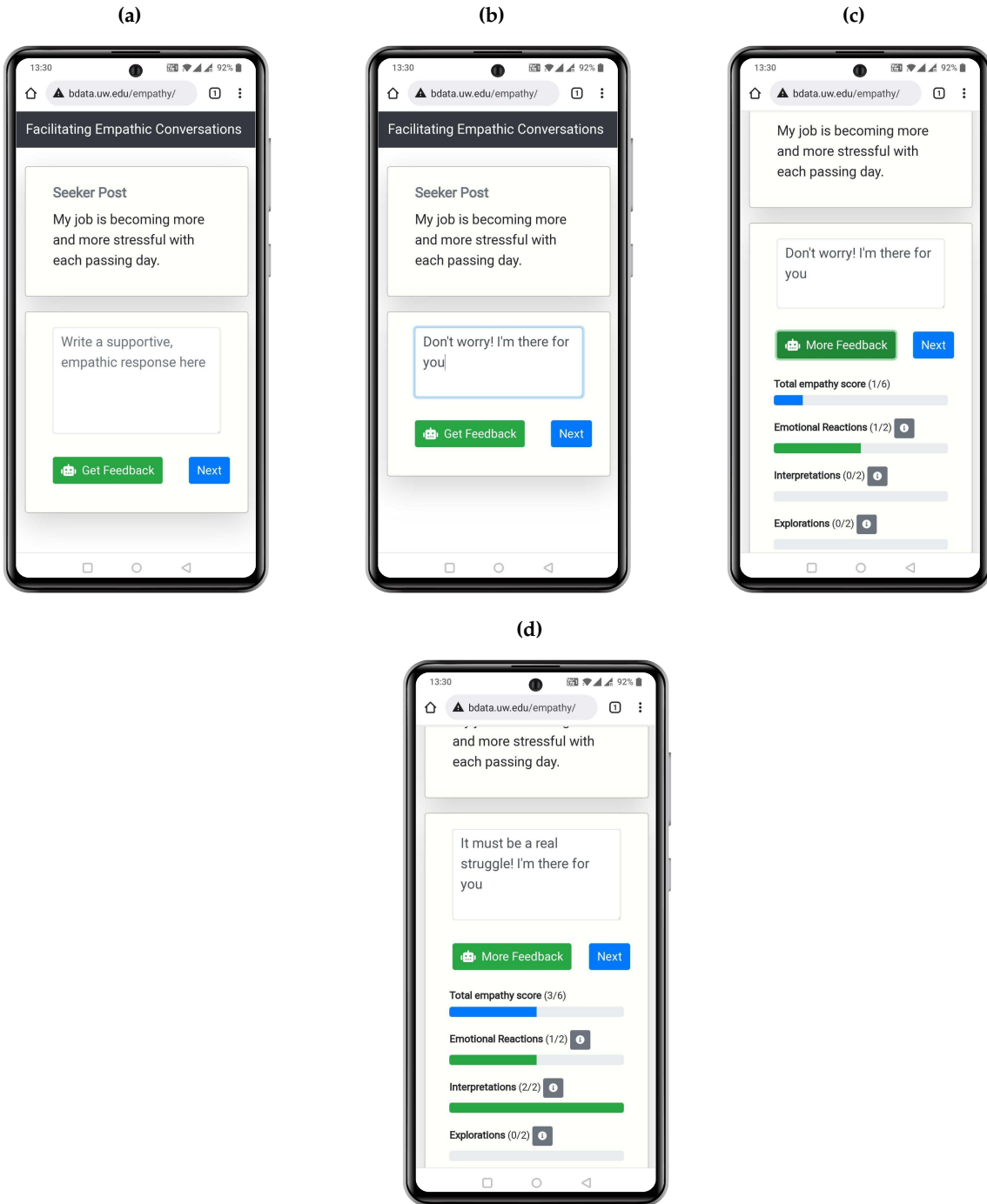
**(b) Automatic/AI-based Evaluation:** Expressed empathy score



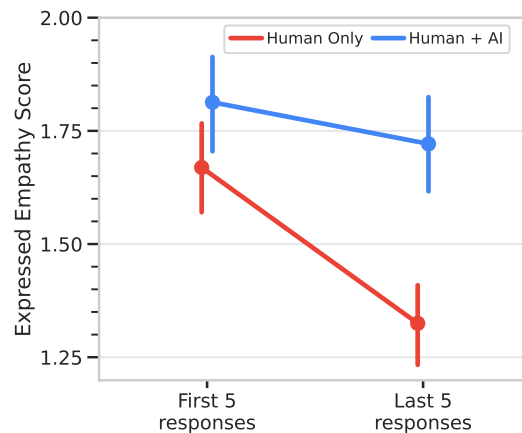
**(c) Study participants’ perceptions (classification-based treatment group)**



**Figure A.6:** Interface of our classification-based AI treatment (Supplementary Figure A.5). **(a)** Participant is asked to write a supportive, empathic response and given an option to receive feedback. **(b)** Participant starts writing the response. **(c)** Participant clicks on the “Get Feedback” button to request classification-based feedback. The feedback consists of classification scores on three empathy communication mechanisms – Emotional Reactions, Interpretations, and Explorations. **(d)** Participant edits the response based on the classification scores, often improving on the communication mechanisms with low scores and requests “More Feedback” if needed.

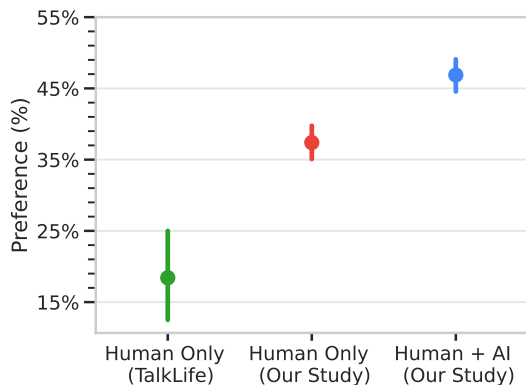


**Figure A.7:** Both Human Only (control) and Human + AI (treatment) group participants showed a significant drop in empathy levels in the last 5 responses of our study. With Human + AI, however, we observed a significantly lower drop in empathy (5.3% vs. 26.0%; N=139; p=0.0062; Two-sided Student's t-test). This indicates the effectiveness of just-in-time AI feedback in alleviating challenges like empathy fatigue, associated with providing mental health support. The empathy differences between Human Only (N=161) and Human + AI (N=139) responses are statistically significant for both first 5 and last 5 responses (p=1.1 \* 10<sup>-8</sup>; Two-sided Student's t-test). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

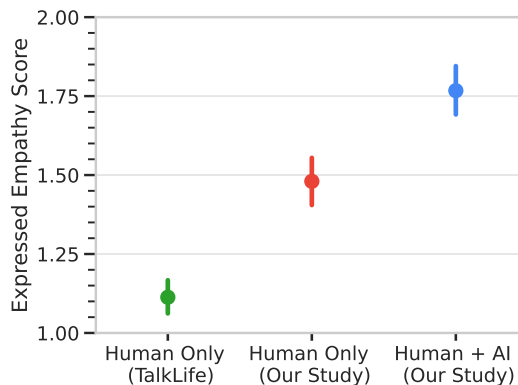


**Figure A.8:** Comparison of existing Human Only responses on TalkLife with Human Only and Human + AI responses in our study. Human Only responses on TalkLife had significantly lower preference for empathy (18.4% vs. 37.4% vs. 46.9%;  $N=139$ ;  $p=1.4 \times 10^{-6}$ ; Two-sided Student's t-test) and significantly lower expressed empathy score (1.11 vs. 1.48 vs. 1.77;  $p=2.8 \times 10^{-46}$ ; Two-sided Student's t-test). This difference might be attributed to the additional initial empathy training provided to participants, as well as a potential selection effect in our study that may have attracted Talklife users who particularly care about expressing empathy in supporting others. As our study shows that Human-AI collaboration improves empathy expression even for those participants who already express empathy more often, practical gains for the average user of the Talklife platform could be even higher. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

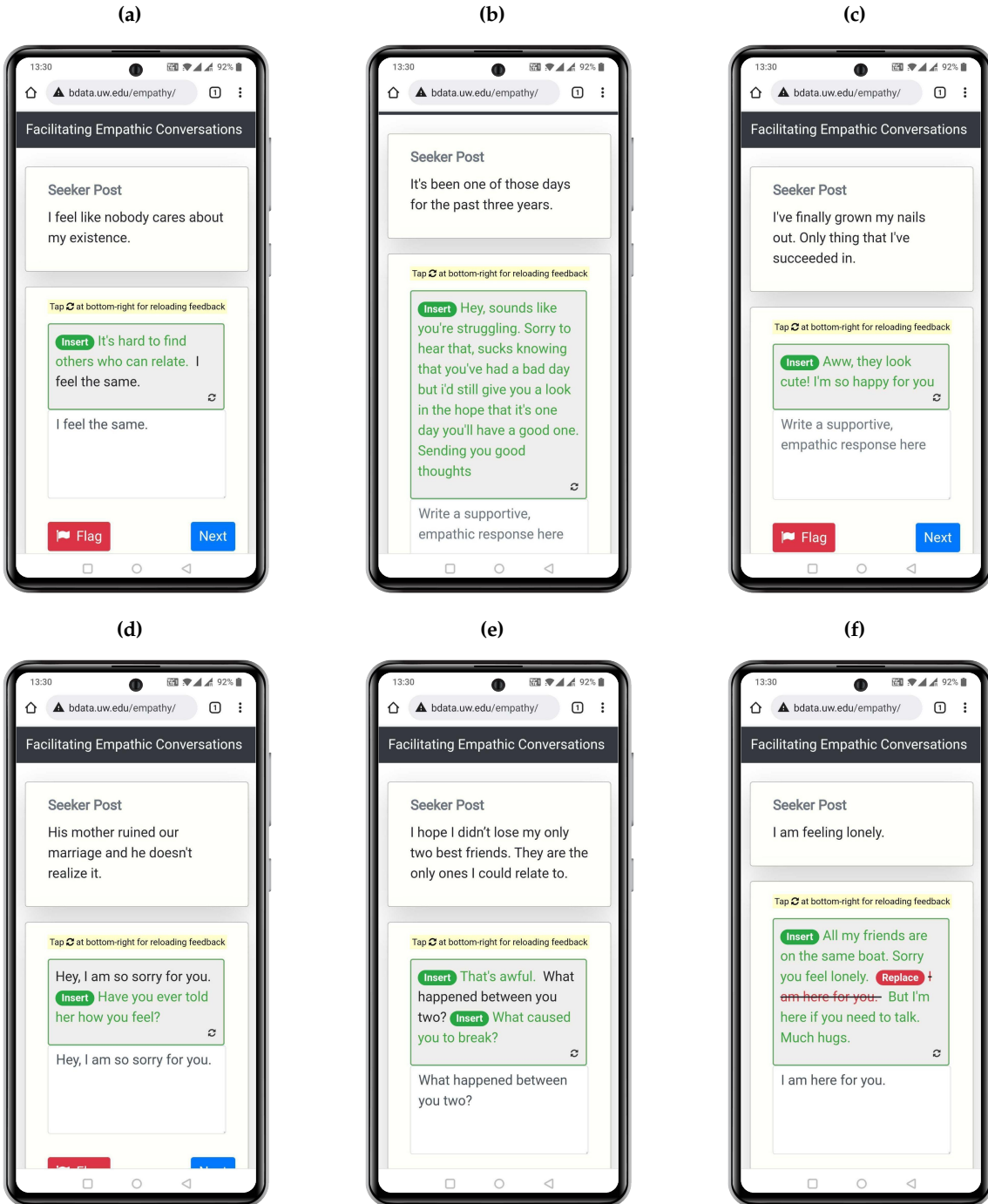
**(a) Human Evaluation:** Which response is more empathic?



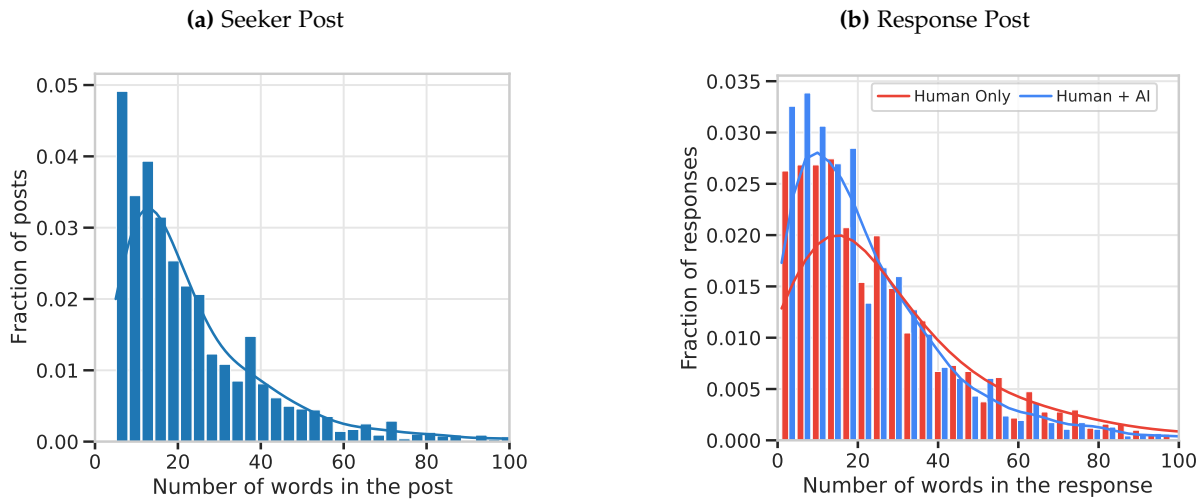
**(b) Automatic/AI-based Evaluation:** Expressed empathy score



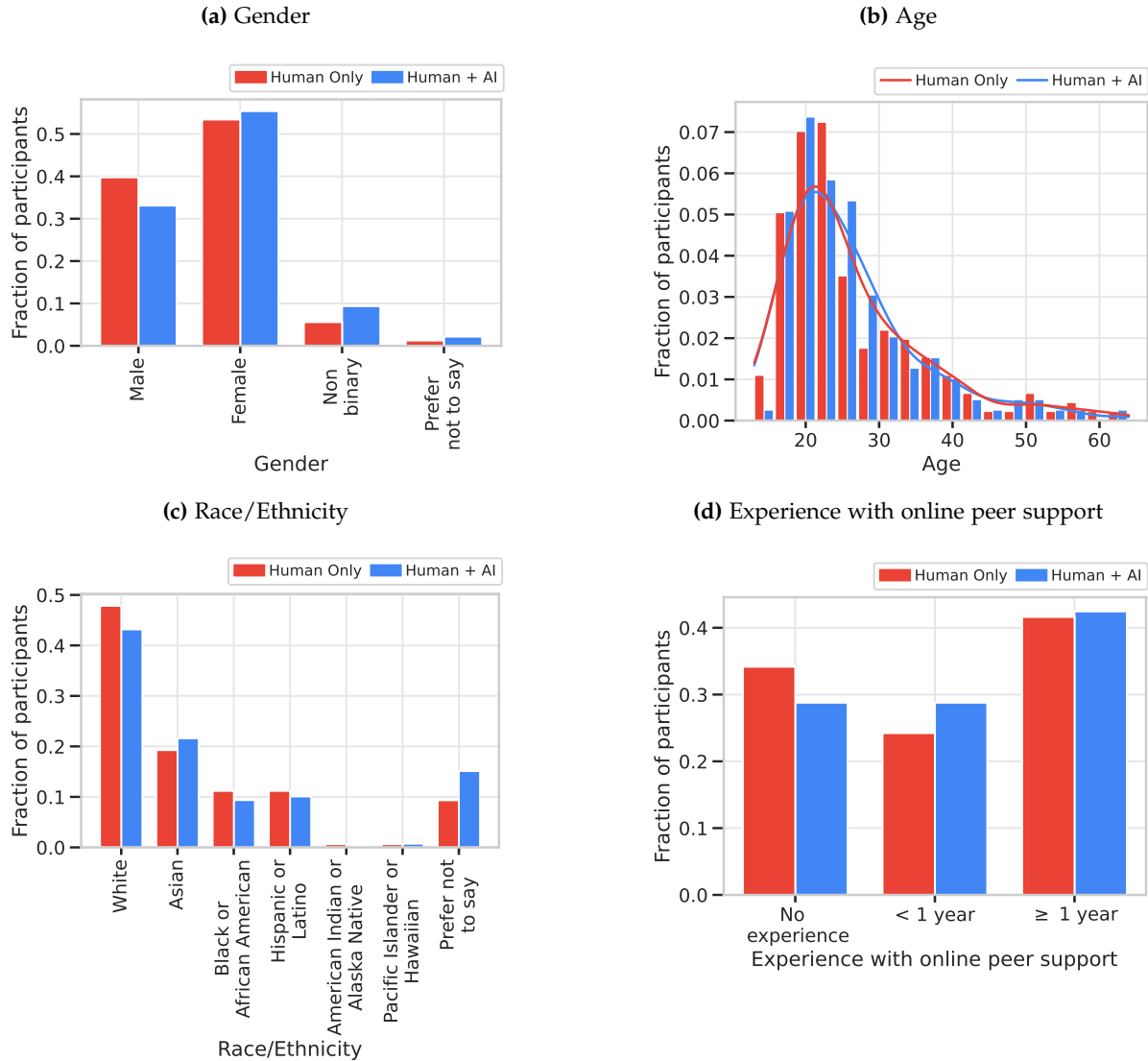
**Figure A.9:** Qualitative examples of just-in-time AI feedback provided to participants by HAILEY. In (b) and (c), the original peer supporter response was empty. Seeker posts in these examples have been paraphrased for anonymization.



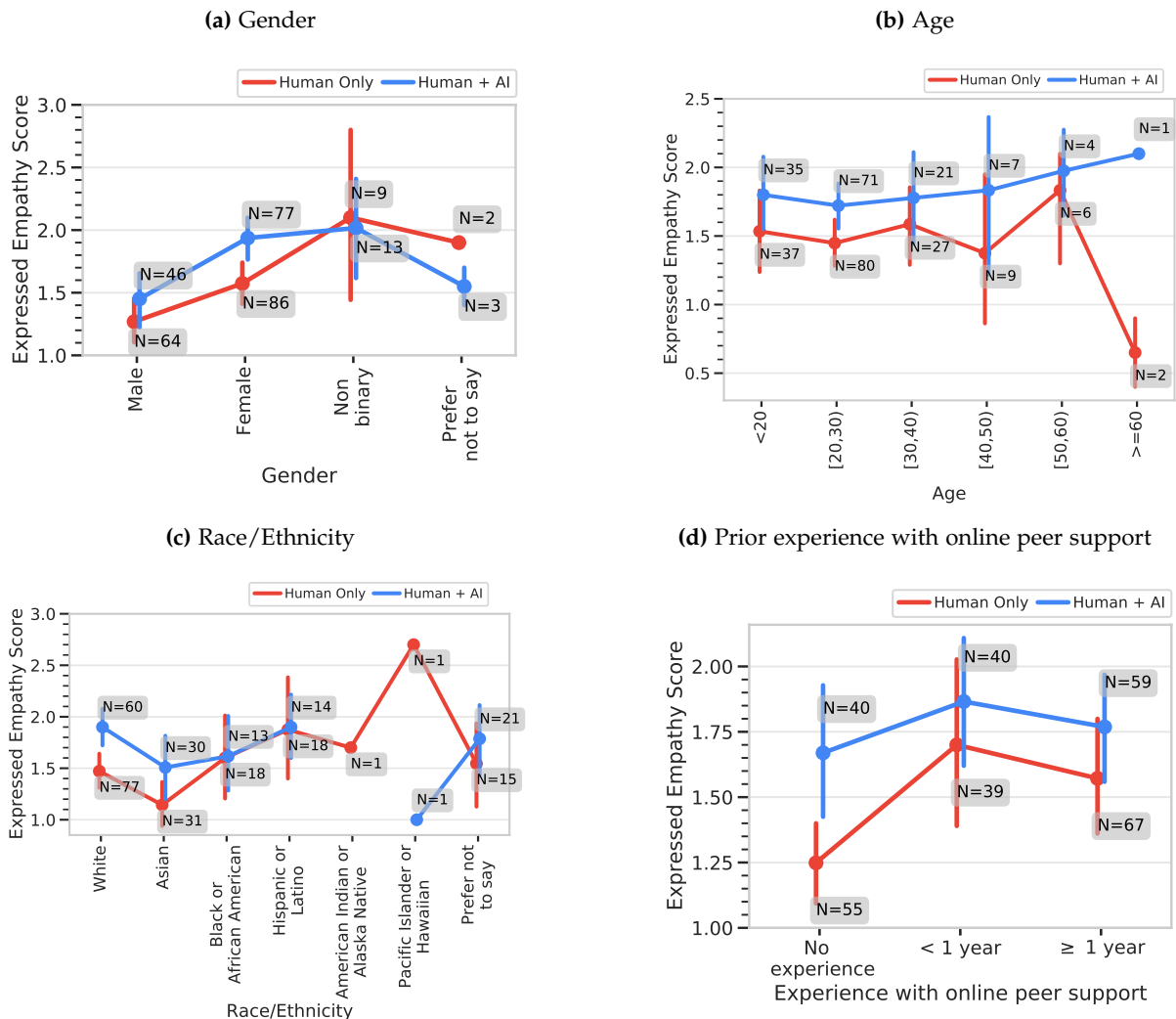
**Figure A.10:** The distribution of post and response lengths. The seeker posts in our dataset had a mean length of 25.9 words, a standard deviation of 25.3 words and a median of 18.0 words. The response posts collected in our study had a mean length of 25.9 words, a standard deviation of 34.6 words and a median of 19.0 words. Also, Human + AI responses (mean = 22.4 words; std = 34.6; median = 19.0) were 28.9% shorter in length compared to Human Only responses on average (mean = 32.1 words; std = 44.4; median = 21.0;  $p < 0.001$ ; Two-sided Student's t-test). In addition, we found that Human + AI responses had 5.2% higher diversity than the Human Only responses based on the Distinct-1 metric (Li et al., 2016a), which computes the number of distinct unigrams divided by the total number of tokens (0.146 vs. 0.139;  $p = 0.019$ ; Two-sided Wilcoxon signed-rank test).



**Figure A.11:** Background and demographics of participants in Human Only (control) and Human + AI (treatment) groups, as reported in phase I (pre-intervention survey).

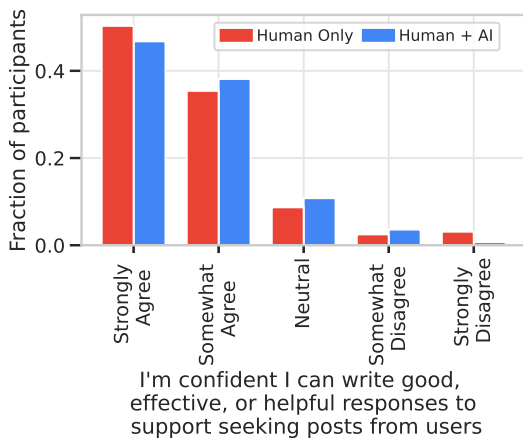


**Figure A.12:** Differences between expressed empathy scores of participants in Human Only (control) and Human + AI (treatment) groups, stratified by demographics of participants and their prior experience with online peer support. The area of the points is proportional to the number of participants in the respective categories. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

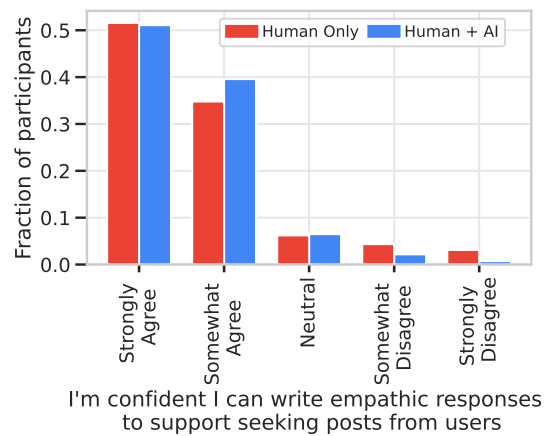


**Figure A.13:** Perceptions of participants in Human Only (control) and Human + AI (treatment) groups, as reported in phase I (pre-intervention survey).

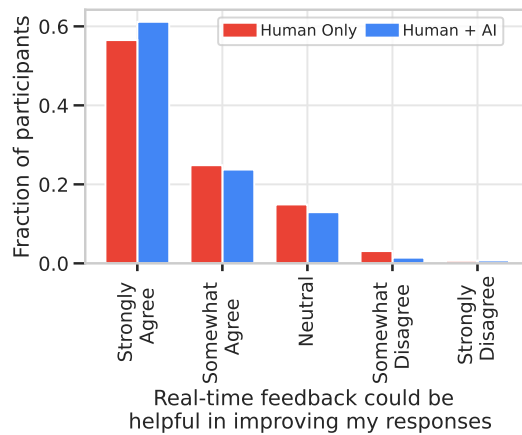
**(a)** Self-efficacy in writing good, effective, or helpful responses



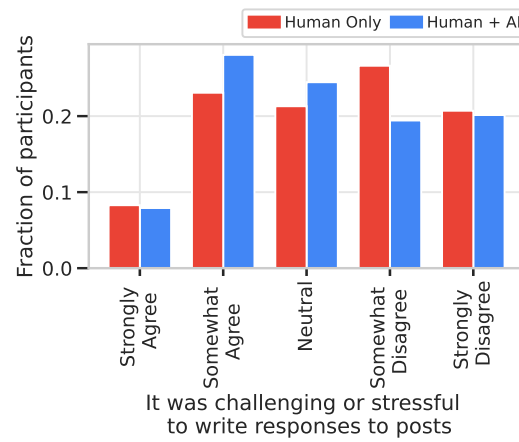
**(b)** Self-efficacy in writing empathic responses



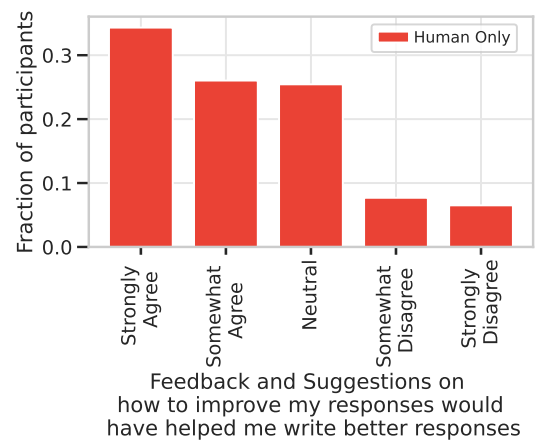
**(c)** Could feedback be helpful?



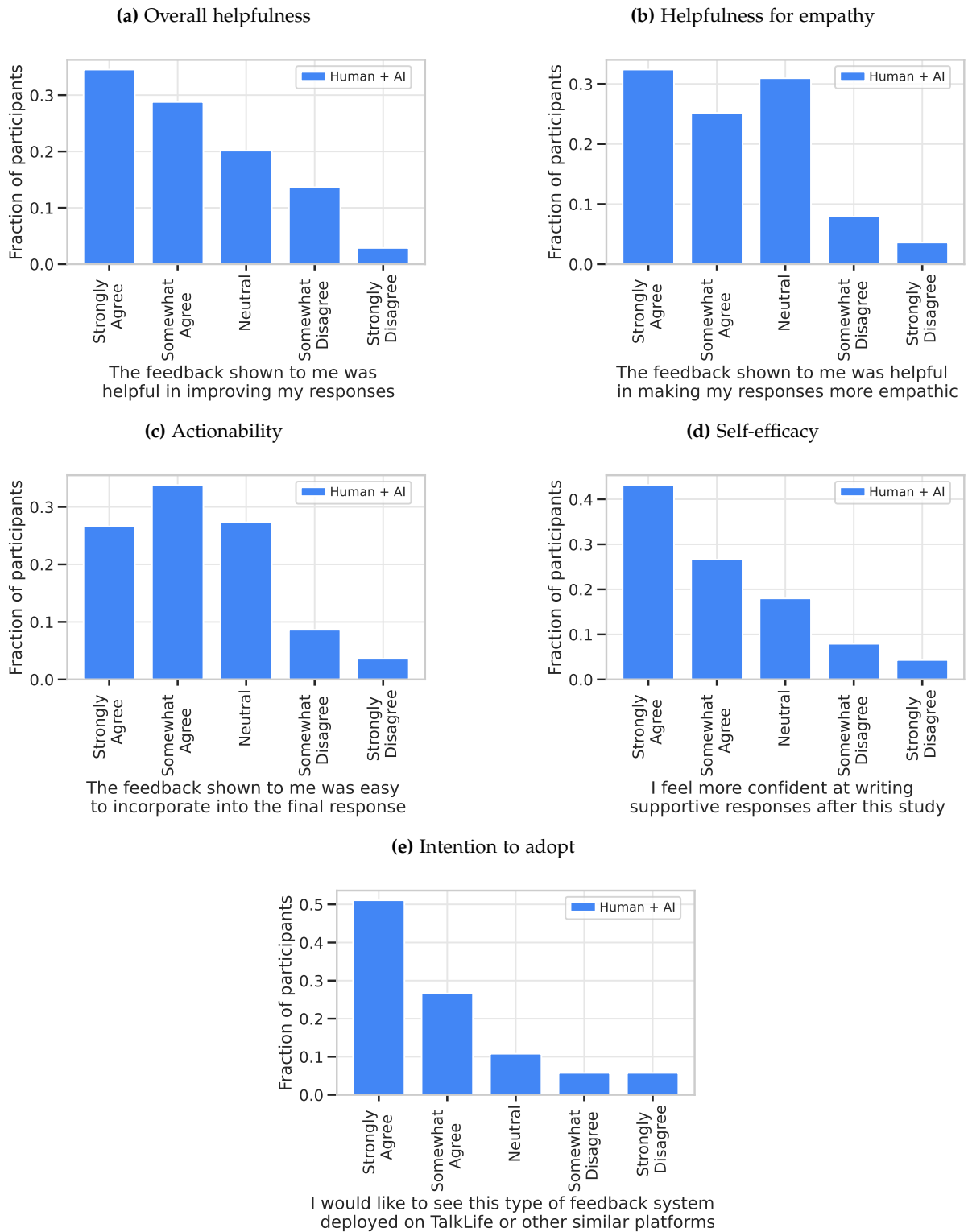
**Figure A.14:** Distribution of participants in Human Only (control) and Human + AI (treatment) groups who report writing responses as challenging or stressful, as reported in phase IV (post-intervention survey).



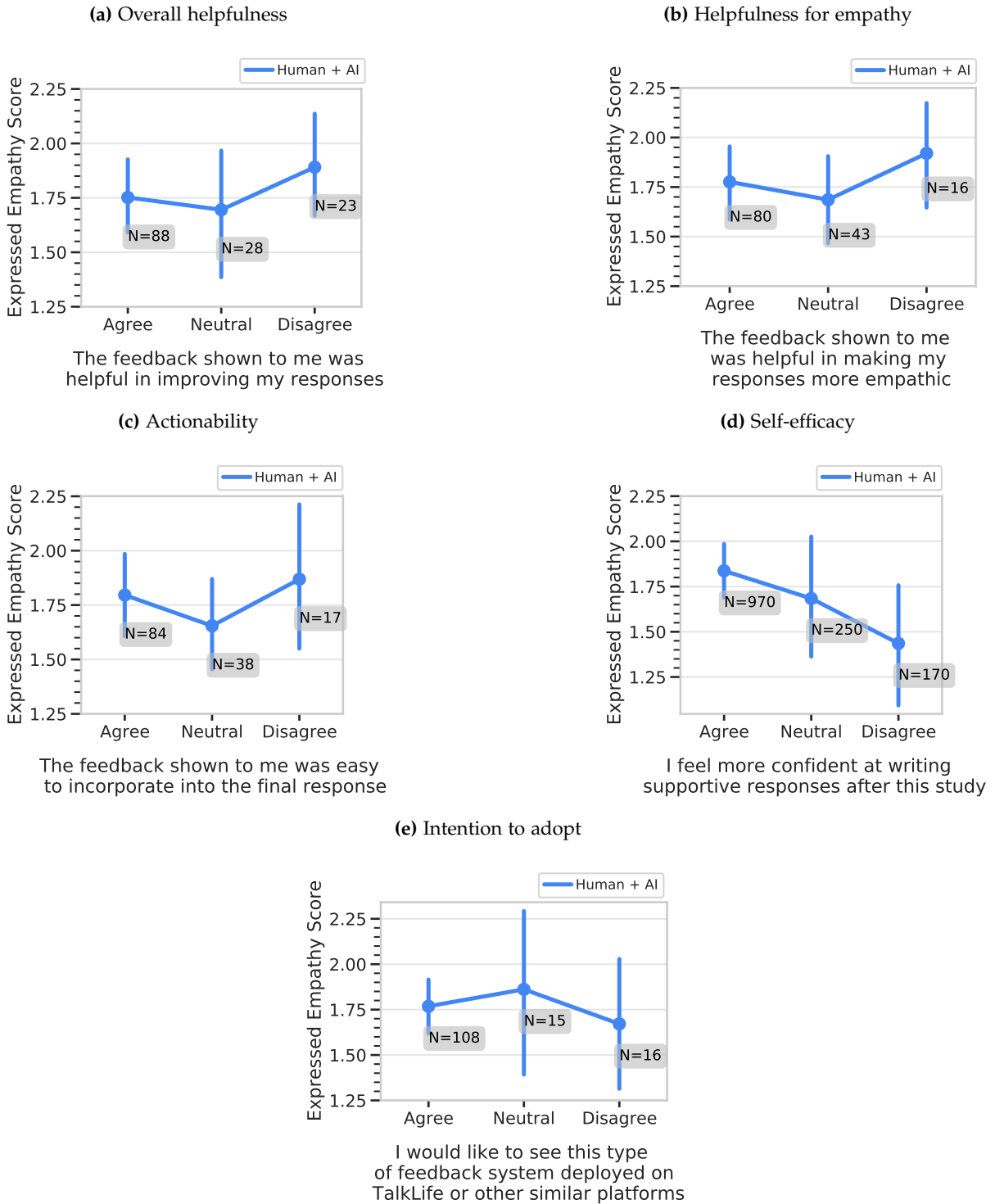
**Figure A.15:** Distribution of participants in the Human Only (control) group who indicate that feedback could have improved responses, as reported in phase IV (post-intervention survey).



**Figure A.16:** Perceptions of participants in the Human + AI (treatment) group, as reported in phase IV (post-intervention survey).

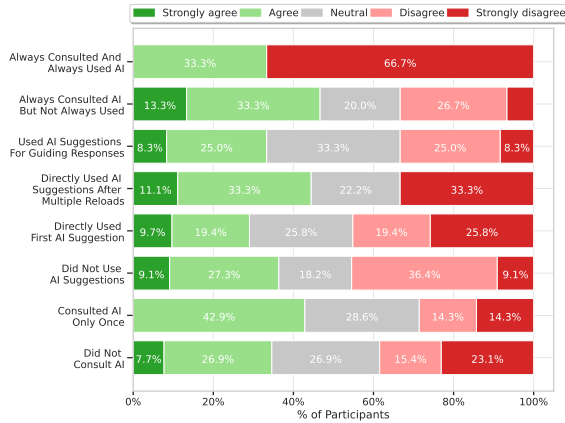


**Figure A.17:** Expressed empathy levels of responses with perceptions of Human + AI (treatment) group participants, as reported in phase IV (post-intervention survey). The area of the points is proportional to the number of participants with respective perceptions. Error bars indicate bootstrapped 95% confidence intervals.

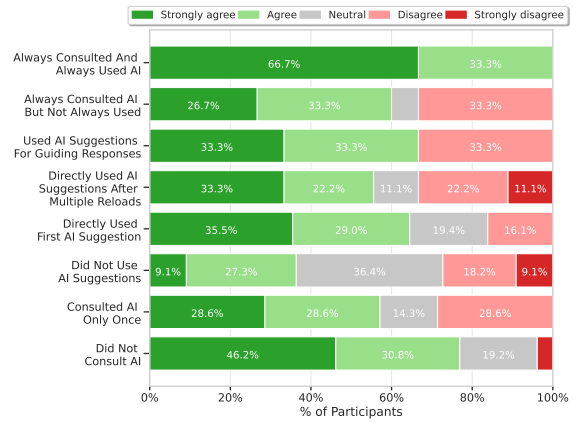


**Figure A.18:** Participant perceptions, as reported in phase IV (post-intervention survey), with different human-AI collaboration categories.

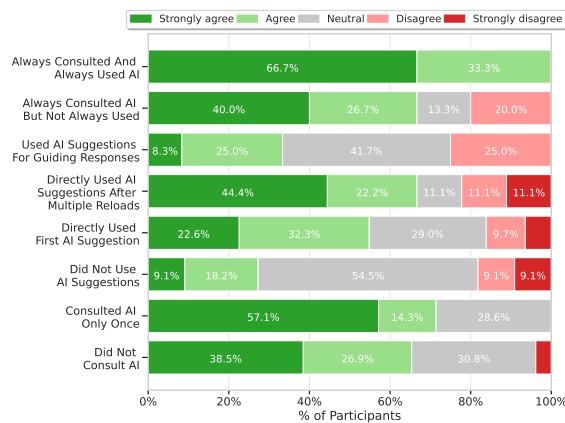
**(a) Challenges:** Feedback and Suggestions on how to improve my responses would have helped me write better responses



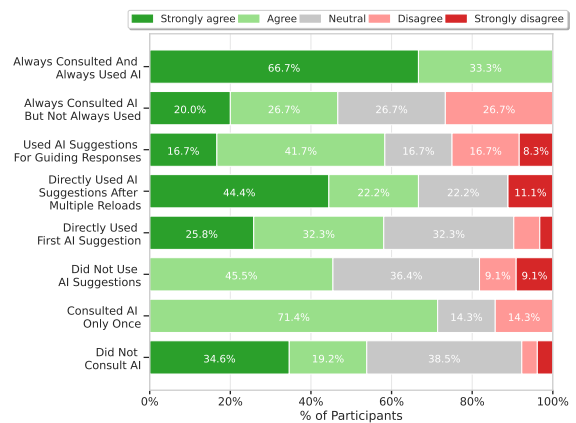
**(b) Overall helpfulness:** The feedback shown to me was helpful in improving my responses



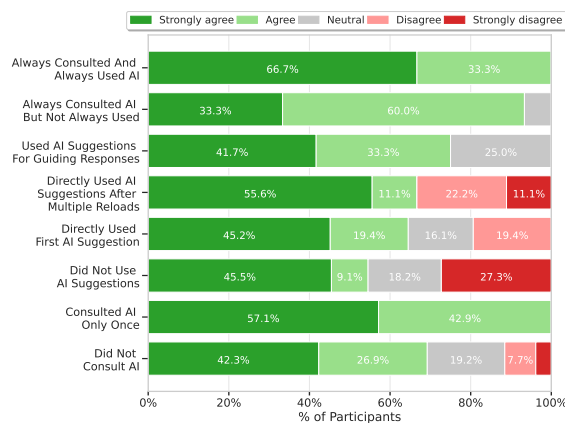
**(c) Helpfulness for empathy:** The feedback shown to me was helpful in making my responses more empathic



**(d) Actionability:** The feedback shown to me was easy to incorporate into the final response



**(e) Self-efficacy:** I feel more confident at writing supportive responses after this study



**(f) Intention to adopt:** I would like to see this type of feedback system deployed on TalkLife or other similar platforms

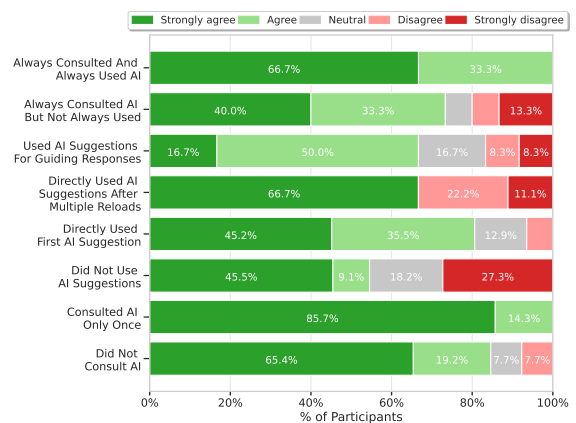


Figure A.19: Consent form used in our study.

## Disclaimer

Thank you for your interest in our study!

- We are **researchers at the University of Washington**, studying peer-to-peer support platforms.
- This study is **not being conducted by TalkLife**. TalkLife is not responsible for any risks/benefits associated with this study.
- As part of this study, we will be collecting (a) **textual responses to support seeking posts**, and (b) **answers to survey questions** describing your background and your assessment of our study. The data from this study will be uploaded to a **secure platform accessible only to the research team**. All collected data will be exclusively used for research purposes. Also, data will only be analyzed in aggregate.
- This study has been determined to be exempt from IRB approval under University of Washington IRB ID STUDY00012706. For concerns or questions, please contact [ashshar@cs.washington.edu](mailto:ashshar@cs.washington.edu).
- This is **not a live conversation** with users. Your responses will not be posted on TalkLife or any other online platform.
- Your participation in this study is **completely voluntary**. You are free to release/quit the study at any time. Refusing to be in the experiment or stopping participation will involve no penalty or loss of benefits to which you are otherwise entitled.
- Poor-quality data (e.g., same response to all posts) **may be removed** without compensation.
- If you are a US citizen or a permanent US resident, you will receive an **Amazon gift card worth 5 USD** after completing this study.
  - Yes**, I'm a US citizen or a permanent US resident.
  - No**, I'm neither a US citizen nor a permanent US resident.
  - Prefer not to say.

## Consent

I agree to participate in this study. I also understand that only US citizens or permanent US residents can be compensated.

Accept and Continue

**Figure A.20:** Form used for collecting demographics and background of participants [phase I: pre-intervention survey].

### Tell us about yourself

Email is **only collected for sending gift cards** and will neither be used for analysis nor be stored by us after the study.

Name

Email

Age

Gender

Country

Race/Ethnicity

Previous experience with online peer support

**Figure A.21:** Onboarding survey used for collecting perceptions of participants [phase I: pre-intervention survey].

### Onboarding Survey

1) I'm confident I can write **good, effective, or helpful responses** to support seeking posts from users.

- Strongly Agree
- Somewhat Agree
- Neutral
- Somewhat Disagree
- Strongly Disagree

2) I'm confident I can write **empathic responses** to support seeking posts from users.

- Strongly Agree
- Somewhat Agree
- Neutral
- Somewhat Disagree
- Strongly Disagree

3) **Real-time feedback** could be helpful in improving my responses.

- Strongly Agree
- Somewhat Agree
- Neutral
- Somewhat Disagree
- Strongly Disagree

[Next](#)

**Figure A.22:** Instructions shown to the control group participants [phase II: empathy training and instructions]. Continued on the next page (1/2).

## Instructions

The study will involve writing textual responses to support seeking posts and answering survey questions. The entire study is expected to take ~30 minutes.

## Content Warning

The study contains posts including but not limited to self-harm and suicidal ideation, which may be disturbing to you. If you have concerns or questions, please send us an email ([ashshar@cs.washington.edu](mailto:ashshar@cs.washington.edu)). If you have strong negative reactions to some of the content, please reach out at [crisis text line](#).

## What will you do?

During the main part of the study, you will be shown online mental health posts through which people seek support. We call them "Seeker Posts" (writers of these posts are called "Seekers"). For each seeker post, you will be asked to **write a supportive, empathic response**.

Also, at various steps, you will be asked **survey questions** describing your background and your assessment of our study.

## Expressing empathy in responses

A key component of your responses should be **empathy** -- You should try and express empathy towards the seeker in your responses.

## Empathy

Empathy is the ability to **understand** or **feel** the emotions and experiences of others. Empathic responses typically involve:

- Reacting with emotions felt after reading a post (e.g., *I feel sorry for you*)
- Communicating an understanding of feelings and experiences (e.g., *This must be terrifying*)
- Improving understanding by exploring feelings and experiences (e.g., *Are you feeling alone right now?*)

## Examples of empathic responses

- **Seeker Post:** My whole family hates me.
- **Empathic Response:** I'm sorry to hear about your situation. If that happened to me, I would feel really isolated.

- **Seeker Post:** I feel like nobody cares about my existence.
- **Empathic Response:** It's hard to find others who can relate. I feel the same.

- **Seeker Post:** I can't deal with this part of my bipolar. I need help.
- **Empathic Response:** Being manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?

Next

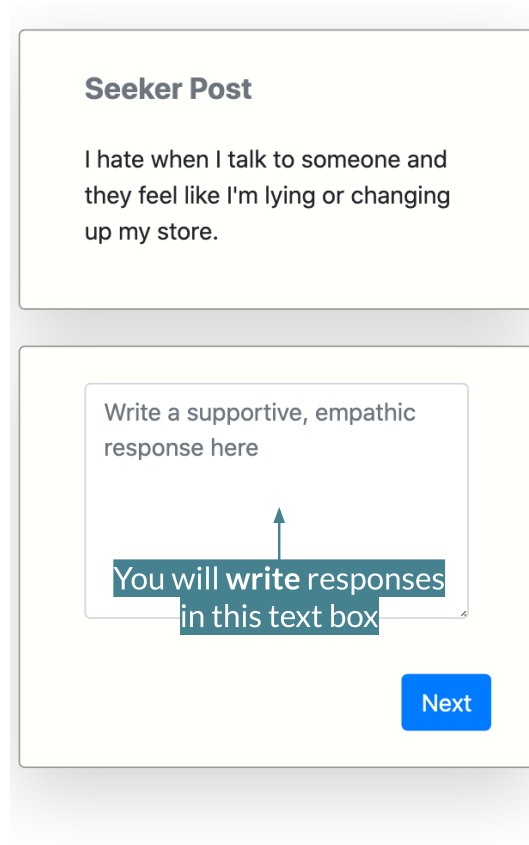
**Figure A.23:** Instructions shown to the control group participants [phase II: empathy training and instructions] (2/2).

Thank you for filling out the survey!

Next, you will **write responses to 10 support seeking posts**. Here, we will give you an overview of our interface.

### The Interface

This is how the main interface will look:



You will read the seeker post and **write a supportive, empathic response** in the space provided.

Start Study

**Figure A.24:** Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (1/6).

## Instructions

The study will involve writing textual responses to support seeking posts and answering survey questions. You will receive real-time feedback with suggestions on how to improve your response. The entire study is expected to take ~30 minutes.

## Content Warning

The study contains posts including but not limited to self-harm and suicidal ideation, which may be disturbing to you. If you have concerns or questions, please send us an email ([ashshar@cs.washington.edu](mailto:ashshar@cs.washington.edu)). If you have strong negative reactions to some of the content, please reach out at [crisis text line](#).

## What will you do?

During the main part of the study, you will be shown online mental health posts through which people seek support. We call them "Seeker Posts" (writers of these posts are called "Seekers"). For each seeker post, you will be asked to **write a supportive, empathic response**. You will get opportunities to **receive "help"** on your responses. We will suggest you ways in which you can improve your responses. You are **strongly recommended to always check these suggestions** and **use** them if they make your response more **supportive** and **empathic**.

Also, at various steps, you will be asked **survey questions** describing your background and your assessment of our study.

## Expressing empathy in responses

A key component of your responses should be **empathy** -- You should try and express empathy towards the seeker in your responses.

### Empathy

Empathy is the ability to **understand** or **feel** the emotions and experiences of others. Empathic responses typically involve:

- Reacting with emotions felt after reading a post (e.g., *I feel sorry for you*)
- Communicating an understanding of feelings and experiences (e.g., *This must be terrifying*)
- Improving understanding by exploring feelings and experiences (e.g., *Are you feeling alone right now?*)

### Examples of empathic responses

- **Seeker Post:** My whole family hates me.
- **Empathic Response:** I'm sorry to hear about your situation. If that happened to me, I would feel really isolated.

- **Seeker Post:** I feel like nobody cares about my existence.
- **Empathic Response:** It's hard to find others who can relate. I feel the same.

- **Seeker Post:** I can't deal with this part of my bipolar. I need help.
- **Empathic Response:** Being manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?

Next

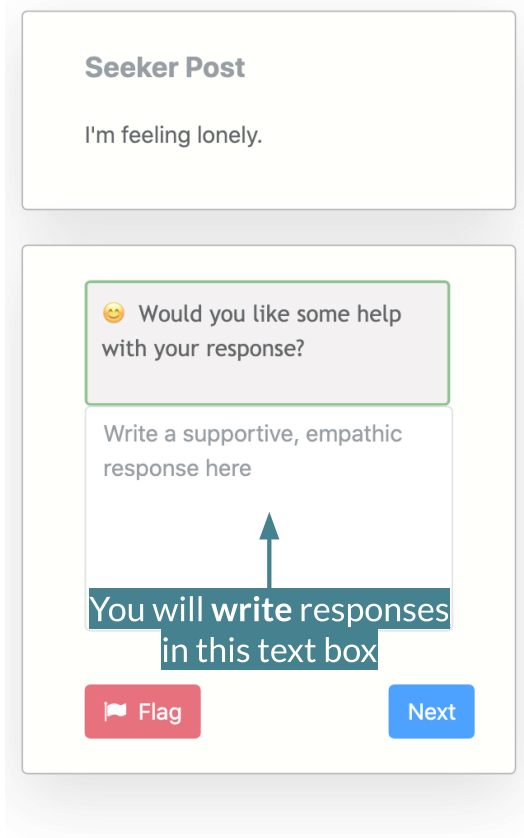
**Figure A.25:** Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (2/6).

Thank you for filling out the survey!

Next, you will **write responses to 10 support seeking posts** while receiving feedback. We will first give you an overview of our interface.

### The Interface

This is how the main interface will look:



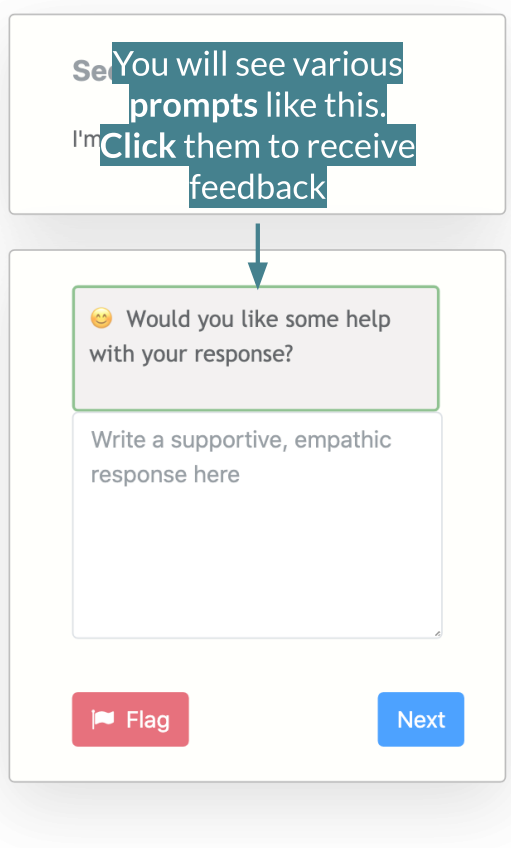
You will read the seeker post and **write a supportive, empathic response** in the space provided.

Next (1/4)

**Figure A.26:** Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (3/6).

### Interface - Receiving feedback via prompts

You will see **prompts to receive real-time feedback** (as shown below). You can click on the prompts to get help with your responses.



You are **strongly recommended to always check these suggestions** and **use** them if they make your response more **supportive** and **empathic**.

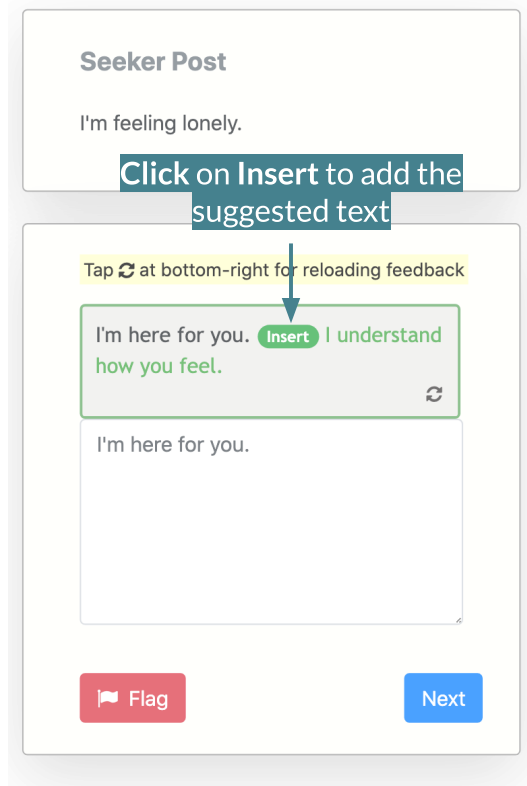
Next (2/4)

**Figure A.27:** Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (4/6).

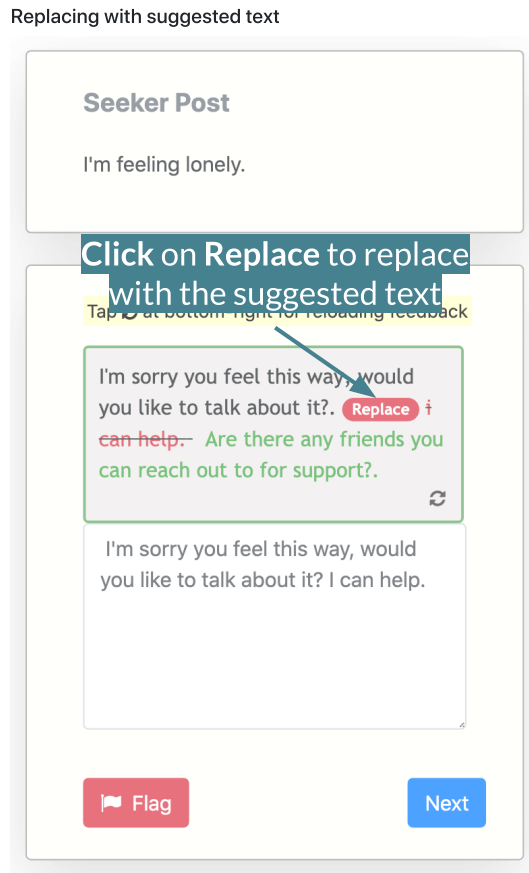
### Interface - Insert and Replace Operations

In our feedback, we will **suggest text** that **you can insert or replace** in your current response to make it more supportive and empathic.

#### Inserting suggested text



**Figure A.28:** Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (5/6).



You can directly incorporate the changes by clicking on **Insert** and **Replace** buttons.

Next (3/4)

**Figure A.29:** Instructions shown to the treatment group participants [phase II: empathy training and instructions] (6/6).

### Interface - Bad feedback

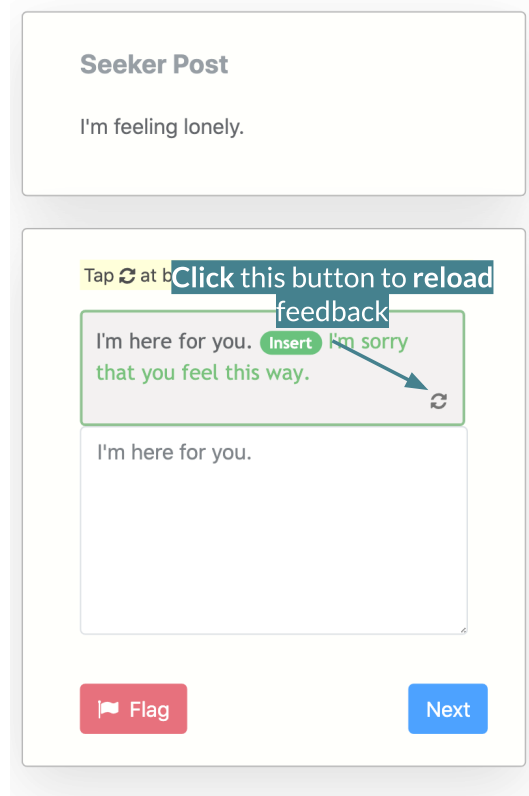
Our feedback will **not** always be perfect. If the feedback is **bad** or **inappropriate**, you may refine, reload, or report the feedback.

#### Refine

You may need to refine the feedback to **correct grammar** or **content**. You should **make relevant changes** such that the feedback can be appropriately integrated in your final response.

#### Reload

Also, whenever the feedback is bad, you can use the  button (see below) to **reload** and **get new feedback**.



You can reload **multiple times** till you see feedback that is helpful.

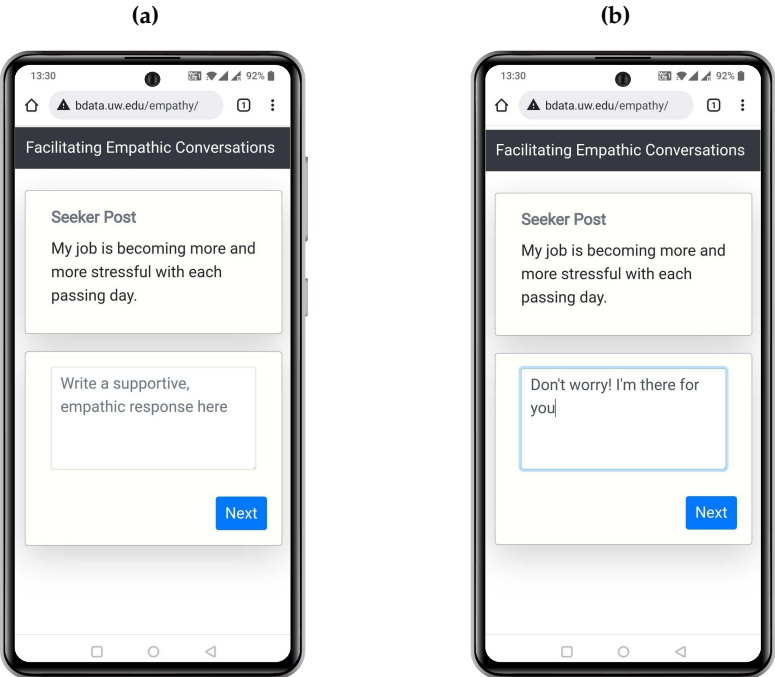
#### Report

If you see feedback that is **inappropriate** or **toxic**, you can report it using the  **Flag** button.

We will now start the study!

Start (4/4)

**Figure A.30:** An example workflow for Human Only (control) participants [phase III: write supportive, empathic responses]. **(a)** Participant is asked to write a supportive, empathic response. **(b)** Participant starts writing the response.



**Figure A.31:** An example workflow for Human + AI (treatment) participants [phase III: write supportive, empathic responses]. **(a)** Participant is asked to write a supportive, empathic response and given an option to receive feedback. **(b)** Participant starts writing the response. **(c)** Participant clicks on the prompt to request feedback from HAILEY. **(d)** Participant accepts the suggested changes and gets an option to request more feedback. **(e)** Participant continues editing the response and requests more feedback as needed. **(f)** When the response is already highly empathic, the participant simply receives a positive feedback.

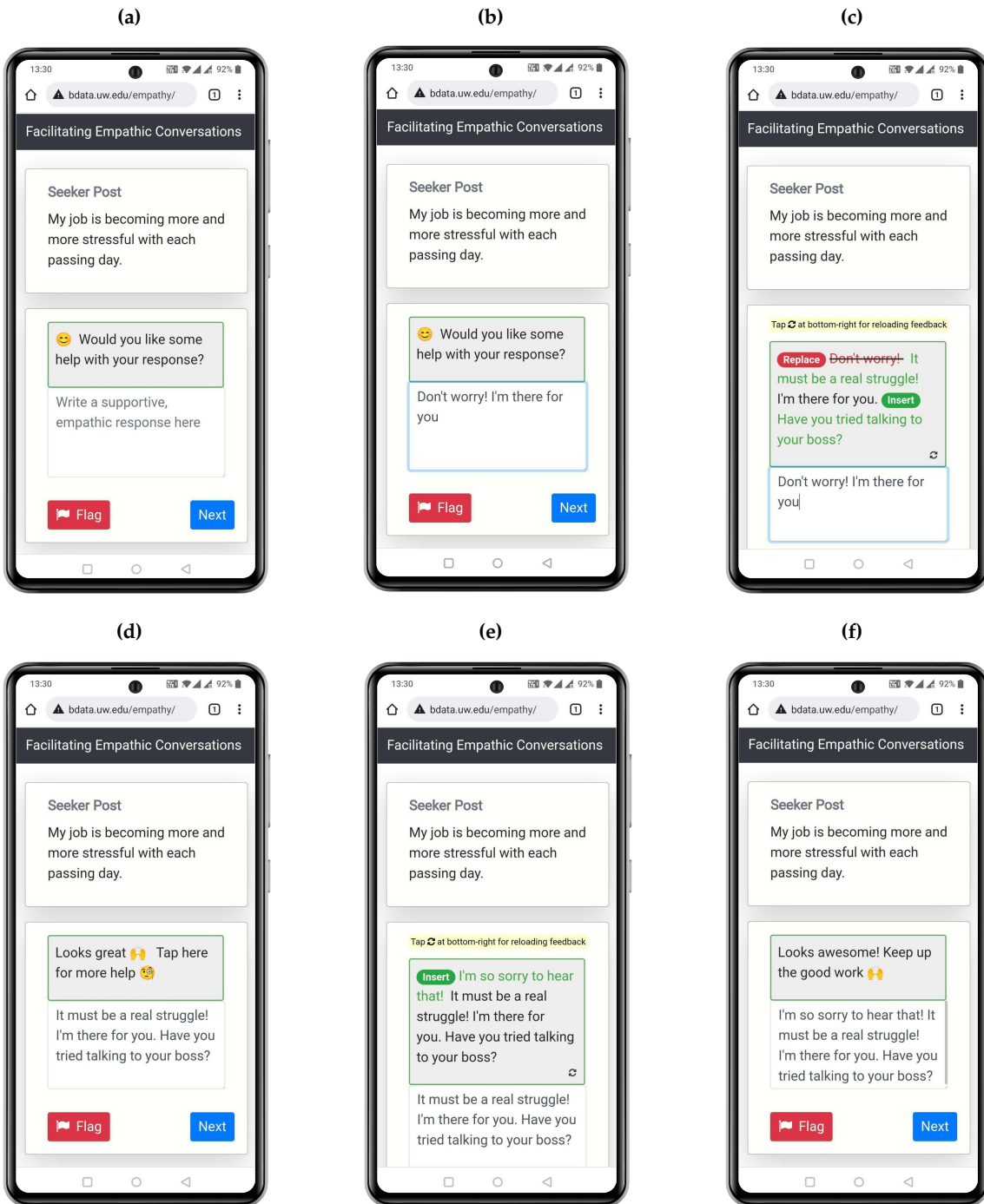
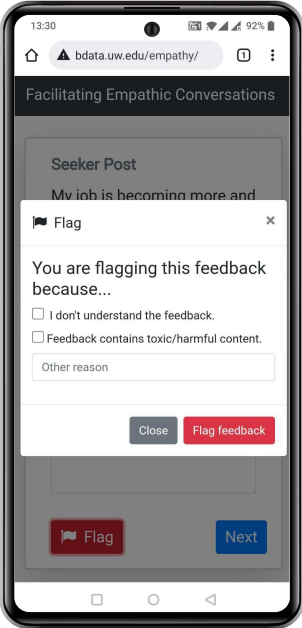


Figure A.32: Interface for flagging feedback [phase III: write supportive, empathic responses].



**Figure A.33:** Exit survey used for collecting perceptions of control group participants [phase IV: post-intervention survey].

### End-of-study Survey

1) It was **challenging** or **stressful** to write responses to posts.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

2) **Feedback** and **Suggestions** on how to improve my responses would have helped me write **better responses**.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

3) Describe the challenges faced while writing responses?

**Figure A.34:** Exit survey used for collecting perceptions of treatment group participants [phase IV: post-intervention survey]. Continued on the next page (1/2).

### End-of-study Survey

1) It was **challenging** or **stressful** to write responses to posts.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

2) The **feedback** shown to me was helpful in **improving** my responses.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

3) The **feedback** shown to me was helpful in making my responses more **empathic**.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

4) The **feedback** shown to me was **easy to incorporate** into the final response.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

5) I feel **more confident** at writing supportive responses after this study.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

6) I would like to see this type of feedback system **deployed on [TalkLife](#)** or other similar platforms.

Strongly Agree  
 Somewhat Agree  
 Neutral  
 Somewhat Disagree  
 Strongly Disagree

7) Describe the challenges faced while writing responses?

**Figure A.35:** Exit survey used for collecting perceptions of treatment group participants [phase IV: post-intervention survey] (2/2).

8) Describe instances where feedback was helpful?  
Why?

9) Describe instances where feedback was not helpful? How could they have been more helpful?

Submit

Figure A.36: Consent form used for human evaluation of responses.

## Disclaimer

Thank you for your interest in our study!

- We are **researchers at the University of Washington**, studying peer-to-peer support platforms.
- This study is **not being conducted by TalkLife**. TalkLife is not responsible for any risks/benefits associated with this study.
- As part of this study, we will be collecting **ratings to online mental health support interactions**. The data from this study will be uploaded to a **secure platform accessible only to the research team**. All collected data will be exclusively used for research purposes. Also, data will only be analyzed in aggregate.
- This study has been determined to be exempt from IRB approval under University of Washington IRB ID STUDY00012706. For concerns or questions, please contact [ashshar@cs.washington.edu](mailto:ashshar@cs.washington.edu).
- Your participation in this study is **completely voluntary**. You are free to release/quit the study at any time. Refusing to be in the experiment or stopping participation will involve no penalty or loss of benefits to which you are otherwise entitled.
- Poor-quality data (e.g., same answers to all posts) **may be removed** without compensation.
- If you are a US citizen or a permanent US resident, you will receive an **Amazon gift card worth 5 USD** after completing this study.
  - Yes**, I'm a US citizen or a permanent US resident.
  - No**, I'm neither a US citizen nor a permanent US resident.
  - Prefer not to say.
- **Note: Top-2 participants** (based on inter-rater agreement) will receive an additional gift card worth 25 USD.

## Consent

I agree to participate in this study. I also understand that only US citizens or permanent US residents can be compensated.

Accept and Continue

Figure A.37: Instructions for human evaluation of responses.

## Instructions

Read and evaluate online mental health support interactions.

### Content Warning

The study contains posts including but not limited to self-harm and suicidal ideation, which may be disturbing to you. If you have concerns or questions, please send us an email ([ashshar@cs.washington.edu](mailto:ashshar@cs.washington.edu)). If you have strong negative reactions to some of the content, please reach out at [crisis text line](#).

### What will you read?

You will be shown three posts to read - a seeker post and two response posts. Details below:

- **Seeker Post:** This would be typically a mental health support seeking post, posted online by a user in distress. The writer of this post is called *Seeker*.
- **Response Post A:** This is a response/reply posted in response to the seeker post, usually in an attempt to provide mental health support to the seeker.
- **Response Post B:** This is *another* response/reply posted in response to the *same* seeker post, usually in an attempt to provide mental health support to the seeker.

### What will you do?

Your task will be to evaluate the two response posts A and B. In particular, for each set of posts, you will answer the following questions:

- Which response is more **empathic** (regardless of appropriateness)?
  - Response A
  - Both are similar
  - Response B
- Which response is more **appropriate/relevant** to the seeker post (regardless of empathy)?
  - Response A
  - Both are similar
  - Response B
- Response A is...
  - Written by a human
  - Generated by a computer
  - Combination of both
- Response B is...
  - Written by a human
  - Generated by a computer
  - Combination of both

You will read **30** such posts and answer the associated questions. **We discourage the use of the "Both are similar" option. Only use it when the posts are actually similar and there is nothing to distinguish the two.**

Next

**Figure A.38:** Interface for human evaluation of responses.

**Seeker Post**

{{seeker\_post}}

**Response Post A**

{{response\_post\_A}}

**Response Post B**

{{response\_post\_B}}

Which response is more **empathic** (regardless of appropriateness)?

Response A  Both are similar  Response B

Which response is more **appropriate/relevant** to the seeker post (regardless of empathy)?

Response A  Both are similar  Response B

Response A is...

Written by a human  Generated by a computer  Combination of both

Response B is...

Written by a human  Generated by a computer  Combination of both

Next (1/30)



## Chapter B

# Supplementary Materials for Human-AI Collaboration to Empower People Seeking Support

### B.1 Method

#### Linguistic Attributes of Reframed Thoughts

We provide additional detail on the linguistic attributes of reframed thoughts.

**Actionability.** We measure actionability using:  $contains\_action(\mathbf{R}_i)$ , and  $next\_action\_coherence(\mathbf{R}_i)$ .

For  $contains\_action(\mathbf{R}_i)$ , our few-shot in-context learning approach proceeds as follows. Using the reframed thoughts that were annotated as high or low actionable in our collected data, we manually create 10 demonstration examples. If a reframed thought contains an action, we ask GPT-3 to extract the action from it. Otherwise, we ask it to generate the text “No Action”. Appendix B.1 shows examples. We then use these 10 demonstrations as in-context examples, followed by the reframe  $\mathbf{R}_i$  which we aim to classify. If GPT-3 predicts an action for  $\mathbf{R}_i$ , we assign  $contains\_action(\mathbf{R}_i) = 1$ ; else we assign 0.

For  $next\_action\_coherence(\mathbf{R}_i)$ , we instruct GPT-3 to generate  $k = 5$  possible next actions given a reframed thought. Given  $(\mathbf{S}_i, \mathbf{T}_i, \mathbf{R}_i)$ , let  $\mathbf{A}_i = a_{i1}, a_{i2}, \dots, a_{ik}$  be the generated set of next actions. Let  $emb(\cdot)$  denote RoBERTa embeddings. Then, we define  $next\_action\_coherence(\mathbf{R}_i)$  as

the average cosine similarity between  $emb(a_i)$  and  $emb(a_j)$  for all  $a_i, a_j \in \mathbf{A}_i$ .

## Action Generation Prompt

We use the following prompt template for extracting actions through GPT-3:

**Statement:** “My bank card could be in many different places and I want to check them first before making any conclusions”

**Proposed Action:** “Check bank card.”

**Statement:** “I cancelled that trip because I had to. It hurts to have done so but it was the right thing”

**Proposed Action:** None

Also, we use the following instruction prompt for generating the next set of actions through GPT-3: “Suggest 5 actions that the person could take based on the following statement:”

## Hyperparameter Choices for our Proposed Retrieval-Enhanced In-Context Learning Method

For the number of examples to retrieve, we experimented with  $k = 1, 5, 10$  and  $20$  and found  $k = 5$  to generate the most effective reframed thoughts, based on a qualitative assessment of 100 manually written situations and thoughts.

## B.2 Reproducibility

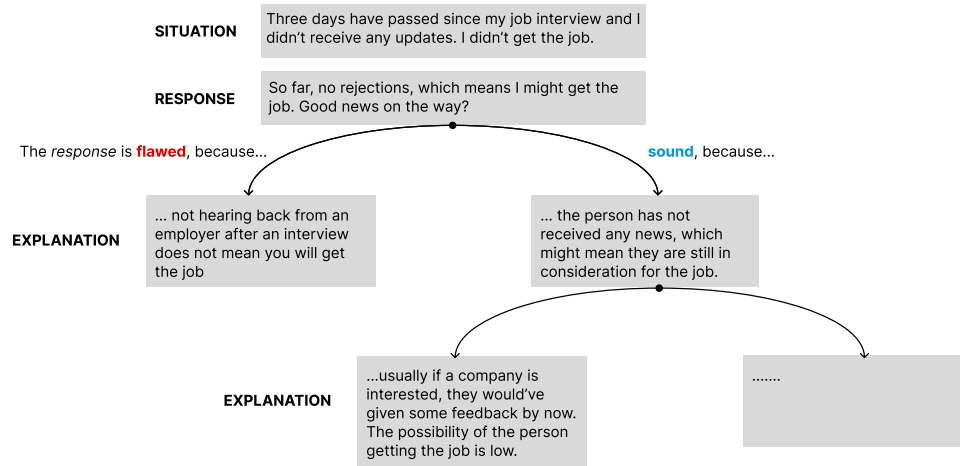
Codes and datasets created in the paper can be found at <https://github.com/behavioral-data/Cognitive-Reframing> under an academic, attribution-only license. The use of existing artifacts was consistent with their intended use. For GPT-3 based models, we will use the OpenAI library. For other deep learning models, we train and them on two NVIDIA Titan RTX GPUs. We use the evaluate python library ([pypi.org/project/evaluate](https://pypi.org/project/evaluate)) for measuring BLEU and ROUGE scores and scipy for statistical tests.

## B.3 List of Thinking Traps

Thinking Trap	Definition	Example	Tips to Overcome
All-or-Nothing Thinking	Thinking in extremes.	"If it isn't perfect, I failed. There's no such thing as 'good enough'."	Things in life are rarely black and white. Focus on what's positive or neutral about the situation.
Overgeneralizing	Jumping to conclusions based on one experience.	"They didn't text me back. Nobody ever texts me back."	Recall times when things went well for you. Imagine what it would be like for things to go well next time.
Labeling	Defining a person based on one action or characteristic.	"I said something embarrassing. I'm such a loser."	Consider all different aspects of a person.
Fortune Telling	Trying to predict the future. Focusing on one possibility and ignoring other, more likely outcomes.	"I'm late for the meeting. I'll make a fool of myself."	Be curious about what's going to happen next. Focus on what you can control and let go of what you can't.
Mind Reading	Assuming you know what someone else is thinking.	"She didn't say hello. She must be mad at me."	Try to imagine other, less negative possibilities. Try to ask the person what they're thinking, rather than just assuming.
Emotional Reasoning	Treating your feelings like facts.	"I woke up feeling anxious. I just know something bad is going to happen today."	Consider all the information you have.
Should Statements	Setting unrealistic expectations for yourself.	"I shouldn't need to ask for help. I should be independent."	Think about where your unrealistic expectations came from. Let your mistakes be an opportunity to learn and grow.
Personalizing	Taking things personally or making them about you.	"He's quiet today. I wonder what I did wrong."	Think about all the other things that could be affecting someone's behavior.
Disqualifying the Positive	When something good happens, you ignore it or think it doesn't count.	"I only won because I got lucky."	Go out of your way to notice the positive side.
Catastrophizing	Focusing on the worst-case scenario.	"My boss asked if I had a few minutes to talk. I'm going to get fired!"	Keep in mind that worst-case scenarios are very unlikely. Try to remind yourself of all the more likely, less severe things that could happen.
Comparing and Despairing	Comparing your worst to someone else's best.	"My niece's birthday party had twice the amount of people"	Remember that what you see on social media and in public is everyone showing off their best.
Blaming	Giving away your own power to other people.	"It's not my fault I yelled. You made me angry!"	Take responsibility for whatever you can—no more, no less.
Negative Feeling or Emotion	I'm having a negative feeling or emotion which isn't a thinking trap.	"I am feeling lonely."	Feeling negative emotions is a normal part of life. Think about what we can control, and what positive things we can be grateful for.

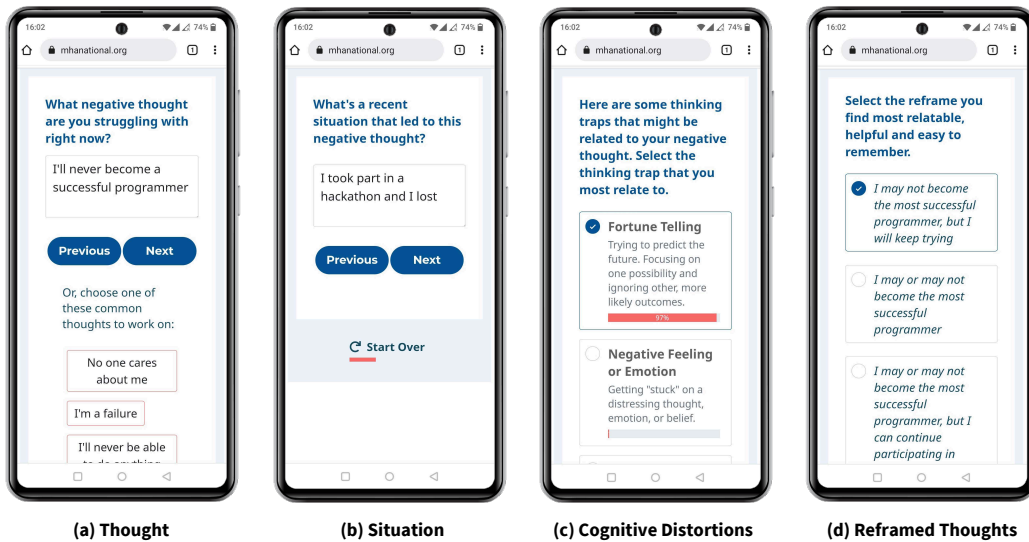
## B.4 Example Illustrating Our Rationality Measurement

**Figure B.1:** To measure reasoning strength, we generate two explanations for each reframe – one for why it might be sound; another for why it may be flawed. To check if the explanations themselves are well-reasoned, we recursively generate explanations for the explanations. Here, we choose a recursive tree depth of 3. Also, at every step, we generate three explanations in favour of a reframe and three explanations against it.



## B.5 Randomized Field-Study Interface

**Figure B.2:** Illustration of the interface used for our randomized field-study: **(a)** Participant starts by writing the negative thought they are struggling with in the moment; **(b)** We ask the participant to describe a recent situation that may have led to their thought; **(c)** An AI model identifies possible cognitive distortion(s) in the thought. Participant selects the cognitive distortions that they most relate to; **(d)** An AI model generates and suggests three different reframed thoughts that may help overcome negative thinking and the associated cognitive distortion. Participant selects the reframe they find the most relatable, helpful and memorable. Some of the instructions provided to the participants, including informed consent and evaluation, have been omitted from this illustration for brevity.



## **B.6 Data Collection Instructions**

**Figure B.3:** Instructions shown during data collection with mental health experts. Continued on the next page (1/3).

# Cognitive Restructuring

## Study Goals

The goal of this study is to collect a dataset for cognitive restructuring.

## Definitions

**Situation** Anything that happens to the person or the circumstance that the person finds themselves in (e.g., "My boss walked past me the hallway without saying hello")

**Thought** What goes through the person's mind in the situation (e.g., "Why are they angry with me?").

**Thinking Traps** [Thinking traps](#), also called *cognitive distortions*, are **exaggerated, biased, and irrational** thoughts which cause individuals to perceive reality inaccurately.

**Thinking Trap Categories** Categories of thinking traps include assuming what others think ("Mind Reading"), thinking in extremes ("All-or-nothing thinking"), focusing on the worst-case scenario ("Catastrophizing"), focusing only on the bad ("Disqualifying the positive"), etc.

### Example Thinking Trap

**Situation:** *My boss walked past me the hallway without saying hello*

**Thought:** *Why are they angry with me?*

**Thinking Trap:** Mind Reading

**Figure B.4:** Instructions shown during data collection with mental health experts. Continued on the next page (2/3).

Here, I'm reading my boss's mind and assuming that they are upset with me. I can't figure this out unless I ask them.

**Thought** A thought response is self-talk (conversation with oneself) that tries to challenge the thinking trap in the original thought

#### Example Thought Responses

**Situation:** *My boss walked past me the hallway without saying hello.*

**Thought:** *Why are they angry with me?*

**Response 1:** *I have no way of figuring out what they might be thinking. Maybe they had a lot on her mind*

**Response 2:** *They are the most wonderful person I know. They must not have noticed me.*

**Response 3:** *They might be mad at me, but atleast they didn't say anything.*

**Cognitive Restructuring** Cognitive Restructuring is a process that helps people notice thinking traps in their thoughts and respond rationally to them.

### Study Steps

In this study, you will perform 20 cognitive restructuring tasks. In each task, you will be shown a situation and a thought. You will be asked to identify thinking traps in the thought and write and rate thought responses.

**Note:** The use of "Both are similar" option (wherever applicable) is discouraged. Use it only when the two responses are truly identical and there is nothing to distinguish the two.

**Figure B.5:** Instructions shown during data collection with mental health experts (3/3).

### Content Warning

This study contains situations and thoughts including but not limited to self-harm and suicidal ideation, which may be disturbing. If you have any questions or concerns, please send us an email. Should you have a strong negative reaction to some of the content, you can reach a crisis counselor at [crisis text line](#) or by texting HOME to 741741.

If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study with someone other than the researcher(s), please contact the Human Subjects Division at xxx.

## **B.7 Consent Form Used in the Randomized Field Study**

**Figure B.6:** Consent form shown to the MHA visitors. Continued on the next page (1/2).

## Consent Form

This tool uses artificial intelligence to generate reframed thoughts and is part of a research study in collaboration with the University of Washington.

**Purpose:** The purpose of the study is to understand how digital tools can help people recognize thinking traps and practice reframing negative thoughts.

**Procedure:** You will be asked to describe a thought and a situation you are struggling with. You will then identify potential "thinking traps" (or cognitive distortions) in the thought and reframe it in a way that is more positive, realistic, or helpful. Finally, you will be asked to take an *optional* demographic survey, which can be skipped as preferred. The tool is expected to take ~5 minutes to complete.

**Benefits:** By using this tool, you may learn about thinking traps. You will practice identifying them and reframing negative thoughts and situations. However, there is no guarantee that the tool will help you reframe your thoughts.

**Data Collection and Sharing:** We will not ask you for your name or any identifiable personal information. Usage data will be made unidentifiable to the best of our extent, will be analyzed to improve the tool, and may be shared and used for future research.

**Risks:** Talking about situations and thoughts you are struggling with may be disturbing to you and may bring up negative emotional reactions. In addition, the tool uses artificial intelligence to generate reframed thoughts. Appropriate steps have been taken to avoid harmful reframes, but there is a possibility that the generated content might be upsetting to you. Also, the *optional* demographic survey asks for information that may be sensitive and could make you feel uncomfortable (e.g., "*What are the main things contributing to your mental health problems right now?*"). This tool is not being actively monitored by a human and should not be used as a "cry for

**Figure B.7:** Consent form shown to the MHA visitors (2/2).

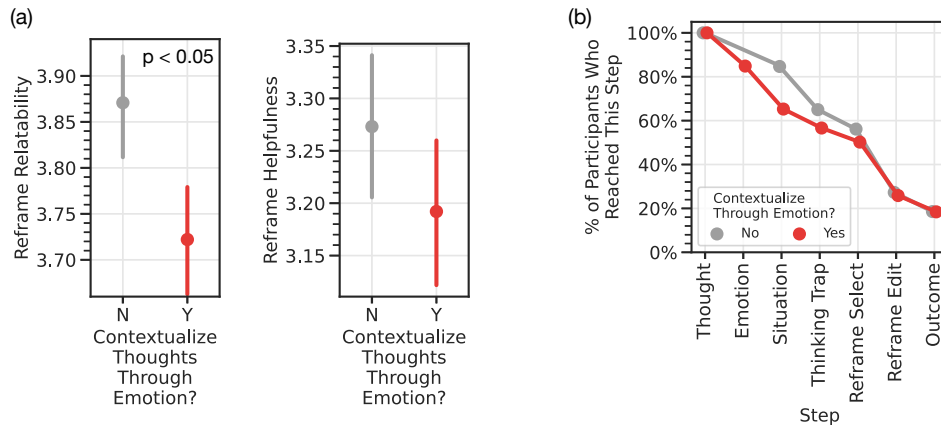
help" outlet. Should you have a strong negative reaction to some of the content, you can text MHA to 741741 or call or text 988.

**Participation:** Participation in this study is completely voluntary. You will not receive any payment for participation. You can refuse participation or stop participating at any time without penalty or loss of benefits to which you are otherwise entitled.

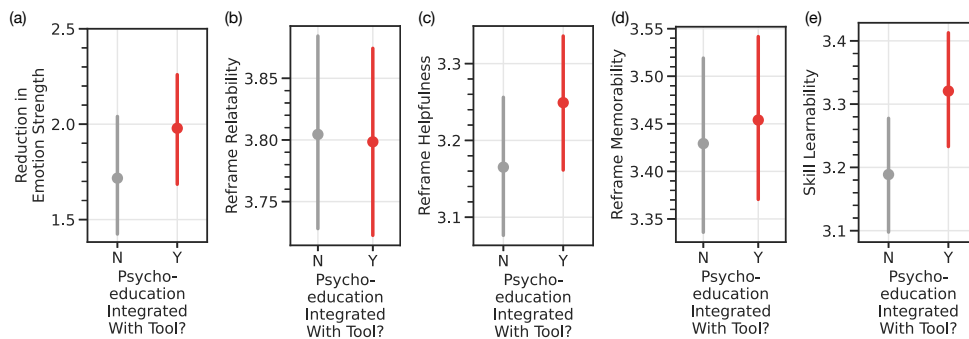
**Contact Us:** If you have questions or concerns about this research, or if you think you have been harmed from being in the study, please email us (Ashish Sharma at ashshar@cs.washington.edu, or Professor Tim Althoff at althoff@cs.washington.edu). If you have questions about your rights as a research participant, you can call the University of Washington Human Subjects Division at (206) 543-0098.

- By ticking this box, you are agreeing to use this tool. You are also confirming that you are at least 13 years old. Be sure that questions about the tool have been answered and that you understand what you are being asked to do. You may contact us if you think of a question later. You are free to stop using the tool at any time. To save a copy of this consent form, you can use [this link](#).

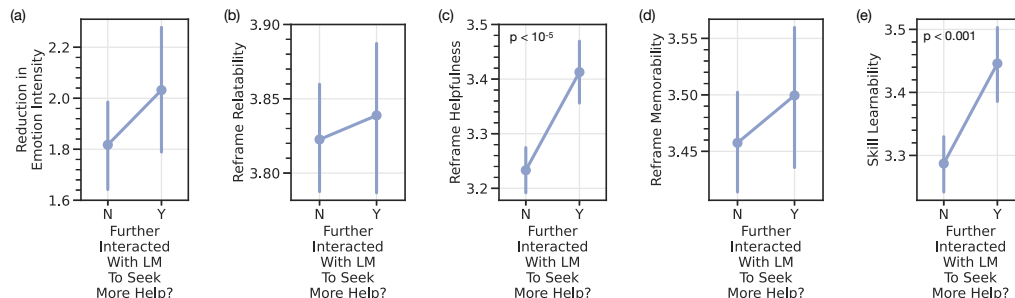
**Figure B.8:** Randomized controlled trial to estimate the effects of contextualizing thoughts through emotions (N=4,016). (a) Contextualizing participant thoughts through their emotions led to 3.86% lower relatable reframes (3.87 vs. 3.72). Note that our language model does not necessarily incorporate emotions unless they are expressed in the thought or situation as well. (b) Asking for additional context did not lead to lower completion rate. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .



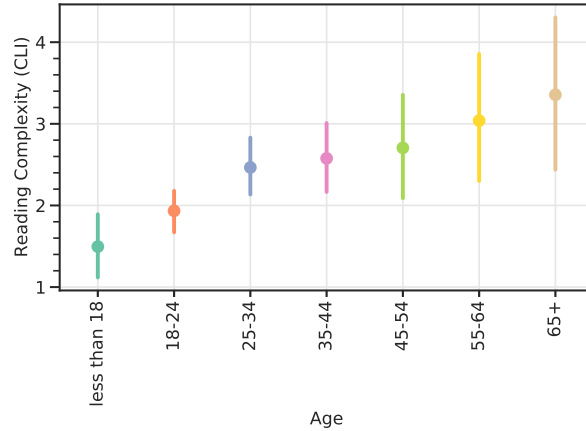
**Figure B.9:** Randomized controlled trial to estimate the effects of integrating psychoeducation (N=1,850). We did not find significant quantitative improvement in outcomes on integrating psychoeducation. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .



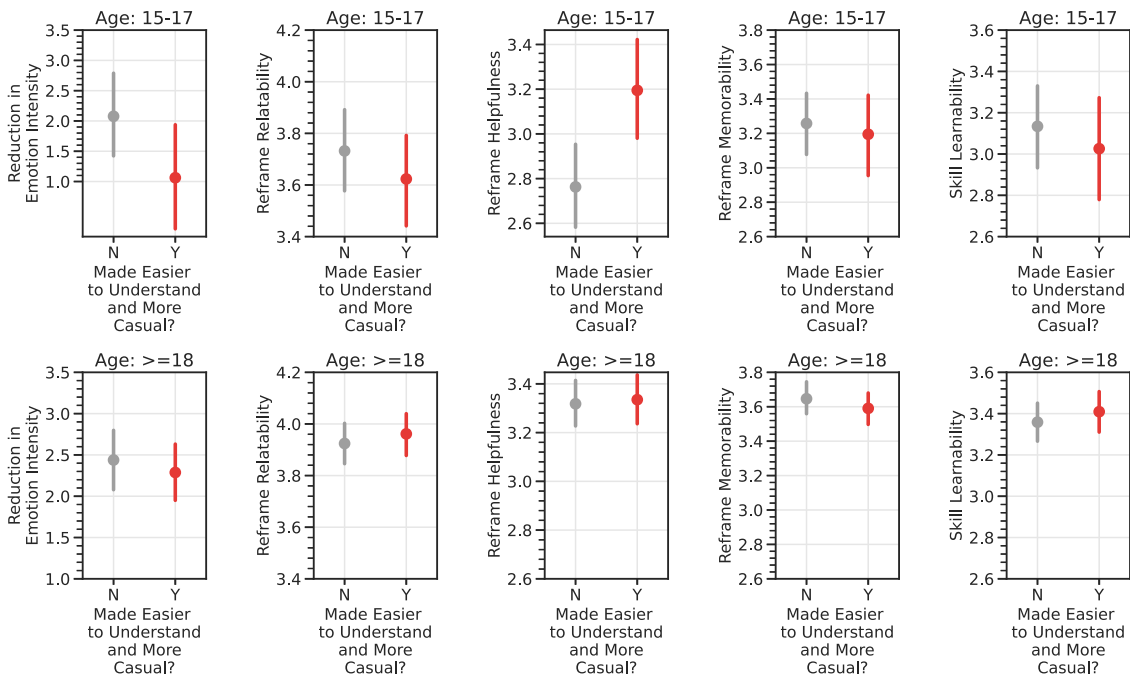
**Figure B.10:** Participants who further interacted with the language model to seek additional reframing suggestions of specific types (actionable, empathic, or personalized) reported 5.57% higher reframe helpfulness (3.41 vs. 3.23) and 4.86% higher skill learnability (3.45 vs. 3.29) and no significant differences in reduction in emotion intensity, reframe relatability, and reframe memorability (at  $\alpha = 0.05$ ; N=992). Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ . For a randomized trial assessing the effects of this intervention, see Figure 3.5.



**Figure B.11:** Reading Complexity (Coleman–Liau Index) of the thoughts written by participants based on their age. We find that adolescents (those below the age of 18) write thoughts with the least reading complexity. Error bars represent 95% bootstrapped confidence intervals.



**Figure B.12:** Randomized controlled trial to estimate the effects of making reframes easier to understand and more casual on age groups 15 to 17 and  $\geq 18$ . Participants in the age group 15 to 17 reported 15.58% higher reframe helpfulness (3.19 vs. 2.76) when they were suggested such reframes. We did not find significant differences for participants in age group  $\geq 18$  due to this intervention. Error bars represent 95% bootstrapped confidence intervals. Effects without p-values were not significant at  $\alpha = 0.05$ .



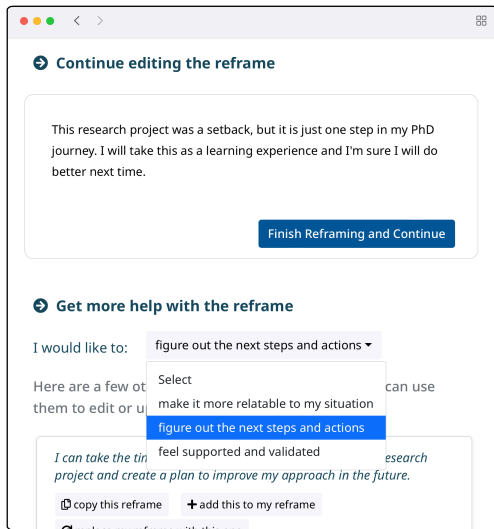
**Table B.1:** Definitions and examples of the list of issues identified in our open coding process.

Issue	Definition	Example Thought
Body image	Feeling ugly, dieting, disordered eating	"I'm fat and ugly"
Dating & marriage	Insecurities around dating and sexuality ("No one will date me" or "I got rejected") or specific situations involving a significant other ("My husband argues with me" or "My wife left me"). Does not include abuse!	"I am scared my girlfriend is going to break up with me because I dont do much for her."
Family	Specific situations involving interactions with family. If the only family member involved is the person's child, that's usually classified as parenting.	"I am the worst daughter in the world."
Fear	Intrusive thoughts about bad things happening; worrying about the future; imagining worst-case scenarios	"Something will go wrong on my flight today"
Friendship	Specific situations involving interactions with specific people. Doesn't have to be hyper-specific, but different from "loneliness" which is more about just the general concept of feeling isolated or unloved.	"My friend doesn't like me anymore."
Habits	Drugs & alcohol, addiction, or just any habits the person is trying to break.	"I made a goal to quit smoking and failed. I keep failing at this"
Health	Real or imagined illness; access to healthcare. Does not include mental health.	"I might be having a serious illness"
Hopelessness	Feeling like things will never get better; feeling like there's no point in trying	"I have lost all hope"
Identity	Discrimination due to race, gender, sexual orientation, etc; or coming to terms with being LGBTQ+.	"I'm not who I want to be. I hate my appearance and my voice."
Loneliness	Not having friends; being isolated from loved ones; general social anxiety	"Why don't i have friends?"
Money	Financial troubles. Includes being jobless if the main worry is finances.	"My financial situation is going out of hand. I'm worried about my future"
Parenting	Feeling like a bad parent, or worried about becoming a parent	"I'm a bad mom"
School	Bad grades, getting into college, fears of graduating, etc.	"I will fail my exam"
Tasks & achievement	Worrying about not being good enough. Examples: hobbies, chores, executive function.	"I'm not good enough"
Trauma	Violence, rape, verbal or physical abuse, etc.	"I can't get through this"
Work	Situations in the workplace. Includes being jobless if the main worry is competency.	"I'm late for the meeting. This shows what a jerk I am"

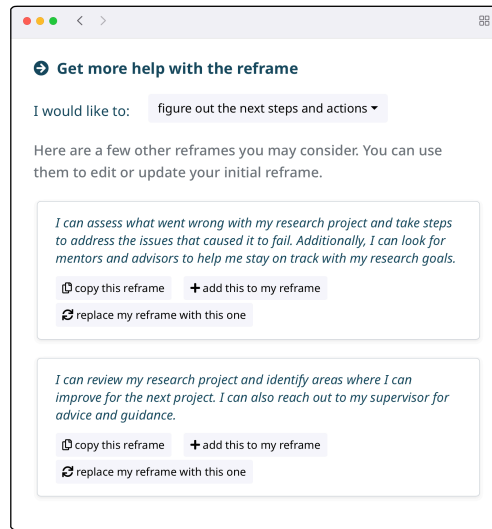
**Table B.2:** Examples illustrating the rewriting of reframes to make them simpler and casual for adolescents. Simpler and casual reframes use more straightforward words and phrases (e.g., “valuable” → “awesome”, “things I have accomplished” → “stuff I’ve already done”, “potential I have” → “how much more I can do”, “improve my physical health” → “make my body healthier”) and use informal language (e.g., “But you know what?”, “Plus...”). This could potentially be more appealing especially to adolescents.

Thought	Original Reframe	Simpler and Casual Reframe
I’m worthless. I haven’t done anything valuable in my life.	I may not have achieved all of my goals yet, but I’m still valuable and capable of doing great things. I can focus on the things I have accomplished and the potential I have to achieve even more.	I haven’t reached all my goals yet, but that’s okay! I’m still awesome and can do awesome things. I can think about the stuff I’ve already done and how much more I can do in the future.
I will be alone my whole life. All of my friends are starting to be in relationships.	It’s normal to feel lonely and to worry about not having a partner. I can focus on developing meaningful relationships with people, and on loving and appreciating myself.	It’s totally normal to feel lonely sometimes and worry about not having a partner. But you know what? Instead of stressing about it, I can focus on building amazing friendships with people who really care about me. Plus, I can also learn to love and appreciate myself just the way I am.
I’m going to fail in school and my body will never change	I may not do well in school right now, but I can still make changes to my lifestyle and work hard to improve my grades. I can also take steps to improve my physical health.	Right now, school isn’t going so great for me. But that doesn’t mean I can’t do things to make it better. I can change the way I live and put in a lot of effort to make my grades improve. I can also do things to make my body healthier.

**Figure B.13:** Detailed interface and process for iterative edits of reframes through further interaction with the language model.



(a) To iteratively refine the reframes, we provide three options to choose from — “make it more relatable to my situation”, “figure out the next steps and actions”, and “feel supported and validated”. Each option enables further interaction with the language model



(b) For the option selected by the participant, a language model generates additional reframing suggestions. Participants can either copy this reframe, add it to their current reframe, replace their original reframe with it, or use it as an inspiration



## **Chapter C**

# **Supplementary Material for Evaluating AI-based Mental Health Interventions**

**Table C.1:** Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS by 40.0% while human therapists responds by 23.4% on average ( $P = 8.0 \times 10^{-73}$ , Cohen’s  $d = 0.4$ , two-sided Student’s  $t$ -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ( $n = 4893$ )			GPT-3.5 ( $n = 4746$ )			Llama2-70b ( $n = 4970$ )			Llama2-13b ( $n = 4820$ )			Average ( $n = 5446$ )	High-quality ( $n = 3907$ )	Low-quality ( $n = 1539$ )
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
Reflections on Needs	40.0 <sub>49.0</sub>	$8.0 \times 10^{-73}$	0.4	32.3 <sub>46.8</sub>	$2.3 \times 10^{-22}$	0.2	36.0 <sub>48.0</sub>	$6.7 \times 10^{-44}$	0.3	34.5 <sub>47.5</sub>	$4.6 \times 10^{-34}$	0.2	23.4 <sub>43.1</sub>	26.2 <sub>44.0</sub>	20.6 <sub>40.5</sub>
Reflections on Emotions	13.8 <sub>34.5</sub>	$4.5 \times 10^{-6}$	0.1	19.4 <sub>39.6</sub>	$2.3 \times 10^{-34}$	0.2	14.9 <sub>35.6</sub>	$6.3 \times 10^{-10}$	0.1	14.4 <sub>35.1</sub>	$4.8 \times 10^{-8}$	0.1	10.5 <sub>32.8</sub>	14.6 <sub>35.3</sub>	6.3 <sub>24.3</sub>
Reflections on Values	5.5 <sub>22.7</sub>	$1.0 \times 10^0$	0.0	6.2 <sub>24.1</sub>	$5.0 \times 10^{-2}$	0.1	4.8 <sub>21.4</sub>	$1.0 \times 10^0$	0.0	5.7 <sub>23.1</sub>	$8.6 \times 10^{-1}$	0.0	4.8 <sub>22.8</sub>	6.3 <sub>24.4</sub>	3.3 <sub>17.9</sub>
Reflections on Consequences	23.4 <sub>42.4</sub>	$5.4 \times 10^{-1}$	0.0	14.5 <sub>35.2</sub>	$5.1 \times 10^{-21}$	0.0	22.4 <sub>41.7</sub>	$1.0 \times 10^0$	0.0	24.0 <sub>42.7</sub>	$8.2 \times 10^{-2}$	0.1	21.8 <sub>40.1</sub>	17.9 <sub>38.4</sub>	25.7 <sub>43.7</sub>
Reflections on Conflicts	5.9 <sub>23.5</sub>	$1.0 \times 10^0$	0.0	4.8 <sub>21.4</sub>	$1.5 \times 10^{-3}$	-0.1	8.1 <sub>27.3</sub>	$6.9 \times 10^{-2}$	0.1	7.8 <sub>26.8</sub>	$3.6 \times 10^{-1}$	0.0	6.6 <sub>26.1</sub>	8.3 <sub>27.6</sub>	5.0 <sub>21.8</sub>
Reflections on Strengths	39.1 <sub>48.8</sub>	$1.9 \times 10^{-281}$	0.7	32.1 <sub>46.7</sub>	$3.8 \times 10^{-176}$	0.6	24.0 <sub>42.7</sub>	$8.0 \times 10^{-85}$	0.4	21.9 <sub>41.3</sub>	$7.0 \times 10^{-64}$	0.3	9.5 <sub>31.3</sub>	12.8 <sub>33.5</sub>	6.2 <sub>24.2</sub>
Questions on Experiences	29.8 <sub>45.7</sub>	$2.1 \times 10^{-49}$	-0.3	29.9 <sub>45.8</sub>	$7.4 \times 10^{-48}$	-0.3	48.2 <sub>50.0</sub>	$2.4 \times 10^{-4}$	0.1	40.7 <sub>49.1</sub>	$1.2 \times 10^{-2}$	-0.1	44.0 <sub>49.9</sub>	48.9 <sub>50.0</sub>	39.1 <sub>48.8</sub>
Questions on Perspectives	5.9 <sub>23.5</sub>	$2.9 \times 10^{-7}$	0.1	2.4 <sub>15.4</sub>	$4.9 \times 10^{-2}$	-0.1	8.4 <sub>27.7</sub>	$3.2 \times 10^{-24}$	0.2	7.3 <sub>26.0</sub>	$1.0 \times 10^{-15}$	0.2	3.5 <sub>19.9</sub>	4.9 <sub>21.7</sub>	2.0 <sub>14.1</sub>
Questions on Emotions	2.2 <sub>14.7</sub>	$7.9 \times 10^{-75}$	-0.4	1.7 <sub>13.0</sub>	$8.4 \times 10^{-83}$	-0.4	8.9 <sub>28.5</sub>	$4.8 \times 10^{-5}$	-0.1	11.7 <sub>32.2</sub>	$1.0 \times 10^0$	0.0	11.7 <sub>33.0</sub>	13.3 <sub>34.0</sub>	10.1 <sub>30.2</sub>
Problem-Solving	57.8 <sub>49.4</sub>	$3.5 \times 10^{-256}$	0.7	49.2 <sub>50.0</sub>	$7.2 \times 10^{-137}$	0.5	45.8 <sub>49.8</sub>	$6.1 \times 10^{-104}$	0.4	34.4 <sub>47.5</sub>	$1.3 \times 10^{-20}$	0.2	25.9 <sub>42.6</sub>	21.2 <sub>40.9</sub>	30.5 <sub>46.0</sub>
Planning	57.1 <sub>49.5</sub>	$0.0 \times 10^0$	0.9	42.6 <sub>49.5</sub>	$1.2 \times 10^{-182}$	0.6	44.6 <sub>49.7</sub>	$1.6 \times 10^{-213}$	0.6	36.2 <sub>48.1</sub>	$7.1 \times 10^{-109}$	0.4	17.1 <sub>37.6</sub>	16.9 <sub>37.5</sub>	17.2 <sub>37.8</sub>
Normalizing	45.5 <sub>49.8</sub>	$2.9 \times 10^{-169}$	0.6	36.9 <sub>48.3</sub>	$1.3 \times 10^{-76}$	0.4	19.3 <sub>39.5</sub>	$1.0 \times 10^0$	0.0	24.9 <sub>43.3</sub>	$8.8 \times 10^{-7}$	0.1	20.5 <sub>39.5</sub>	17.8 <sub>38.3</sub>	23.2 <sub>42.2</sub>
Psychoeducation	16.4 <sub>37.0</sub>	$2.5 \times 10^{-45}$	0.3	24.5 <sub>43.0</sub>	$1.0 \times 10^{-130}$	0.5	12.6 <sub>33.2</sub>	$2.2 \times 10^{-17}$	0.2	8.9 <sub>28.4</sub>	$1.7 \times 10^{-1}$	0.0	7.6 <sub>24.8</sub>	5.2 <sub>22.3</sub>	9.9 <sub>29.8</sub>

**Table C.2:** Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS by 31.1% and GPT-4 with full conversation simulation responds by 42.5% when compared with the human therapists responds by 23.4% on average from Table C.1. (Single:  $P = 2.3 \times 10^{-6}$ , Cohen’s  $d = 0.2$ ; two-sided student t-test) (Full:  $P = 6.4 \times 10^{-84}$ , Cohen’s  $d = 0.4$ ; two-sided student t-test)

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4 ( $n = 3837$ )			GPT-3.5 ( $n = 3907$ )			Llama2-70b ( $n = 3906$ )			Llama2-13b ( $n = 3895$ )			GPT-4 ( $n = 1056$ )			GPT-3.5 ( $n = 839$ )			Llama2-70b ( $n = 1064$ )			Llama2-13b ( $n = 925$ )		
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$
Reflections on Needs	31.1 <sub>46.3</sub>	$2.3 \times 10^{-6}$	0.2	35.4 <sub>47.8</sub>	$2.0 \times 10^{-12}$	0.3	20.9 <sub>40.7</sub>	$1.0 \times 10^0$	-0.1	19.9 <sub>39.9</sub>	$2.7 \times 10^{-1}$	-0.1	42.5 <sub>49.4</sub>	$6.4 \times 10^{-84}$	0.4	31.6 <sub>48.5</sub>	$1.6 \times 10^{-17}$	0.2	40.1 <sub>49.0</sub>	$2.2 \times 10^{-66}$	0.4	38.0 <sub>48.5</sub>	$4.7 \times 10^{-51}$	0.3
Reflections on Emotions	8.4 <sub>27.8</sub>	$7.7 \times 10^{-1}$	-0.1	8.5 <sub>27.8</sub>	$1.0 \times 10^0$	-0.1	9.4 <sub>29.2</sub>	$1.0 \times 10^0$	0.0	9.8 <sub>29.8</sub>	$1.0 \times 10^0$	0.0	15.3 <sub>36.0</sub>	$2.0 \times 10^{-10}$	0.1	21.8 <sub>41.3</sub>	$1.9 \times 10^{-47}$	0.3	16.4 <sub>37.0</sub>	$6.3 \times 10^{-15}$	0.2	15.5 <sub>36.2</sub>	$3.1 \times 10^{-11}$	0.1
Reflections on Values	4.5 <sub>20.6</sub>	$1.0 \times 10^0$	0.0	5.1 <sub>22.1</sub>	$1.0 \times 10^0$	0.0	3.1 <sub>17.3</sub>	$2.5 \times 10^{-1}$	-0.1	3.5 <sub>18.3</sub>	$1.0 \times 10^0$	-0.1	5.7 <sub>23.3</sub>	$8.1 \times 10^{-1}$	0.0	6.4 <sub>24.5</sub>	$1.9 \times 10^{-2}$	0.1	5.3 <sub>22.4</sub>	$1.0 \times 10^0$	0.0	6.2 <sub>24.1</sub>	$7.3 \times 10^{-2}$	0.1
Reflections on Values	4.5 <sub>20.6</sub>	$6.1 \times 10^{-1}$	-0.0	5.1 <sub>22.1</sub>	$7.3 \times 10^{-1}$	0.0	3.1 <sub>17.3</sub>	$1.9 \times 10^{-2}$	-0.1	3.5 <sub>18.3</sub>	$8.2 \times 10^{-2}$	-0.1	5.7 <sub>23.3</sub>	$6.2 \times 10^{-2}$	0.0	6.4 <sub>24.5</sub>	$1.5 \times 10^{-3}$	0.1	5.3 <sub>22.4</sub>	$3.5 \times 10^{-1}$	0.0	6.2 <sub>24.1</sub>	$5.6 \times 10^{-3}$	0.1
Reflections on Conflicts	3.0 <sub>12.2</sub>	$2.0 \times 10^{-4}$	-0.1	2.0 <sub>14.1</sub>	$7.1 \times 10^{-6}$	-0.2	3.2 <sub>17.6</sub>	$4.7 \times 10^{-4}$	-0.1	3.6 <sub>18.6</sub>	$7.5 \times 10^{-3}$	-0.1	6.7 <sub>25.0</sub>	$1.0 \times 10^0$	0.0	5.4 <sub>22.6</sub>	$2.1 \times 10^{-1}$	-0.1	9.4 <sub>29.3</sub>	$1.6 \times 10^{-5}$	0.1	8.8 <sub>33.1</sub>	$1.9 \times 10^{-3}$	0.1
Reflections on Strengths	68.3 <sub>46.6</sub>	$0.0 \times 10^0$	1.7	54.6 <sub>49.8</sub>	$4.0 \times 10^{-249}$	1.3	35.4 <sub>47.9</sub>	$9.5 \times 10^{-106}$	0.7	37.7 <sub>48.5</sub>	$2.4 \times 10^{-112}$	0.8	31.1 <sub>46.3</sub>	$3.0 \times 10^{-151}$	0.6	27.3 <sub>44.5</sub>	$4.8 \times 10^{-109}$	0.5	20.9 <sub>40.7</sub>	$9.2 \times 10^{-51}$	0.3	18.1 <sub>38.5</sub>	$6.1 \times 10^{-31}$	0.2
Questions on Experiences	29.4 <sub>45.6</sub>	$1.4 \times 10^{-17}$	-0.3	20.9 <sub>40.7</sub>	$6.0 \times 10^{-36}$	-0.5	32.7 <sub>46.9</sub>	$1.3 \times 10^{-10}$	-0.2	35.5 <sub>47.9</sub>	$1.7 \times 10^{-5}$	-0.2	29.9 <sub>45.8</sub>	$4.5 \times 10^{-42}$	-0.3	31.8 <sub>46.6</sub>	$1.5 \times 10^{-31}$	-0.3	52.4 <sub>49.9</sub>	$1.3 \times 10^{-14}$	0.2	42.0 <sub>49.4</sub>	$7.3 \times 10^{-1}$	0.0
Questions on Perspectives	3.2 <sub>17.7</sub>	$1.0 \times 10^0$	0.0	2.0 <sub>14.1</sub>	$5.4 \times 10^{-1}$	-0.1	6.9 <sub>25.3</sub>	$1.7 \times 10^{-5}$	0.2	5.3 <sub>22.4</sub>	$1.5 \times 10^{-1}$	0.1	6.6 <sub>24.8</sub>	$2.7 \times 10^{-10}$	0.1	2.5 <sub>15.7</sub>	$1.8 \times 10^{-1}$	-0.1	8.8 <sub>33.1</sub>	$3.2 \times 10^{-25}$	0.2	7.7 <sub>26.7</sub>	$1.7 \times 10^{-17}$	0.2
Questions on Emotions	1.3 <sub>14.1</sub>	$8.1 \times 10^{-23}$	-0.3	2.0 <sub>14.1</sub>	$7.1 \times 10^{-16}$	-0.3	3.7 <sub>18.8</sub>	$1.6 \times 10^{-13}$	-0.3	5.4 <sub>22.6</sub>	$2.8 \times 10^{-7}$	-0.2	2.4 <sub>15.5</sub>	$3.6 \times 10^{-57}$	-0.3	1.7 <sub>12.8</sub>	$5.4 \times 10^{-71}$	-0.4	10.3 <sub>30.5</sub>	$5.1 \times 10^{-1}$	0.0	13.2 <sub>33.9</sub>	$3.8 \times 10^{-1}$	0.0
Problem-Solving	44.2 <sub>49.7</sub>	$4.2 \times 10^{-34}$	0.4	55.2 <sub>49.8</sub>	$1.9 \times 10^{-70}$	0.7	54.1 <sub>49.9</sub>	$4.9 \times 10^{-79}$	0.6	34.7 <sub>47.6</sub>	$1.3 \times 10^{-7}$	0.2	61.5 <sub>48.7</sub>	$3.2 \times 10^{-284}$	0.8	47.9 <sub>49.0</sub>	$7.5 \times 10^{-112}$	0.5	43.5 <sub>49.6</sub>	$8.9 \times 10^{-74}$	0.4	34.3 <sub>47.5</sub>	$2.8 \times 10^{-18}$	0.2
Planning	66.8 <sub>47.1</sub>	$4.7 \times 10^{-279}$	1.3	70.2 <sub>45.8</sub>	$2.4 \times 10^{-269}$	1.4	64.1 <sub>48.0</sub>	$7.9 \times 10^{-252}$	1.2	44.8 <sub>49.8</sub>	$1.8 \times 10^{-82}$	0.7	54.4 <sub>49.8</sub>	$0.0 \times 10^0$	0.9	36.7 <sub>48.2</sub>	$6.2 \times 10^{-104}$	0.5	39.3 <sub>48.8</sub>	$1.4 \times 10^{-130}$	0.5	34.1 <sub>47.4</sub>	$6.5 \times 10^{-81}$	0.4
Normalizing	65.7 <sub>47.5</sub>	$2.5 \times 10^{-218}$	1.1	70.8 <sub>45.5</sub>	$9.2 \times 10^{-227}$	1.2	22.5 <sub>41.8</sub>	$1.0 \times 10^0$	0.0	28.6 <sub>45.2</sub>	$1.9 \times 10^{-7}$	0.2	39.9 <sub>49.0</sub>	$5.4 \times 10^{-95}$	0.4	29.6 <sub>45.6</sub>	$1.7 \times 10^{-23}$	0.2	18.4 <sub>38.8</sub>	$1.5 \times 10^{-1}$	-0.1	24.0 <sub>42.7</sub>	$5.2 \times 10^{-4}$	0.1
Psychoeducation	20.5 <sub>40.4</sub>	$7.3 \times 10^{-42}$	0.5	25.3 <sub>43.5</sub>	$1.0 \times 10^{-62}$	0.6	26.9 <sub>44.4</sub>	$6.0 \times 10^{-85}$	0.7	18.1 <sub>38.5</sub>	$4.2 \times 10^{-26}$	0.4	15.2 <sub>36.0</sub>	$7.7 \times 10^{-33}$	0.3	24.4 <sub>43.0</sub>	$1.5 \times 10^{-121}$	0.5	8.7 <sub>28.2</sub>	$5.5 \times 10^{-1}$	0.0	6.7 <sub>25.0</sub>	$1.0 \times 10^0$	0.0

**Table C.3:** Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS at 3.1 turns earlier than average human therapists responds. (GPT-4: 3.1 turn, Average human therapists: 6.2 turn;  $P = 4.3 \times 10^{-15}$ , Cohen’s  $d = 0.7$ , two-sided Student’s  $t$ -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ( $n = 4893$ )			GPT-3.5 ( $n = 4746$ )			Llama2-70b ( $n = 4970$ )			Llama2-13b ( $n = 4820$ )			Average ( $n = 5446$ )	High-quality ( $n = 3907$ )	Low-quality ( $n = 1539$ )
	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	turn	turn
Reflections on Needs	3.1 <sub>2,9</sub>	$4.3 \times 10^{-15}$	0.7	6.4 <sub>5,4</sub>	$1.0 \times 10^0$	0.0	5.1 <sub>5,0</sub>	$2.5 \times 10^{-1}$	0.2	5.3 <sub>5,5</sub>	$9.1 \times 10^{-1}$	0.2	6.2 <sub>5,5</sub>	4.4 <sub>4,5</sub>	8.1 <sub>6,2</sub>
Reflections on Emotions	6.1 <sub>5,0</sub>	$1.0 \times 10^0$	0.0	5.7 <sub>4,8</sub>	$1.0 \times 10^0$	0.1	7.2 <sub>5,6</sub>	$1.0 \times 10^0$	-0.2	6.1 <sub>5,4</sub>	$1.0 \times 10^0$	0.0	6.3 <sub>4,9</sub>	5.2 <sub>4,8</sub>	7.4 <sub>5,0</sub>
Reflections on Values	8.3 <sub>5,1</sub>	$3.6 \times 10^{-1}$	0.3	9.1 <sub>5,5</sub>	$1.0 \times 10^0$	0.1	10.3 <sub>5,7</sub>	$1.0 \times 10^0$	-0.1	10.1 <sub>5,6</sub>	$1.0 \times 10^0$	0.0	9.9 <sub>5,5</sub>	8.8 <sub>5,3</sub>	11.0 <sub>5,8</sub>
Reflections on Consequences	6.5 <sub>4,2</sub>	$1.0 \times 10^0$	0.1	7.6 <sub>4,9</sub>	$1.0 \times 10^0$	-0.1	6.5 <sub>4,5</sub>	$1.0 \times 10^0$	0.1	7.0 <sub>5,4</sub>	$1.0 \times 10^0$	0.0	7.0 <sub>5,0</sub>	6.2 <sub>4,6</sub>	7.8 <sub>5,4</sub>
Reflections on Conflicts	7.2 <sub>4,6</sub>	$3.9 \times 10^{-4}$	0.6	8.1 <sub>4,6</sub>	$1.3 \times 10^{-1}$	0.4	7.8 <sub>4,9</sub>	$7.9 \times 10^{-3}$	0.4	7.8 <sub>5,3</sub>	$9.1 \times 10^{-3}$	0.4	10.0 <sub>5,2</sub>	8.2 <sub>4,7</sub>	11.8 <sub>5,6</sub>
Reflections on Strengths	7.9 <sub>5,4</sub>	$5.9 \times 10^{-4}$	0.4	9.1 <sub>6,3</sub>	$1.0 \times 10^0$	0.2	8.7 <sub>6,3</sub>	$2.1 \times 10^{-1}$	0.2	9.4 <sub>6,5</sub>	$1.0 \times 10^0$	0.1	10.2 <sub>6,1</sub>	8.4 <sub>6,0</sub>	11.9 <sub>5,7</sub>
Questions on Experiences	2.7 <sub>2,5</sub>	$1.0 \times 10^0$	0.1	3.7 <sub>4,3</sub>	$1.5 \times 10^{-1}$	-0.2	2.9 <sub>3,0</sub>	$1.0 \times 10^0$	0.0	3.8 <sub>4,0</sub>	$2.8 \times 10^{-2}$	-0.3	2.9 <sub>2,5</sub>	2.2 <sub>2,1</sub>	3.6 <sub>2,9</sub>
Questions on Perspectives	8.5 <sub>4,7</sub>	$3.1 \times 10^{-7}$	0.8	9.7 <sub>5,3</sub>	$1.6 \times 10^{-2}$	0.6	8.7 <sub>5,7</sub>	$4.2 \times 10^{-6}$	0.7	8.7 <sub>5,4</sub>	$2.8 \times 10^{-6}$	0.7	12.6 <sub>5,1</sub>	10.0 <sub>4,7</sub>	15.3 <sub>4,2</sub>
Questions on Emotions	6.2 <sub>4,9</sub>	$1.0 \times 10^0$	0.1	6.0 <sub>5,9</sub>	$1.0 \times 10^0$	0.2	5.3 <sub>6,5</sub>	$1.6 \times 10^{-1}$	0.3	5.4 <sub>5,5</sub>	$1.6 \times 10^{-1}$	0.3	7.0 <sub>6,4</sub>	6.7 <sub>6,3</sub>	7.3 <sub>6,5</sub>
Problem-Solving	7.8 <sub>5,2</sub>	$1.0 \times 10^0$	0.1	7.1 <sub>5,1</sub>	$1.5 \times 10^{-1}$	0.2	7.5 <sub>4,8</sub>	$7.1 \times 10^{-1}$	0.2	9.0 <sub>5,8</sub>	$1.0 \times 10^0$	-0.1	8.3 <sub>5,4</sub>	8.7 <sub>5,6</sub>	8.0 <sub>5,1</sub>
Planning	9.3 <sub>5,4</sub>	$4.2 \times 10^{-5}$	0.4	10.2 <sub>6,1</sub>	$3.8 \times 10^{-2}$	0.3	8.7 <sub>5,3</sub>	$7.2 \times 10^{-8}$	0.5	10.7 <sub>6,1</sub>	$6.0 \times 10^{-1}$	0.2	11.9 <sub>6,5</sub>	11.5 <sub>6,8</sub>	12.4 <sub>6,0</sub>
Normalizing	7.9 <sub>6,0</sub>	$9.0 \times 10^{-1}$	-0.2	7.0 <sub>6,0</sub>	$1.0 \times 10^0$	0.0	9.3 <sub>6,8</sub>	$2.8 \times 10^{-3}$	-0.4	9.0 <sub>6,3</sub>	$5.3 \times 10^{-3}$	-0.3	6.9 <sub>6,2</sub>	6.2 <sub>6,3</sub>	7.7 <sub>6,0</sub>
Psychoeducation	11.1 <sub>5,5</sub>	$2.5 \times 10^{-2}$	-0.4	8.2 <sub>5,7</sub>	$1.0 \times 10^0$	0.1	11.8 <sub>5,7</sub>	$2.3 \times 10^{-4}$	-0.5	11.5 <sub>5,8</sub>	$5.6 \times 10^{-3}$	-0.5	8.9 <sub>5,5</sub>	8.5 <sub>6,6</sub>	9.2 <sub>5,5</sub>

**Table C.4:** Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS at 4.3 turn which is 1.9 turns earlier than average human therapists. Similarly, GPT-4 with full conversation simulation responds at 2.0 turn which is 4.2 turns earlier than average human therapists. (Single: 4.3 turn, Average human therapists: 6.2 turn from Table C.3;  $P = 1.1 \times 10^{-3}$ , Cohen’s  $d = 0.4$ , two-sided Student’s  $t$ -test) (Full: 2.0 turn, Average human therapists: 6.2 turn from Table C.3;  $P = 9.2 \times 10^{-16}$ , Cohen’s  $d = 0.7$ ).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4 ( $n = 3837$ )			GPT-3.5 ( $n = 3907$ )			Llama2-70b ( $n = 3906$ )			Llama2-13b ( $n = 3895$ )			GPT-4 ( $n = 1056$ )			GPT-3.5 ( $n = 839$ )			Llama2-70b ( $n = 1064$ )			Llama2-13b ( $n = 925$ )					
	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$
Reflections on Needs	4.3 <sub>2,6</sub>	$1.1 \times 10^{-3}$	0.4	8.2 <sub>6,2</sub>	$1.7 \times 10^{-2}$	-0.3	8.3 <sub>5,1</sub>	$1.4 \times 10^{-2}$	-0.4	9.6 <sub>5,9</sub>	$6.5 \times 10^{-6}$	-0.6	2.0 <sub>2,8</sub>	$9.2 \times 10^{-16}$	0.9	4.7 <sub>3,9</sub>	$4.5 \times 10^{-2}$	0.3	2.9 <sub>3,4</sub>	$1.0 \times 10^{-9}$	0.7	2.4 <sub>2,5</sub>	$6.7 \times 10^{-14}$	0.8	4.8 <sub>5,1</sub>	$9.3 \times 10^{-3}$	-0.5
Reflections on Emotions	8.8 <sub>5,1</sub>	$9.3 \times 10^{-3}$	-0.5	7.8 <sub>5,3</sub>	$8.1 \times 10^{-1}$	-0.3	9.3 <sub>5,2</sub>	$3.1 \times 10^{-4}$	-0.6	8.8 <sub>5,3</sub>	$8.7 \times 10^{-3}$	-0.5	4.6 <sub>4,2</sub>	$3.3 \times 10^{-2}$	0.4	4.7 <sub>4,2</sub>	$7.1 \times 10^{-2}$	0.3	6.0 <sub>5,4</sub>	$1.0 \times 10^0$	0.1	4.8 <sub>5,0</sub>	$1.4 \times 10^{-1}$	0.1	9.3 <sub>5,5</sub>	$1.0 \times 10^0$	0.3
Reflections on Values	8.2 <sub>4,1</sub>	$1.0 \times 10^0$	0.3	9.5 <sub>6,7</sub>	$1.0 \times 10^0$	0.1	13.6 <sub>6,1</sub>	$2.7 \times 10^{-2}$	-0.7	13.1 <sub>5,3</sub>	$1.6 \times 10^{-1}$	-0.6	8.4 <sub>5,6</sub>	$9.3 \times 10^{-1}$	0.3	8.9 <sub>5,0</sub>	$1.0 \times 10^0$	0.2	9.2 <sub>5,5</sub>	$1.0 \times 10^0$	0.1	9.3 <sub>5,5</sub>	$1.0 \times 10^0$	0.1	9.3 <sub>5,5</sub>	$1.0 \times 10^0$	0.1
Reflections on Consequences	7.5 <sub>3,4</sub>	$1.0 \times 10^0$	-0.1	9.1 <sub>4,7</sub>	$3.7 \times 10^{-2}$	-0.4	9.0 <sub>4,6</sub>	$1.1 \times 10^{-1}$	-0.4	11.3 <sub>5,3</sub>	$6.5 \times 10^{-7}$	-0.8	6.0 <sub>4,6</sub>	$8.3 \times 10^{-1}$	0.2	6.7 <sub>4,8</sub>	$1.0 \times 10^0$	0.1	5.7 <sub>4,2</sub>	$1.0 \times 10^{-1}$	0.3	5.5 <sub>4,6</sub>	$4.5 \times 10^{-2}$	0.3	5.5 <sub>4,6</sub>	$4.5 \times 10^{-2}$	0.3
Reflections on Conflicts	7.1 <sub>3,5</sub>	$1.8 \times 10^{-1}$	0.6	6.9 <sub>2,3</sub>	$3.4 \times 10^{-1}$	0.6	8.0 <sub>3,2</sub>	$1.0 \times 10^0$	0.4	11.0 <sub>5,5</sub>	$1.0 \times 10^0$	-0.2	7.2 <sub>4,9</sub>	$2.7 \times 10^{-3}$	0.6	8.4 <sub>5,1</sub>	$6.7 \times 10^{-1}$	0.3	7.8 <sub>5,3</sub>	$1.7 \times 10^{-2}$	0.4	7.1 <sub>5,0</sub>	$2.1 \times 10^{-4}$	0.6	7.1 <sub>5,0</sub>	$2.1 \times 10^{-4}$	0.6
Reflections on Strengths	9.0 <sub>5,2</sub>	$1.0 \times 10^0$	0.2	11.4 <sub>6,1</sub>	$1.0 \times 10^0$	-0.2	11.2 <sub>6,1</sub>	$1.0 \times 10^0$	-0.2	12.2 <sub>6,6</sub>	$1.1 \times 10^{-1}$	-0.3	6.8 <sub>4,4</sub>	$5.4 \times 10^{-6}$	0.6	7.0 <sub>5,7</sub>	$7.2 \times 10^{-5}$	0.5	6.1 <sub>5,6</sub>	$1.1 \times 10^{-7}$	0.7	7.1 <sub>5,6</sub>	$1.1 \times 10^{-4}$	0.5	6.1 <sub>5,6</sub>	$1.1 \times 10^{-4}$	0.5
Questions on Experiences	3.8 <sub>2,1</sub>	$5.7 \times 10^{-3}$	-0.4	5.9 <sub>3,2</sub>	$8.3 \times 10^{-12}$	-0.9	4.5 <sub>3,5</sub>	$5.4 \times 10^{-8}$	-0.5	6.1 <sub>4,5</sub>	$6.0 \times 10^{-17}$	-0.9	1.7 <sub>2,3</sub>	$3.1 \times 10^{-5}$	0.5	2.0 <sub>2,5</sub>	$1.6 \times 10^{-2}$	0.3	1.4 <sub>1,2</sub>	$4.5 \times 10^{-10}$	0.7	1.6 <sub>1,3</sub>	$3.5 \times 10^{-8}$	0.6	1.6 <sub>1,3</sub>	$3.5 \times 10^{-8}$	0.6
Questions on Perspectives	8.0 <sub>5,1</sub>	$3.6 \times 10^{-3}$	0.9	6.3 <sub>2,0</sub>	$4.1 \times 10^{-2}$	1.3	11.2 <sub>4,5</sub>	$1.0 \times 10^0$	0.3	10.4 <sub>4,7</sub>	$4.4 \times 10^{-1}$	0.5	8.6 <sub>4,6</sub>	$3.9 \times 10^{-6}$	0.8	10.1 <sub>5,4</sub>	$1.0 \times 10^{-1}$	0.5	7.6 <sub>5,8</sub>	$3.4 \times 10^{-8}$	0.9	8.2 <sub>5,5</sub>	$7.4 \times 10^{-7}$	0.8	8.2 <sub>5,5</sub>	$7.4 \times 10^{-7}$	0.8
Questions on Emotions	4.9 <sub>3,1</sub>	$1.0 \times 10^0$	0.3	7.0 <sub>5,9</sub>	$1.0 \times 10^0$	0.0	6.9 <sub>6,0</sub>	$1.0 \times 10^0$	0.0	5.7 <sub>4,2</sub>	$1.0 \times 10^0$	0.2	6.4 <sub>5,1</sub>	$1.0 \times 10^0$	0.1	5.8 <sub>6,0</sub>	$1.0 \times 10^0$	0.2	4.9 <sub>5,5</sub>	$4.6 \times 10^{-2}$	0.4	5.3 <sub>5,9</sub>	$2.3 \times 10^{-1}$	0.3	5.3 <sub>5,9</sub>	$2.3 \times 10^{-1}$	0.3
Problem-Solving	10.3 <sub>4,9</sub>	$3.9 \times 10^{-3}$	-0.4	8.9 <sub>4,6</sub>	$1.0 \times 10^0$	-0.1	9.2 <sub>4,5</sub>	$1.0 \times 10^0$	-0.2	11.9 <sub>5,8</sub>	$8.8 \times 10^{-8}$	-0.7	5.4 <sub>4,3</sub>	$1.0 \times 10^{-6}$	0.6	5.6 <sub>5,0</sub>	$1.7 \times 10^{-5}$	0.5	5.7 <sub>4,5</sub>	$2.7 \times 10^{-5}$	0.5	6.5 <sub>4,4</sub>	$7.1 \times 10^{-3}$	0.4	6.5 <sub>4,4</sub>	$7.1 \times 10^{-3}$	0.4
Planning	11.4 <sub>5,1</sub>	$1.0 \times 10^0$	0.1	12.2 <sub>5,4</sub>	$1.0 \times 10^0$	0.0	9.8 <sub>5,1</sub>	$1.6 \times 10^{-2}$	0.4	12.7 <sub>6,5</sub>	$1.0 \times 10^0$	-0.1	7.5 <sub>5,0</sub>	$5.0 \times 10^{-10}$	0.7	8.4 <sub>6,2</sub>	$1.2 \times 10^{-5}$	0.5	7.7 <sub>5,3</sub>	$5.0 \times 10^{-9}$	0.7	9.0 <sub>5,2</sub>	$1.4 \times 10^{-4}$	0.5	9.0 <sub>5,2</sub>	$1.4 \times 10^{-4}$	0.5
Normalizing	10.3 <sub>6,2</sub>	$1.6 \times 10^{-5}$	-0.5	9.6 <sub>6,1</sub>	$2.4 \times 10^{-3}$	-0.4	12.4 <sub>6,7</sub>	$2.0 \times 10^{-10}$	-0.9	11.2 <sub>6,3</sub>	$1.3 \times 10^{-6}$	-0.7	5.7 <sub>4,9</sub>	$5.5 \times 10^{-1}$	0.2	4.7 <sub>5,0</sub>	$6.1 \times 10^{-3}$	0.4	6.7 <sub>5,7</sub>	$1.0 \times 10^0$	0.0	7.6 <sub>6,0</sub>	$1.0 \times 10^0$	-0.1	7.6 <sub>6,0</sub>	$1.0 \times 10^0$	-0.1
Psychoeducation	12.0 <sub>5,1</sub>	$7.3 \times 10^{-3}$	-0.6	12.2 <sub>5,4</sub>	$3.8 \times 10^{-3}$	-0.6	13.2 <sub>5,3</sub>	$3.3 \times 10^{-7}$	-0.8	13.5 <sub>5,9</sub>	$1.1 \times 10^{-5}$	-0.8	10.4 <sub>5,7</sub>	$7.0 \times 10^{-1}$	-0.3	6.2 <sub>4,7</sub>	$1.7 \times 10^{-3}$	0.5	10.1 <sub>5,8</sub>	$1.0 \times 10^0$	-0.2	9.6 <sub>5,1</sub>	$1.0 \times 10^0$	-0.1	9.6 <sub>5,1</sub>	$1.0 \times 10^0$	-0.1

**Table C.5:** Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON CONSEQUENCES by 19.9% and human therapists responds by 27.3% on average when client expresses CHANGING UNHEALTHY BEHAVIOR. ( $P = 4.1 \times 10^{-6}$ , Cohen’s  $d = -0.2$ , two-sided Student’s  $t$ -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
<b>Client behavior: Changing unhealthy behavior</b>															
<i>n</i>	2284			2148			2211			2082			1553	1198	355
Reflections on Needs	29.3 <sub>45.5</sub>	$1.0 \times 10^0$	0.0	29.1 <sub>45.4</sub>	$1.0 \times 10^0$	0.0	26.9 <sub>44.4</sub>	$1.0 \times 10^0$	0.0	27.8 <sub>44.8</sub>	$1.0 \times 10^0$	0.0	27.7 <sub>45.5</sub>	30.6 <sub>46.1</sub>	24.8 <sub>43.2</sub>
Reflections on Emotions	8.2 <sub>27.4</sub>	$1.0 \times 10^0$	-0.1	11.8 <sub>32.2</sub>	$1.0 \times 10^0$	0.0	8.7 <sub>28.2</sub>	$1.0 \times 10^0$	-0.1	8.6 <sub>28.0</sub>	$1.0 \times 10^0$	-0.1	10.4 <sub>33.3</sub>	14.6 <sub>35.3</sub>	6.2 <sub>24.1</sub>
Reflections on Values	4.1 <sub>19.9</sub>	$1.0 \times 10^0$	0.0	4.9 <sub>21.7</sub>	$1.0 \times 10^0$	0.0	3.6 <sub>18.6</sub>	$1.0 \times 10^0$	-0.1	4.5 <sub>20.8</sub>	$1.0 \times 10^0$	0.0	4.8 <sub>23.7</sub>	7.0 <sub>25.5</sub>	2.5 <sub>15.7</sub>
Reflections on Consequences	19.9 <sub>39.9</sub>	$4.1 \times 10^{-6}$	-0.2	12.3 <sub>32.9</sub>	$4.9 \times 10^{-31}$	-0.4	16.6 <sub>37.2</sub>	$3.1 \times 10^{-14}$	-0.3	19.6 <sub>39.7</sub>	$1.8 \times 10^{-6}$	-0.2	27.3 <sub>42.3</sub>	19.9 <sub>39.9</sub>	34.6 <sub>47.7</sub>
Reflections on Conflicts	4.1 <sub>19.9</sub>	$1.3 \times 10^{-2}$	-0.1	3.3 <sub>17.9</sub>	$5.2 \times 10^{-5}$	-0.2	4.4 <sub>20.5</sub>	$6.7 \times 10^{-2}$	-0.1	6.1 <sub>24.0</sub>	$1.0 \times 10^0$	0.0	6.9 <sub>26.6</sub>	8.3 <sub>27.5</sub>	5.6 <sub>23.1</sub>
Reflections on Strengths	44.4 <sub>49.7</sub>	$1.0 \times 10^{-121}$	0.8	35.6 <sub>47.9</sub>	$5.5 \times 10^{-73}$	0.6	25.9 <sub>43.8</sub>	$2.7 \times 10^{-34}$	0.4	22.6 <sub>41.8</sub>	$4.7 \times 10^{-23}$	0.3	9.3 <sub>32.4</sub>	14.1 <sub>34.8</sub>	4.5 <sub>20.8</sub>
Questions on Experiences	15.2 <sub>35.9</sub>	$2.6 \times 10^{-47}$	-0.5	16.3 <sub>36.9</sub>	$5.1 \times 10^{-40}$	-0.5	27.7 <sub>44.8</sub>	$2.0 \times 10^{-5}$	-0.2	24.3 <sub>42.9</sub>	$8.5 \times 10^{-12}$	-0.3	35.6 <sub>48.6</sub>	40.2 <sub>49.1</sub>	31.0 <sub>46.3</sub>
Questions on Perspectives	2.8 <sub>16.4</sub>	$1.0 \times 10^0$	0.0	0.8 <sub>9.1</sub>	$3.2 \times 10^{-6}$	-0.2	3.9 <sub>19.3</sub>	$1.0 \times 10^0$	0.0	4.8 <sub>21.4</sub>	$1.0 \times 10^0$	0.1	3.5 <sub>20.2</sub>	4.8 <sub>21.5</sub>	2.3 <sub>14.9</sub>
Questions on Emotions	0.7 <sub>8.1</sub>	$2.6 \times 10^{-37}$	-0.4	0.3 <sub>5.3</sub>	$2.9 \times 10^{-40}$	-0.5	3.2 <sub>17.5</sub>	$4.6 \times 10^{-14}$	-0.3	4.2 <sub>20.0</sub>	$1.3 \times 10^{-8}$	-0.2	9.6 <sub>30.7</sub>	11.3 <sub>31.6</sub>	7.9 <sub>27.0</sub>
Problem-Solving	40.7 <sub>49.1</sub>	$8.0 \times 10^{-14}$	0.3	38.1 <sub>48.6</sub>	$1.9 \times 10^{-8}$	0.2	37.3 <sub>48.4</sub>	$4.5 \times 10^{-7}$	0.2	29.4 <sub>45.6</sub>	$1.0 \times 10^0$	0.0	28.4 <sub>42.7</sub>	20.4 <sub>40.3</sub>	36.3 <sub>48.2</sub>
Planning	32.4 <sub>46.8</sub>	$1.2 \times 10^{-39}$	0.4	32.4 <sub>46.8</sub>	$2.4 \times 10^{-39}$	0.5	32.4 <sub>46.8</sub>	$1.4 \times 10^{-39}$	0.5	25.9 <sub>43.8</sub>	$2.6 \times 10^{-18}$	0.3	13.7 <sub>32.6</sub>	10.7 <sub>30.9</sub>	16.6 <sub>37.3</sub>
Normalizing	28.0 <sub>44.9</sub>	$1.1 \times 10^{-23}$	0.3	23.4 <sub>42.3</sub>	$2.7 \times 10^{-11}$	0.2	9.9 <sub>29.9</sub>	$5.6 \times 10^{-3}$	-0.1	13.8 <sub>34.5</sub>	$1.0 \times 10^0$	0.0	14.0 <sub>32.6</sub>	10.5 <sub>30.7</sub>	17.5 <sub>38.0</sub>
Psychoeducation	6.1 <sub>24.0</sub>	$1.0 \times 10^0$	0.0	12.0 <sub>32.5</sub>	$1.7 \times 10^{-10}$	0.2	8.5 <sub>27.8</sub>	$1.9 \times 10^{-2}$	0.1	5.8 <sub>23.4</sub>	$1.0 \times 10^0$	0.0	5.4 <sub>20.5</sub>	3.5 <sub>18.4</sub>	7.3 <sub>26.1</sub>
<b>Client behavior: Sustaining unhealthy behavior</b>															
<i>n</i>	436			419			553			445			389	275	114
Reflections on Needs	36.5 <sub>48.2</sub>	$5.1 \times 10^{-1}$	0.2	35.8 <sub>48.0</sub>	$9.7 \times 10^{-1}$	0.2	31.5 <sub>46.5</sub>	$1.0 \times 10^0$	0.1	30.1 <sub>45.9</sub>	$1.0 \times 10^0$	0.1	27.6 <sub>45.3</sub>	30.5 <sub>46.1</sub>	24.6 <sub>43.2</sub>
Reflections on Emotions	9.2 <sub>28.9</sub>	$1.0 \times 10^0$	-0.1	14.8 <sub>35.5</sub>	$1.0 \times 10^0$	0.1	11.0 <sub>31.4</sub>	$1.0 \times 10^0$	0.0	14.2 <sub>34.9</sub>	$1.0 \times 10^0$	0.1	11.7 <sub>33.2</sub>	13.8 <sub>34.6</sub>	9.6 <sub>29.7</sub>
Reflections on Values	4.1 <sub>19.9</sub>	$1.0 \times 10^0$	0.0	3.6 <sub>18.6</sub>	$1.0 \times 10^0$	-0.1	4.7 <sub>21.2</sub>	$1.0 \times 10^0$	-0.1	3.8 <sub>19.2</sub>	$1.0 \times 10^0$	-0.1	6.8 <sub>25.0</sub>	6.5 <sub>24.8</sub>	7.0 <sub>25.7</sub>
Reflections on Consequences	35.1 <sub>47.8</sub>	$1.0 \times 10^0$	-0.1	27.4 <sub>44.7</sub>	$5.8 \times 10^{-1}$	-0.2	26.6 <sub>44.2</sub>	$1.2 \times 10^{-1}$	-0.2	31.5 <sub>46.5</sub>	$1.0 \times 10^0$	-0.1	36.1 <sub>46.8</sub>	26.5 <sub>44.2</sub>	45.0 <sub>50.0</sub>
Reflections on Conflicts	11.2 <sub>31.6</sub>	$1.0 \times 10^0$	0.1	6.9 <sub>25.4</sub>	$1.0 \times 10^0$	0.0	6.7 <sub>25.0</sub>	$1.0 \times 10^0$	-0.1	8.1 <sub>27.3</sub>	$1.0 \times 10^0$	0.0	8.2 <sub>27.1</sub>	7.6 <sub>26.6</sub>	8.8 <sub>28.4</sub>
Reflections on Strengths	11.7 <sub>32.2</sub>	$8.8 \times 10^{-2}$	0.2	8.1 <sub>27.3</sub>	$1.0 \times 10^0$	0.1	7.6 <sub>26.5</sub>	$1.0 \times 10^0$	0.1	4.7 <sub>21.2</sub>	$1.0 \times 10^0$	0.0	5.2 <sub>23.6</sub>	6.9 <sub>25.4</sub>	3.5 <sub>18.5</sub>
Questions on Experiences	19.5 <sub>39.7</sub>	$2.3 \times 10^{-8}$	-0.4	20.5 <sub>40.4</sub>	$3.5 \times 10^{-7}$	-0.4	29.7 <sub>45.7</sub>	$1.6 \times 10^{-1}$	-0.2	30.6 <sub>46.1</sub>	$6.5 \times 10^{-1}$	-0.2	39.3 <sub>49.4</sub>	46.2 <sub>49.9</sub>	32.5 <sub>47.0</sub>
Questions on Perspectives	7.6 <sub>26.5</sub>	$9.9 \times 10^{-1}$	0.2	2.6 <sub>16.0</sub>	$1.0 \times 10^0$	0.0	6.9 <sub>25.3</sub>	$1.0 \times 10^0$	0.1	6.1 <sub>23.9</sub>	$1.0 \times 10^0$	0.1	3.5 <sub>19.3</sub>	4.4 <sub>20.5</sub>	2.6 <sub>16.1</sub>
Questions on Emotions	0.5 <sub>6.8</sub>	$2.5 \times 10^{-7}$	-0.4	0.5 <sub>6.9</sub>	$5.5 \times 10^{-7}$	-0.4	2.5 <sub>15.7</sub>	$6.7 \times 10^{-4}$	-0.3	5.8 <sub>23.5</sub>	$1.0 \times 10^0$	-0.1	9.2 <sub>29.7</sub>	10.5 <sub>30.8</sub>	7.9 <sub>27.1</sub>
Problem-Solving	31.4 <sub>46.5</sub>	$8.2 \times 10^{-1}$	0.2	27.9 <sub>44.9</sub>	$1.0 \times 10^0$	0.1	34.0 <sub>47.4</sub>	$3.5 \times 10^{-2}$	0.2	21.8 <sub>41.3</sub>	$1.0 \times 10^0$	0.0	23.6 <sub>40.3</sub>	15.6 <sub>36.4</sub>	31.6 <sub>46.7</sub>
Planning	12.8 <sub>33.5</sub>	$4.7 \times 10^{-2}$	0.2	9.3 <sub>29.1</sub>	$1.0 \times 10^0$	0.1	19.9 <sub>40.0</sub>	$5.1 \times 10^{-8}$	0.4	10.1 <sub>30.2</sub>	$1.0 \times 10^0$	0.2	5.9 <sub>22.1</sub>	4.0 <sub>19.6</sub>	7.9 <sub>27.1</sub>
Normalizing	14.7 <sub>35.4</sub>	$1.0 \times 10^0$	0.0	15.5 <sub>36.2</sub>	$1.0 \times 10^0$	0.0	6.9 <sub>25.3</sub>	$1.4 \times 10^{-2}$	-0.2	7.4 <sub>26.2</sub>	$1.1 \times 10^{-1}$	-0.2	14.0 <sub>33.2</sub>	10.5 <sub>30.8</sub>	17.5 <sub>38.2</sub>
Psychoeducation	2.1 <sub>14.2</sub>	$1.3 \times 10^{-5}$	-0.4	11.9 <sub>32.5</sub>	$1.0 \times 10^0$	0.1	5.2 <sub>22.3</sub>	$3.4 \times 10^{-1}$	-0.2	2.9 <sub>16.9</sub>	$5.6 \times 10^{-4}$	-0.3	9.8 <sub>26.7</sub>	4.7 <sub>21.3</sub>	14.9 <sub>35.8</sub>
<b>Client behavior: Sharing positive emotions</b>															
<i>n</i>	658			554			408			534			294	226	68
Reflections on Needs	12.9 <sub>33.6</sub>	$1.0 \times 10^0$	-0.1	17.9 <sub>38.3</sub>	$1.0 \times 10^0$	0.0	20.6 <sub>40.5</sub>	$1.0 \times 10^0$	0.1	16.7 <sub>37.3</sub>	$1.0 \times 10^0$	0.0	17.3 <sub>40.9</sub>	24.3 <sub>43.0</sub>	10.3 <sub>30.6</sub>
Reflections on Emotions	5.3 <sub>22.5</sub>	$1.0 \times 10^0$	-0.1	8.1 <sub>27.3</sub>	$1.0 \times 10^0$	0.0	5.6 <sub>23.1</sub>	$1.0 \times 10^0$	-0.1	5.8 <sub>23.4</sub>	$1.0 \times 10^0$	-0.1	7.5 <sub>32.0</sub>	15.0 <sub>35.8</sub>	0.0 <sub>0.0</sub>
Reflections on Values	2.9 <sub>16.8</sub>	$1.0 \times 10^0$	0.0	4.0 <sub>19.5</sub>	$1.0 \times 10^0$	0.0	5.1 <sub>22.1</sub>	$1.0 \times 10^0$	0.1	6.4 <sub>24.4</sub>	$1.0 \times 10^0$	0.1	3.4 <sub>20.6</sub>	5.3 <sub>22.5</sub>	1.5 <sub>12.1</sub>
Reflections on Consequences	6.2 <sub>24.2</sub>	$6.6 \times 10^{-5}$	-0.3	5.2 <sub>22.3</sub>	$5.1 \times 10^{-6}$	-0.4	14.2 <sub>35.0</sub>	$1.0 \times 10^0$	0.0	7.9 <sub>26.9</sub>	$1.8 \times 10^{-2}$	-0.3	15.8 <sub>34.0</sub>	11.1 <sub>31.4</sub>	20.6 <sub>40.7</sub>
Reflections on Conflicts	1.8 <sub>13.4</sub>	$1.0 \times 10^0$	-0.1	2.0 <sub>14.0</sub>	$1.0 \times 10^0$	-0.1	4.9 <sub>21.6</sub>	$1.0 \times 10^0$	0.1	1.9 <sub>13.6</sub>	$1.0 \times 10^0$	-0.1	3.4 <sub>20.6</sub>	5.3 <sub>22.5</sub>	1.5 <sub>12.1</sub>
Reflections on Strengths	66.1 <sub>47.4</sub>	$3.7 \times 10^{-62}$	1.3	52.3 <sub>50.0</sub>	$2.7 \times 10^{-33}$	0.9	29.2 <sub>45.5</sub>	$1.9 \times 10^{-7}$	0.5	39.3 <sub>48.9</sub>	$5.8 \times 10^{-17}$	0.7	10.2 <sub>34.0</sub>	15.9 <sub>36.7</sub>	4.4 <sub>20.7</sub>
Questions on Experiences	7.9 <sub>27.0</sub>	$1.5 \times 10^{-19}$	-0.7	10.6 <sub>30.9</sub>	$2.2 \times 10^{-12}$	-0.6	24.3 <sub>42.9</sub>	$1.0 \times 10^0$	-0.2	15.7 <sub>36.4</sub>	$8.3 \times 10^{-6}$	-0.4	31.6 <sub>47.6</sub>	36.7 <sub>48.3</sub>	26.5 <sub>44.4</sub>
Questions on Perspectives	2.1 <sub>14.4</sub>	$1.0 \times 10^0$	0.1	0.7 <sub>8.5</sub>	$1.0 \times 10^0$	-0.1	3.9 <sub>19.4</sub>	$1.0 \times 10^0$	0.1	3.4 <sub>18.1</sub>	$1.0 \times 10^0$	0.1	1.3 <sub>14.2</sub>	2.7 <sub>16.1</sub>	0.0 <sub>0.0</sub>
Questions on Emotions	0.5 <sub>6.7</sub>	$3.6 \times 10^{-7}$	-0.4	0.5 <sub>7.3</sub>	$8.4 \times 10^{-6}$	-0.4	4.9 <sub>21.6</sub>	$1.0 \times 10^0$	-0.1	3.6 <sub>18.5</sub>	$1.0 \times 10^0$	-0.2	7.5 <sub>28.9</sub>	10.6 <sub>30.9</sub>	4.4 <sub>20.7</sub>
Problem-Solving	19.8 <sub>39.8</sub>	$9.3 \times 10^{-1}$	0.2	23.1 <sub>42.2</sub>	$3.3 \times 10^{-2}$	0.3	26.0 <sub>43.9</sub>	$2.0 \times 10^{-3}$	0.3	14.2 <sub>35.0</sub>	$1.0 \times 10^0$	0.0	13.0 <sub>33.6</sub>	12.8 <sub>33.5</sub>	13.2 <sub>34.1</sub>
Planning	47.9 <sub>50.0</sub>	$7.0 \times 10^{-33}$	0.9	39.4 <sub>48.9</sub>	$5.1 \times 10^{-21}$	0.7	32.6 <sub>46.9</sub>	$3.4 \times 10^{-13}$	0.6	25.1 <sub>43.4</sub>	$1.5 \times 10^{-7}$	0.4	8.2 <sub>26.9</sub>	7.5 <sub>26.4</sub>	8.8 <sub>28.6</sub>
Normalizing	59.0 <sub>49.2</sub>	$7.1 \times 10^{-25}$	0.8	42.6 <sub>49.5</sub>	$4.0 \times 10^{-7}$	0.4	12.5 <sub>33.1</sub>	$2.0 \times 10^{-2}$	-0.3	29.0 <sub>45.4</sub>	$1.0 \times 10^0$	0.1	22.7 <sub>40.6</sub>	19.0 <sub>39.3</sub>	26.5 <sub>44.4</sub>
Psychoeducation	4.3 <sub>20.2</sub>	$1.0 \times 10^0$	0.1	7.4 <sub>26.2</sub>	$7.5 \times 10^{-1}$	0.2	5.6 <sub>23.1</sub>	$1.0 \times 10^0$	0.1	4.7 <sub>21.1</sub>	$1.0 \times 10^0$	0.1	3.1 <sub>15.3</sub>	1.8 <sub>13.2</sub>	4.4 <sub>20.7</sub>

**Table C.6:** (Continue Table C.5) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the Cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 38.9% and human therapists responds by 24.5% on average when client expresses SHARING NEGATIVE EMOTIONS. ( $P = 4.0 \times 10^{-9}$ , Cohen’s  $d = 0.3$ , two-sided Student’s  $t$ -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
<b>Client behavior: Sharing negative emotions</b>															
<i>n</i>	1063			1016			1236			1114			820	646	174
Reflections on Needs	38.9 <sub>48.8</sub>	$4.0 \times 10^{-9}$	0.3	32.3 <sub>46.8</sub>	$2.5 \times 10^{-2}$	0.2	28.9 <sub>45.3</sub>	$1.0 \times 10^0$	0.1	31.9 <sub>46.6</sub>	$3.9 \times 10^{-2}$	0.2	24.5 <sub>44.5</sub>	28.9 <sub>45.4</sub>	20.1 <sub>40.2</sub>
Reflections on Emotions	27.4 <sub>44.6</sub>	$8.0 \times 10^{-3}$	0.2	33.9 <sub>47.3</sub>	$1.4 \times 10^{-9}$	0.3	25.1 <sub>43.4</sub>	$2.9 \times 10^{-1}$	0.1	24.0 <sub>42.7</sub>	$1.0 \times 10^0$	0.1	19.4 <sub>42.7</sub>	27.4 <sub>44.6</sub>	11.5 <sub>32.0</sub>
Reflections on Values	4.5 <sub>20.8</sub>	$1.0 \times 10^0$	-0.1	4.9 <sub>21.6</sub>	$1.0 \times 10^0$	-0.1	4.0 <sub>19.5</sub>	$1.0 \times 10^0$	-0.1	5.6 <sub>22.9</sub>	$1.0 \times 10^0$	0.0	6.3 <sub>25.4</sub>	7.4 <sub>26.2</sub>	5.2 <sub>22.2</sub>
Reflections on Consequences	21.3 <sub>40.9</sub>	$1.0 \times 10^0$	0.0	14.0 <sub>34.7</sub>	$6.1 \times 10^{-4}$	-0.2	16.7 <sub>37.4</sub>	$2.8 \times 10^{-1}$	-0.1	22.2 <sub>41.6</sub>	$1.0 \times 10^0$	0.0	21.8 <sub>39.8</sub>	18.3 <sub>38.7</sub>	25.3 <sub>43.6</sub>
Reflections on Conflicts	7.9 <sub>27.0</sub>	$1.0 \times 10^0$	0.0	5.3 <sub>22.4</sub>	$8.2 \times 10^{-1}$	-0.1	8.4 <sub>27.8</sub>	$1.0 \times 10^0$	0.0	8.6 <sub>28.1</sub>	$1.0 \times 10^0$	0.0	8.4 <sub>29.5</sub>	10.5 <sub>30.7</sub>	6.3 <sub>24.4</sub>
Reflections on Strengths	23.1 <sub>42.2</sub>	$2.0 \times 10^{-20}$	0.5	17.2 <sub>37.8</sub>	$8.9 \times 10^{-10}$	0.3	17.8 <sub>38.3</sub>	$2.8 \times 10^{-11}$	0.3	11.8 <sub>32.2</sub>	$1.3 \times 10^{-2}$	0.2	6.6 <sub>26.6</sub>	8.5 <sub>27.9</sub>	4.6 <sub>21.0</sub>
Questions on Experiences	22.3 <sub>41.6</sub>	$2.8 \times 10^{-7}$	-0.3	18.4 <sub>38.8</sub>	$2.2 \times 10^{-13}$	-0.4	32.9 <sub>47.0</sub>	$1.0 \times 10^0$	0.0	34.9 <sub>47.7</sub>	$1.0 \times 10^0$	0.0	34.6 <sub>48.6</sub>	40.6 <sub>49.1</sub>	28.7 <sub>45.4</sub>
Questions on Perspectives	3.3 <sub>17.9</sub>	$1.0 \times 10^0$	0.0	0.8 <sub>8.8</sub>	$3.1 \times 10^{-2}$	-0.2	5.6 <sub>23.0</sub>	$1.0 \times 10^0$	0.1	4.9 <sub>21.7</sub>	$1.0 \times 10^0$	0.1	3.3 <sub>19.9</sub>	4.8 <sub>21.4</sub>	1.7 <sub>13.1</sub>
Questions on Emotions	1.6 <sub>12.6</sub>	$2.8 \times 10^{-12}$	-0.4	1.1 <sub>10.4</sub>	$3.6 \times 10^{-14}$	-0.4	3.2 <sub>17.7</sub>	$3.1 \times 10^{-7}$	-0.3	4.5 <sub>20.7</sub>	$1.1 \times 10^{-3}$	-0.2	9.6 <sub>31.1</sub>	11.8 <sub>32.2</sub>	7.5 <sub>26.4</sub>
Problem-Solving	37.1 <sub>48.3</sub>	$4.4 \times 10^{-17}$	0.4	35.5 <sub>47.9</sub>	$2.8 \times 10^{-14}$	0.4	33.2 <sub>47.1</sub>	$2.3 \times 10^{-11}$	0.3	24.4 <sub>43.0</sub>	$2.5 \times 10^{-1}$	0.1	19.0 <sub>35.7</sub>	12.1 <sub>32.6</sub>	25.9 <sub>43.9</sub>
Planning	13.9 <sub>34.6</sub>	$8.8 \times 10^{-4}$	0.2	12.8 <sub>33.4</sub>	$1.9 \times 10^{-2}$	0.2	19.3 <sub>39.4</sub>	$7.0 \times 10^{-12}$	0.3	13.5 <sub>34.2</sub>	$2.7 \times 10^{-3}$	0.2	7.6 <sub>25.0</sub>	6.0 <sub>23.8</sub>	9.2 <sub>29.0</sub>
Normalizing	23.0 <sub>42.1</sub>	$3.4 \times 10^{-2}$	0.2	30.3 <sub>46.0</sub>	$2.8 \times 10^{-10}$	0.3	12.9 <sub>33.6</sub>	$1.0 \times 10^0$	-0.1	14.3 <sub>35.0</sub>	$1.0 \times 10^0$	-0.1	16.5 <sub>36.2</sub>	14.7 <sub>35.4</sub>	18.4 <sub>38.9</sub>
Psychoeducation	4.5 <sub>20.8</sub>	$1.0 \times 10^0$	0.0	10.3 <sub>30.5</sub>	$2.5 \times 10^{-4}$	0.2	6.3 <sub>24.3</sub>	$1.0 \times 10^0$	0.1	3.0 <sub>17.0</sub>	$1.0 \times 10^0$	-0.1	4.7 <sub>17.5</sub>	2.0 <sub>14.1</sub>	7.5 <sub>26.4</sub>
<b>Client behavior: Sharing experiences</b>															
<i>n</i>	865			842			940			879			759	533	226
Reflections on Needs	33.3 <sub>47.2</sub>	$1.4 \times 10^{-4}$	0.2	28.0 <sub>44.9</sub>	$9.6 \times 10^{-1}$	0.1	26.0 <sub>43.9</sub>	$1.0 \times 10^0$	0.1	27.0 <sub>44.4</sub>	$1.0 \times 10^0$	0.1	22.5 <sub>42.8</sub>	26.5 <sub>44.2</sub>	18.6 <sub>39.0</sub>
Reflections on Emotions	17.9 <sub>38.4</sub>	$1.0 \times 10^0$	0.1	23.9 <sub>42.7</sub>	$1.8 \times 10^{-5}$	0.3	20.0 <sub>40.0</sub>	$4.3 \times 10^{-2}$	0.2	19.3 <sub>39.5</sub>	$1.6 \times 10^{-1}$	0.2	13.5 <sub>36.8</sub>	19.9 <sub>40.0</sub>	7.1 <sub>25.7</sub>
Reflections on Values	5.7 <sub>23.1</sub>	$1.0 \times 10^0$	0.0	5.5 <sub>22.7</sub>	$1.0 \times 10^0$	0.0	4.5 <sub>20.7</sub>	$1.0 \times 10^0$	0.0	5.1 <sub>22.1</sub>	$1.0 \times 10^0$	0.0	4.7 <sub>23.1</sub>	7.1 <sub>25.8</sub>	2.2 <sub>14.7</sub>
Reflections on Consequences	19.4 <sub>39.6</sub>	$1.0 \times 10^0$	0.0	10.8 <sub>31.1</sub>	$1.8 \times 10^{-3}$	-0.2	14.1 <sub>34.9</sub>	$1.0 \times 10^0$	-0.1	16.3 <sub>36.9</sub>	$1.0 \times 10^0$	-0.1	18.2 <sub>38.5</sub>	17.8 <sub>38.3</sub>	18.6 <sub>39.0</sub>
Reflections on Conflicts	6.4 <sub>24.4</sub>	$1.0 \times 10^0$	0.0	3.4 <sub>18.2</sub>	$1.0 \times 10^0$	-0.1	7.2 <sub>25.9</sub>	$1.0 \times 10^0$	0.1	5.5 <sub>22.7</sub>	$1.0 \times 10^0$	0.0	5.8 <sub>25.1</sub>	8.1 <sub>27.3</sub>	3.5 <sub>18.5</sub>
Reflections on Strengths	21.5 <sub>41.1</sub>	$5.4 \times 10^{-11}$	0.4	18.3 <sub>38.7</sub>	$1.4 \times 10^{-6}$	0.3	13.5 <sub>34.2</sub>	$1.0 \times 10^{-1}$	0.2	12.5 <sub>33.1</sub>	$7.3 \times 10^{-1}$	0.1	8.4 <sub>29.9</sub>	12.0 <sub>32.5</sub>	9.4 <sub>21.6</sub>
Questions on Experiences	23.5 <sub>42.4</sub>	$3.6 \times 10^{-18}$	-0.5	25.3 <sub>43.5</sub>	$1.3 \times 10^{-14}$	-0.4	35.1 <sub>47.8</sub>	$4.1 \times 10^{-3}$	-0.2	36.4 <sub>48.1</sub>	$4.7 \times 10^{-2}$	-0.2	44.7 <sub>49.8</sub>	45.2 <sub>49.8</sub>	44.2 <sub>49.8</sub>
Questions on Perspectives	4.4 <sub>20.5</sub>	$7.9 \times 10^{-1}$	0.1	1.4 <sub>11.9</sub>	$1.0 \times 10^0$	0.0	5.0 <sub>21.8</sub>	$1.3 \times 10^{-1}$	0.2	4.6 <sub>20.9</sub>	$5.1 \times 10^{-1}$	0.1	2.0 <sub>15.6</sub>	3.2 <sub>17.6</sub>	0.9 <sub>9.4</sub>
Questions on Emotions	2.2 <sub>14.7</sub>	$2.4 \times 10^{-7}$	-0.3	1.8 <sub>13.2</sub>	$1.5 \times 10^{-8}$	-0.3	3.9 <sub>19.5</sub>	$2.6 \times 10^{-3}$	-0.2	5.8 <sub>23.4</sub>	$1.0 \times 10^0$	-0.1	8.8 <sub>28.6</sub>	9.2 <sub>28.9</sub>	8.4 <sub>27.8</sub>
Problem-Solving	29.6 <sub>45.7</sub>	$7.1 \times 10^{-13}$	0.4	23.9 <sub>42.7</sub>	$2.1 \times 10^{-5}$	0.3	21.2 <sub>40.9</sub>	$6.7 \times 10^{-3}$	0.2	15.1 <sub>35.9</sub>	$1.0 \times 10^0$	0.0	14.0 <sub>32.7</sub>	9.4 <sub>29.2</sub>	18.6 <sub>39.0</sub>
Planning	12.9 <sub>33.6</sub>	$5.9 \times 10^{-3}$	0.2	10.5 <sub>30.6</sub>	$1.0 \times 10^0$	0.1	13.1 <sub>33.7</sub>	$3.3 \times 10^{-3}$	0.2	8.3 <sub>27.6</sub>	$1.0 \times 10^0$	0.0	7.0 <sub>25.5</sub>	6.9 <sub>25.4</sub>	7.1 <sub>25.7</sub>
Normalizing	16.2 <sub>36.9</sub>	$1.0 \times 10^0$	0.1	18.2 <sub>38.6</sub>	$1.0 \times 10^0$	0.1	8.8 <sub>28.4</sub>	$5.5 \times 10^{-2}$	-0.2	11.0 <sub>31.4</sub>	$1.0 \times 10^0$	-0.1	14.0 <sub>34.7</sub>	13.9 <sub>34.6</sub>	14.2 <sub>34.9</sub>
Psychoeducation	3.1 <sub>17.4</sub>	$4.5 \times 10^{-1}$	-0.1	5.7 <sub>23.2</sub>	$1.0 \times 10^0$	0.0	2.8 <sub>16.4</sub>	$7.7 \times 10^{-2}$	-0.2	1.1 <sub>10.6</sub>	$7.4 \times 10^{-7}$	-0.3	5.7 <sub>20.7</sub>	2.6 <sub>16.0</sub>	8.8 <sub>28.5</sub>
<b>Client behavior: Gained insights</b>															
<i>n</i>	450			365			430			437			258	214	44
Reflections on Needs	28.0 <sub>44.9</sub>	$1.0 \times 10^0$	0.0	29.3 <sub>45.6</sub>	$1.0 \times 10^0$	0.1	27.9 <sub>44.9</sub>	$1.0 \times 10^0$	0.0	22.9 <sub>42.1</sub>	$1.0 \times 10^0$	-0.1	27.0 <sub>45.1</sub>	29.0 <sub>45.5</sub>	25.0 <sub>43.8</sub>
Reflections on Emotions	9.1 <sub>28.8</sub>	$1.0 \times 10^0$	-0.1	11.8 <sub>32.3</sub>	$1.0 \times 10^0$	0.0	10.5 <sub>30.6</sub>	$1.0 \times 10^0$	-0.1	6.9 <sub>25.3</sub>	$1.0 \times 10^0$	-0.2	12.5 <sub>35.5</sub>	15.9 <sub>36.6</sub>	9.1 <sub>29.1</sub>
Reflections on Values	5.6 <sub>22.9</sub>	$1.0 \times 10^0$	-0.1	8.8 <sub>28.3</sub>	$1.0 \times 10^0$	0.0	6.7 <sub>25.1</sub>	$1.0 \times 10^0$	-0.1	7.8 <sub>26.8</sub>	$1.0 \times 10^0$	0.0	8.1 <sub>28.6</sub>	9.3 <sub>29.2</sub>	6.8 <sub>25.5</sub>
Reflections on Consequences	18.2 <sub>38.6</sub>	$1.0 \times 10^0$	0.0	12.6 <sub>33.2</sub>	$1.0 \times 10^0$	-0.2	15.8 <sub>36.5</sub>	$1.0 \times 10^0$	-0.1	14.9 <sub>35.6</sub>	$1.0 \times 10^0$	-0.1	18.8 <sub>37.0</sub>	15.0 <sub>35.7</sub>	22.7 <sub>42.4</sub>
Reflections on Conflicts	4.2 <sub>20.1</sub>	$1.0 \times 10^0$	0.0	3.8 <sub>19.2</sub>	$1.0 \times 10^0$	0.0	6.3 <sub>24.3</sub>	$1.0 \times 10^0$	0.1	4.8 <sub>21.4</sub>	$1.0 \times 10^0$	0.0	3.9 <sub>21.9</sub>	5.6 <sub>23.1</sub>	2.3 <sub>15.1</sub>
Reflections on Strengths	51.3 <sub>50.0</sub>	$4.0 \times 10^{-24}$	0.9	36.4 <sub>48.2</sub>	$2.1 \times 10^{-9}$	0.6	31.4 <sub>46.5</sub>	$2.1 \times 10^{-6}$	0.4	32.7 <sub>47.0</sub>	$2.8 \times 10^{-7}$	0.5	12.5 <sub>35.5</sub>	15.9 <sub>36.6</sub>	9.1 <sub>29.1</sub>
Questions on Experiences	13.1 <sub>33.8</sub>	$2.3 \times 10^{-5}$	-0.4	16.2 <sub>36.9</sub>	$1.2 \times 10^{-2}$	-0.3	24.7 <sub>43.1</sub>	$1.0 \times 10^0$	-0.1	17.4 <sub>37.9</sub>	$3.0 \times 10^{-2}$	-0.3	29.2 <sub>48.6</sub>	42.5 <sub>49.6</sub>	15.9 <sub>37.0</sub>
Questions on Perspectives	3.1 <sub>17.4</sub>	$1.0 \times 10^0$	-0.1	2.2 <sub>14.7</sub>	$1.0 \times 10^0$	-0.1	4.4 <sub>20.6</sub>	$1.0 \times 10^0$	0.0	4.3 <sub>20.4</sub>	$1.0 \times 10^0$	0.0	4.2 <sub>22.7</sub>	6.1 <sub>23.9</sub>	2.3 <sub>15.1</sub>
Questions on Emotions	0.7 <sub>8.1</sub>	$5.2 \times 10^{-9}$	-0.5	1.9 <sub>13.7</sub>	$9.7 \times 10^{-5}$	-0.4	4.2 <sub>20.1</sub>	$5.0 \times 10^{-2}$	-0.3	4.1 <sub>19.9</sub>	$4.0 \times 10^{-2}$	-0.3	10.8 <sub>30.7</sub>	10.3 <sub>30.4</sub>	11.4 <sub>32.1</sub>
Problem-Solving	41.1 <sub>49.3</sub>	$2.0 \times 10^{-2}$	0.3	37.5 <sub>48.5</sub>	$7.0 \times 10^{-1}$	0.2	37.0 <sub>48.3</sub>	$8.9 \times 10^{-1}$	0.2	25.9 <sub>43.8</sub>	$1.0 \times 10^0$	0.0	27.9 <sub>39.6</sub>	15.0 <sub>35.7</sub>	40.9 <sub>49.7</sub>
Planning	34.9 <sub>47.7</sub>	$4.4 \times 10^{-10}$	0.5	28.2 <sub>45.1</sub>	$3.5 \times 10^{-5}$	0.4	38.4 <sub>48.7</sub>	$9.5 \times 10^{-13}$	0.6	25.9 <sub>43.8</sub>	$4.4 \times 10^{-4}$	0.4	11.5 <sub>32.1</sub>	11.7 <sub>32.2</sub>	11.4 <sub>32.1</sub>
Normalizing	32.4 <sub>46.9</sub>	$3.1 \times 10^{-3}$	0.3	30.1 <sub>45.9</sub>	$5.6 \times 10^{-2}$	0.3	13.7 <sub>34.4</sub>	$1.0 \times 10^0$	-0.1	21.7 <sub>41.3</sub>	$1.0 \times 10^0$	0.1	18.6 <sub>35.1</sub>	12.1 <sub>32.7</sub>	25.0 <sub>43.8</sub>
Psychoeducation	8.7 <sub>28.2</sub>	$1.0 \times 10^0$	0.0	12.6 <sub>33.2</sub>	$1.0 \times 10^0$	0.1	9.5 <sub>29.4</sub>	$1.0 \times 10^0$	0.0	9.2 <sub>28.9</sub>	$1.0 \times 10^0$	0.0	8.4 <sub>18.4</sub>	0.9 <sub>9.6</sub>	15.9 <sub>37.0</sub>

**Table C.7:** Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON CONSEQUENCES by 38.9% which is 15.8% less than average human therapists. (Single: 11.5%, Average human therapists: 27.3% from Table C.5;  $P = 5.3 \times 10^{-17}$ , Cohen’s  $d = -0.4$ , two-sided Student’s  $t$ -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b		
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$
<b>Client behavior: Changing unhealthy behavior</b>																								
$n$	1553			731			1553			595			1553			658			1553			529		
Reflections on Needs	25.2434	$1.0 \times 10^0$	-0.1	30.8462	$1.0 \times 10^0$	0.1	17.0376	$9.7 \times 10^{-6}$	-0.2	21.0408	$2.0 \times 10^{-3}$	-0.2	31.2464	$1.0 \times 10^0$	0.1	28.4451	$1.0 \times 10^0$	0.0	31.1463	$1.0 \times 10^0$	0.1	30.1459	$1.0 \times 10^0$	0.1
Reflections on Emotions	7.5264	$1.0 \times 10^0$	-0.1	7.9270	$1.0 \times 10^0$	-0.1	5.6231	$6.2 \times 10^{-2}$	-0.2	8.3274	$1.0 \times 10^0$	-0.1	8.5279	$1.0 \times 10^0$	-0.1	13.2333	$1.0 \times 10^0$	0.1	10.0300	$1.0 \times 10^0$	0.0	8.8283	$1.0 \times 10^0$	-0.1
Reflections on Values	3.8192	$1.0 \times 10^0$	0.0	1.8135	$3.6 \times 10^{-1}$	-0.1	1.8134	$2.2 \times 10^{-1}$	-0.1	4.3204	$1.0 \times 10^0$	0.0	4.2302	$1.0 \times 10^0$	0.0	6.1240	$1.0 \times 10^0$	0.1	4.3203	$1.0 \times 10^0$	0.0	4.6209	$1.0 \times 10^0$	0.0
Reflections on Consequences	11.5319	$5.3 \times 10^{-17}$	-0.4	9.2290	$2.2 \times 10^{-19}$	-0.5	5.2232	$8.8 \times 10^{-34}$	-0.6	8.1274	$4.7 \times 10^{-20}$	-0.5	23.0217	$1.0 \times 10^0$	-0.1	13.5342	$4.1 \times 10^{-21}$	-0.4	21.4411	$8.0 \times 10^{-3}$	-0.1	23.5324	$1.0 \times 10^0$	-0.1
Reflections on Conflicts	2.3351	$9.8 \times 10^{-4}$	-0.2	1.3115	$5.9 \times 10^{-8}$	-0.2	1.5222	$4.5 \times 10^{-5}$	-0.2	3.2177	$2.1 \times 10^{-1}$	-0.2	5.0217	$1.0 \times 10^0$	-0.1	4.1197	$4.6 \times 10^{-2}$	-0.1	5.6230	$1.0 \times 10^0$	-0.1	7.1258	$1.0 \times 10^0$	0.0
Reflections on Strengths	76.2426	$2.0 \times 10^{-19}$	1.9	62.7484	$5.0 \times 10^{-19}$	1.4	44.497	$2.0 \times 10^{-77}$	0.9	42.3275	$2.9 \times 10^{-62}$	0.9	29.5356	$1.4 \times 10^{-42}$	0.5	25.2434	$2.5 \times 10^{-28}$	0.4	18.1385	$5.6 \times 10^{-2}$	0.2	15.8265	$1.1 \times 10^{-5}$	0.2
Questions on Experiences	16.3369	$2.8 \times 10^{-19}$	-0.4	8.4278	$1.0 \times 10^{-34}$	-0.6	21.3410	$3.2 \times 10^{-9}$	-0.3	27.4466	$5.0 \times 10^{-2}$	-0.2	14.7354	$9.7 \times 10^{-40}$	-0.5	19.2395	$2.1 \times 10^{-22}$	-0.4	30.8462	$1.7 \times 10^{-1}$	-0.1	23.2422	$2.8 \times 10^{-12}$	-0.3
Questions on Perspectives	3.2146	$1.0 \times 10^0$	-0.1	0.0201	$1.5 \times 10^{-3}$	-0.2	4.3202	$1.0 \times 10^0$	0.0	4.3201	$1.0 \times 10^0$	0.0	3.0171	$1.0 \times 10^0$	0.0	1.2107	$3.0 \times 10^{-3}$	-0.1	3.7190	$2.1 \times 10^0$	0.0	5.0217	$1.0 \times 10^0$	0.1
Questions on Emotions	0.3352	$4.7 \times 10^{-14}$	-0.4	0.2411	$9.8 \times 10^{-12}$	-0.4	1.1103	$3.5 \times 10^{-10}$	-0.3	0.461	$7.4 \times 10^{-10}$	-0.3	0.861	$1.1 \times 10^{-24}$	-0.4	0.357	$4.5 \times 10^{-29}$	-0.4	4.1197	$5.1 \times 10^{-7}$	-0.2	5.5228	$1.8 \times 10^{-3}$	-0.2
Problem-Solving	38.3486	$5.8 \times 10^{-5}$	0.2	48.241	$6.9 \times 10^{-18}$	0.4	56.2496	$3.0 \times 10^{-37}$	0.6	41.4803	$5.2 \times 10^{-7}$	0.3	41.9493	$4.0 \times 10^{-14}$	0.3	34.3475	$2.1 \times 10^{-2}$	0.1	29.2455	$1.0 \times 10^0$	0.0	25.4335	$1.0 \times 10^0$	-0.1
Planning	47.6500	$5.8 \times 10^{-76}$	0.9	55.049	$9.3 \times 10^{-99}$	1.1	64.0480	$1.9 \times 10^{-151}$	1.3	48.0500	$4.5 \times 10^{-66}$	0.9	25.2434	$6.7 \times 10^{-15}$	0.3	23.8426	$7.8 \times 10^{-12}$	0.3	19.1393	$2.4 \times 10^{-3}$	0.1	18.4387	$2.0 \times 10^{-2}$	0.1
Normalizing	52.7500	$2.6 \times 10^{-96}$	1.0	52.8500	$1.2 \times 10^{-87}$	1.0	18.5389	$3.7 \times 10^{-1}$	0.1	21.7413	$9.1 \times 10^{-4}$	0.2	16.4370	$1.0 \times 10^0$	0.1	12.1324	$1.0 \times 10^0$	-0.1	6.2442	$5.9 \times 10^{-12}$	-0.3	11.3151	$1.0 \times 10^0$	-0.1
Psychoeducation	9.2289	$2.9 \times 10^{-2}$	0.2	8.9285	$1.3 \times 10^{-1}$	0.2	21.4411	$2.7 \times 10^{-31}$	0.6	14.0347	$7.7 \times 10^{-10}$	0.3	4.7212	$1.0 \times 10^0$	0.0	13.2339	$9.3 \times 10^{-13}$	0.3	3.0170	$2.2 \times 10^{-2}$	-0.1	3.0171	$3.3 \times 10^{-2}$	-0.1
<b>Client behavior: Sustaining unhealthy behavior</b>																								
$n$	389			47			389			30			389			164			389			56		
Reflections on Needs	59.6496	$6.0 \times 10^{-4}$	0.7	43.3304	$1.0 \times 10^0$	0.3	34.1476	$1.0 \times 10^0$	0.1	41.1496	$1.0 \times 10^0$	0.3	33.7473	$1.0 \times 10^0$	0.1	35.2478	$1.0 \times 10^0$	0.2	30.3460	$1.0 \times 10^0$	0.1	28.5432	$1.0 \times 10^0$	0.0
Reflections on Emotions	17.0380	$1.0 \times 10^0$	0.2	13.3346	$1.0 \times 10^0$	0.0	9.8298	$1.0 \times 10^0$	-0.1	25.0437	$6.0 \times 10^{-1}$	0.4	8.2275	$1.0 \times 10^0$	-0.1	14.9357	$1.0 \times 10^0$	0.1	11.6320	$1.0 \times 10^0$	0.0	12.6332	$1.0 \times 10^0$	0.0
Reflections on Values	4.3204	$1.0 \times 10^0$	-0.1	3.3183	$1.0 \times 10^0$	-0.1	5.5228	$1.0 \times 10^0$	-0.1	0.010	$1.0 \times 10^0$	-0.3	4.1199	$1.0 \times 10^0$	-0.1	3.6187	$1.0 \times 10^0$	-0.1	4.4205	$1.0 \times 10^0$	-0.1	4.4205	$1.0 \times 10^0$	-0.1
Reflections on Consequences	34.0479	$1.0 \times 10^0$	0.0	43.3304	$1.0 \times 10^0$	0.2	23.2423	$1.9 \times 10^{-1}$	-0.3	32.1471	$1.0 \times 10^0$	-0.1	35.2478	$1.0 \times 10^0$	0.0	26.2440	$2.0 \times 10^{-1}$	-0.2	28.0450	$1.0 \times 10^0$	-0.2	31.4465	$1.0 \times 10^0$	-0.1
Reflections on Conflicts	21.3414	$2.8 \times 10^{-1}$	0.5	10.0305	$1.0 \times 10^0$	0.1	7.3261	$1.0 \times 10^0$	0.0	14.3353	$1.0 \times 10^0$	0.2	10.0301	$1.0 \times 10^0$	0.1	6.7250	$1.0 \times 10^0$	-0.1	6.4246	$1.0 \times 10^0$	-0.1	7.2239	$1.0 \times 10^0$	0.0
Reflections on Strengths	14.9360	$1.0 \times 10^0$	0.4	13.3346	$1.0 \times 10^0$	0.3	13.4342	$9.8 \times 10^{-2}$	0.3	5.4227	$1.0 \times 10^0$	0.0	11.3317	$1.9 \times 10^{-1}$	0.2	7.7267	$1.0 \times 10^0$	-0.1	5.1221	$1.0 \times 10^0$	0.0	4.6210	$1.0 \times 10^0$	0.1
Questions on Experiences	34.0479	$1.0 \times 10^0$	-0.1	23.2423	$1.0 \times 10^0$	-0.3	29.9459	$1.0 \times 10^0$	-0.2	50.9505	$1.0 \times 10^0$	0.2	17.9382	$1.5 \times 10^{-9}$	-0.5	20.3403	$4.8 \times 10^{-7}$	-0.4	29.6457	$3.4 \times 10^{-1}$	-0.2	27.8448	$5.2 \times 10^{-2}$	-0.2
Questions on Perspectives	14.9360	$5.6 \times 10^{-2}$	0.5	6.7254	$1.0 \times 10^0$	0.2	10.4306	$1.2 \times 10^{-1}$	0.3	12.7334	$2.8 \times 10^{-1}$	0.4	6.7250	$1.0 \times 10^0$	0.1	2.3151	$1.0 \times 10^0$	-0.1	5.4226	$1.0 \times 10^0$	-0.1	5.1221	$1.0 \times 10^0$	0.1
Questions on Emotions	0.010	$1.0 \times 10^0$	-0.3	0.010	$1.0 \times 10^0$	-0.3	1.2110	$6.7 \times 10^{-2}$	-0.3	0.010	$1.0 \times 10^0$	-0.3	0.522	$2.1 \times 10^{-6}$	-0.4	3.1173	$3.6 \times 10^{-2}$	-0.3	6.7250	$1.0 \times 10^0$	-0.3	6.7250	$1.0 \times 10^0$	-0.1
Problem-Solving	40.4496	$6.8 \times 10^{-1}$	0.4	50.9309	$6.0 \times 10^{-2}$	0.6	58.5494	$3.3 \times 10^{-18}$	0.8	46.4033	$1.1 \times 10^{-2}$	0.5	30.3460	$1.0 \times 10^0$	0.2	26.2440	$1.0 \times 10^0$	0.1	23.7425	$1.0 \times 10^0$	0.0	18.3387	$1.0 \times 10^0$	-0.1
Planning	14.9360	$1.0 \times 10^0$	0.4	26.7450	$7.5 \times 10^{-4}$	0.8	40.9493	$1.1 \times 10^{-25}$	1.1	21.4414	$1.8 \times 10^{-3}$	0.6	12.6332	$8.3 \times 10^{-2}$	0.2	8.0271	$1.0 \times 10^0$	0.1	11.3114	$6.9 \times 10^{-1}$	0.2	8.5279	$1.0 \times 10^0$	0.1
Normalizing	31.9471	$7.9 \times 10^{-2}$	0.5	46.7507	$8.1 \times 10^{-5}$	0.9	11.0314	$1.0 \times 10^0$	-0.1	5.4227	$1.0 \times 10^0$	-0.3	12.6332	$1.0 \times 10^0$	0.0	13.1338	$1.0 \times 10^0$	0.0	5.1221	$9.6 \times 10^{-4}$	-0.3	7.7267	$2.7 \times 10^{-1}$	-0.2
Psychoeducation	4.3204	$1.0 \times 10^0$	-0.2	3.3183	$1.0 \times 10^0$	-0.2	11.6321	$1.0 \times 10^0$	0.0	10.7312	$1.0 \times 10^0$	0.0	1.8133	$1.2 \times 10^{-5}$	-0.4	12.6332	$1.0 \times 10^0$	0.1	2.6158	$3.8 \times 10^{-4}$	-0.3	1.8133	$1.2 \times 10^{-5}$	-0.4
<b>Client behavior: Sharing positive emotions</b>																								
$n$	294			364			294			260			294			114			294			240		
Reflections on Needs	4.4205	$1.5 \times 10^{-5}$	-0.4	17.3739	$1.0 \times 10^0$	0.0	11.4319	$1.0 \times 10^0$	-0.2	5.8235	$9.8 \times 10^{-3}$	-0.3	23.5225	$1.0 \times 10^0$	0.1	18.4388	$1.0 \times 10^0$	0.0	24.1429	$1.0 \times 10^0$	0.2	25.5337	$1.0 \times 10^0$	0.2
Reflections on Emotions	3.6186	$1.0 \times 10^0$	-0.2	3.1173	$1.0 \times 10^0$	-0.2	1.8132	$1.0 \times 10^0$	-0.2	2.5156	$1.0 \times 10^0$	-0.2	7.5264	$1.0 \times 10^0$	0.0	12.6332	$1.0 \times 10^0$	0.2	7.1258	$1.0 \times 10^0$	0.0	8.5279	$1.0 \times 10^0$	0.0
Reflections on Values	0.891	$1.0 \times 10^0$	-0.2	1.9138	$1.0 \times 10^0$	-0.1	2.6161	$1.0 \times 10^0$	0.0	3.8190	$1.0 \times 10^0$	0.0	5.4227	$1.0 \times 10^0$	0.1	5.8234	$1.0 \times 10^0$	0.1	6.1240	$1.0 \times 10^0$	0.1	8.5279	$9.2 \times 10^{-1}$	0.2
Reflections on Consequences	0.891	$2.5 \times 10^{-13}$	-0.6	1.5123	$2.4 \times 10^{-8}$	-0.5	7.0257	$9.9 \times 10^{-1}$	-0.3	0.891	$6.0 \times 10^{-9}$	-0.6	12.9336	$1.0 \times 10^0$	-0.1	8.5279	$3.5 \times 10^{-1}$	-0.2	17.0376	$1.0 \times 10^0$	0.0	13.6343	$1.0 \times 10^0$	-0.1
Reflections on Conflicts	0.010	$1.4 \times 10^{-1}$	-0.2	0.010	$6.4 \times 10^{-1}$	-0.2	1.8132	$1.0 \times 10^0$	-0.1	0.010	$8.6 \times 10^{-1}$	-0.2	4.1198	$1.0 \times 10^0$	0.0	3.7190	$1.0 \times 10^0$	0.0	6.1240	$1.0 \times 10^0$	0.0	3.4182	$1.0 \times 10^0$	0.0
Reflections on Strengths	97.8427	$1.0 \times 10^{-197}$	3.5	83.8369	$3.2 \times 10^{-88}$	2.1	58.8494	$4.0 \times 10^{-24}$	1.2	67.4211	$2.4 \times 10^{-46}$	1.4	26.9444	$3.2 \times 10^{-5}$	0.4	24.5331	$7.2 \times 10^{-4}$	0.4	17.7382					



**Table C.9:** Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 14$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits POSITIVE EMOTIONS by 1.5% while human therapists responds by 1.0% on average ( $P = 8.4 \times 10^{-13}$ , Cohen’s  $d = 0.1$ , two-sided Student’s  $t$ -test).

Linguistic Attributes	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ( $n = 4893$ )			GPT-3.5 ( $n = 4746$ )			Llama2-70b ( $n = 4970$ )			Llama2-13b ( $n = 4820$ )			Average ( $n = 5446$ )	High-quality ( $n = 3907$ )	Low-quality ( $n = 1539$ )
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
Positive Emotions	1.5 <sub>2,1</sub>	$8.4 \times 10^{-13}$	0.1	1.2 <sub>1,8</sub>	$9.2 \times 10^{-5}$	0.1	1.1 <sub>3,1</sub>	$1.0 \times 10^0$	0.0	1.1 <sub>3,3</sub>	$2.9 \times 10^{-1}$	0.0	1.0 <sub>4,2</sub>	1.2 <sub>4,3</sub>	0.7 <sub>3,6</sub>
Negative Emotions	0.8 <sub>1,5</sub>	$1.0 \times 10^0$	0.0	0.9 <sub>1,6</sub>	$9.3 \times 10^{-6}$	0.1	0.8 <sub>1,9</sub>	$6.4 \times 10^{-2}$	0.1	0.6 <sub>1,8</sub>	$1.0 \times 10^0$	0.0	0.7 <sub>3,6</sub>	0.7 <sub>3,4</sub>	0.6 <sub>4,3</sub>
BigWords	24.5 <sub>7,1</sub>	$0.0 \times 10^0$	1.6	23.1 <sub>7,7</sub>	$0.0 \times 10^0$	1.4	23.7 <sub>11,0</sub>	$0.0 \times 10^0$	1.2	19.8 <sub>12,6</sub>	$0.0 \times 10^0$	0.8	10.7 <sub>9,9</sub>	11.0 <sub>9,9</sub>	10.4 <sub>10,1</sub>
Self	1.5 <sub>2,3</sub>	$1.1 \times 10^{-84}$	-0.4	1.8 <sub>2,6</sub>	$4.3 \times 10^{-68}$	-0.3	1.3 <sub>2,4</sub>	$7.9 \times 10^{-134}$	-0.5	1.9 <sub>3,0</sub>	$9.3 \times 10^{-61}$	-0.3	3.1 <sub>5,1</sub>	2.5 <sub>4,5</sub>	3.8 <sub>6,1</sub>
Other	8.7 <sub>3,3</sub>	$8.2 \times 10^{-1}$	0.0	8.2 <sub>3,3</sub>	$8.7 \times 10^{-12}$	-0.1	7.5 <sub>4,8</sub>	$2.4 \times 10^{-38}$	-0.2	6.7 <sub>4,9</sub>	$2.0 \times 10^{-85}$	-0.4	8.9 <sub>7,1</sub>	8.7 <sub>6,9</sub>	9.2 <sub>7,4</sub>
Health	2.7 <sub>2,8</sub>	$1.1 \times 10^{-30}$	0.2	3.0 <sub>3,3</sub>	$1.6 \times 10^{-60}$	0.3	2.0 <sub>2,9</sub>	$4.8 \times 10^{-2}$	0.1	1.7 <sub>3,9</sub>	$6.1 \times 10^{-1}$	0.0	1.8 <sub>4,3</sub>	1.2 <sub>3,2</sub>	2.5 <sub>6,1</sub>
Wellness	0.9 <sub>1,6</sub>	$8.8 \times 10^{-62}$	0.3	0.8 <sub>1,6</sub>	$2.9 \times 10^{-51}$	0.3	0.6 <sub>1,3</sub>	$1.1 \times 10^{-14}$	0.1	0.5 <sub>1,9</sub>	$1.2 \times 10^{-3}$	0.1	0.3 <sub>2,1</sub>	0.2 <sub>1,3</sub>	0.4 <sub>3,5</sub>
Behavioral Activation	6.8 <sub>4,0</sub>	$7.8 \times 10^{-260}$	0.7	5.9 <sub>4,3</sub>	$4.5 \times 10^{-175}$	0.5	5.7 <sub>5,2</sub>	$7.6 \times 10^{-129}$	0.5	5.1 <sub>5,7</sub>	$4.6 \times 10^{-69}$	0.3	3.2 <sub>5,9</sub>	3.5 <sub>5,9</sub>	2.9 <sub>5,9</sub>
Meaning	3.5 <sub>2,3</sub>	$1.4 \times 10^{-253}$	0.7	3.1 <sub>2,2</sub>	$1.1 \times 10^{-209}$	0.6	2.1 <sub>2,4</sub>	$4.7 \times 10^{-38}$	0.2	1.9 <sub>2,7</sub>	$6.5 \times 10^{-15}$	0.2	1.4 <sub>3,5</sub>	1.4 <sub>3,4</sub>	1.4 <sub>3,7</sub>
Purpose	3.8 <sub>2,6</sub>	$1.3 \times 10^{-285}$	0.7	3.5 <sub>2,5</sub>	$1.8 \times 10^{-257}$	0.7	2.5 <sub>2,6</sub>	$3.5 \times 10^{-63}$	0.3	2.2 <sub>3,0</sub>	$2.4 \times 10^{-27}$	0.2	1.5 <sub>3,6</sub>	1.5 <sub>3,5</sub>	1.5 <sub>3,8</sub>
Motivation	3.4 <sub>2,4</sub>	$6.8 \times 10^{-264}$	0.7	3.1 <sub>2,3</sub>	$9.2 \times 10^{-248}$	0.6	2.1 <sub>2,3</sub>	$7.1 \times 10^{-56}$	0.3	1.9 <sub>2,8</sub>	$4.8 \times 10^{-26}$	0.2	1.3 <sub>3,3</sub>	1.3 <sub>3,3</sub>	1.3 <sub>3,4</sub>
Sadness	0.7 <sub>1,3</sub>	$1.4 \times 10^{-37}$	0.3	0.8 <sub>1,4</sub>	$3.5 \times 10^{-67}$	0.3	0.6 <sub>1,5</sub>	$1.5 \times 10^{-33}$	0.2	0.5 <sub>1,6</sub>	$9.3 \times 10^{-13}$	0.1	0.3 <sub>1,5</sub>	0.3 <sub>1,5</sub>	0.2 <sub>1,6</sub>
Sympathy	0.3 <sub>0,7</sub>	$4.1 \times 10^{-11}$	0.1	0.4 <sub>0,9</sub>	$8.5 \times 10^{-41}$	0.3	0.3 <sub>1,1</sub>	$6.7 \times 10^{-18}$	0.2	0.3 <sub>1,1</sub>	$1.4 \times 10^{-9}$	0.1	0.1 <sub>1,0</sub>	0.2 <sub>1,1</sub>	0.1 <sub>0,8</sub>
Stress	1.1 <sub>1,8</sub>	$1.0 \times 10^0$	0.0	1.4 <sub>1,9</sub>	$2.7 \times 10^{-7}$	0.1	1.4 <sub>2,2</sub>	$3.6 \times 10^{-8}$	0.1	1.1 <sub>2,2</sub>	$1.0 \times 10^0$	0.0	1.1 <sub>3,5</sub>	1.1 <sub>3,2</sub>	1.1 <sub>4,2</sub>

**Table C.10:** Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in High-low quality dataset (Pérez-Rosas et al., 2019) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with POSITIVE EMOTIONS by 2.0% and GPT-4 with full conversation simulation responds by 1.3% when compared with the human therapists responds by 1.0% on average from Table C.9. (Single:  $P = 2.0 \times 10^{-14}$ , Cohen’s  $d = 0.3$ ; two-sided student t-test) (Full:  $P = 3.7 \times 10^{-5}$ , Cohen’s  $d = 0.1$ ; two-sided student t-test).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4 ( $n = 3837$ )			GPT-3.5 ( $n = 3907$ )			Llama2-70b ( $n = 3906$ )			Llama2-13b ( $n = 3895$ )			GPT-4 ( $n = 1056$ )			GPT-3.5 ( $n = 839$ )			Llama2-70b ( $n = 1064$ )			Llama2-13b ( $n = 925$ )		
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$
Positive Emotions	2.0 <sub>1,7</sub>	$2.0 \times 10^{-14}$	0.3	1.7 <sub>1,7</sub>	$1.5 \times 10^{-5}$	0.2	1.3 <sub>1,4</sub>	$5.2 \times 10^{-2}$	0.1	1.7 <sub>2,1</sub>	$3.2 \times 10^{-6}$	0.2	1.3 <sub>2,2</sub>	$3.7 \times 10^{-5}$	0.1	1.2 <sub>1,8</sub>	$2.5 \times 10^{-2}$	0.1	1.0 <sub>3,3</sub>	$1.0 \times 10^0$	0.0	1.0 <sub>3,5</sub>	$1.0 \times 10^0$	0.0
Negative Emotions	0.5 <sub>1,0</sub>	$1.0 \times 10^0$	-0.1	0.6 <sub>1,1</sub>	$1.0 \times 10^0$	0.0	0.7 <sub>1,1</sub>	$1.0 \times 10^0$	0.0	0.6 <sub>1,1</sub>	$1.0 \times 10^0$	0.0	0.8 <sub>1,6</sub>	$2.7 \times 10^{-1}$	0.0	1.0 <sub>1,7</sub>	$1.7 \times 10^{-7}$	0.1	0.9 <sub>2,0</sub>	$2.5 \times 10^{-2}$	0.1	0.6 <sub>1,9</sub>	$1.0 \times 10^0$	0.0
BigWords	28.2 <sub>4,8</sub>	$0.0 \times 10^0$	1.9	25.1 <sub>6,8</sub>	$0.0 \times 10^0$	1.5	32.5 <sub>11,1</sub>	$0.0 \times 10^0$	2.2	32.0 <sub>16,0</sub>	$0.0 \times 10^0$	1.9	23.5 <sub>7,3</sub>	$0.0 \times 10^0$	1.4	22.8 <sub>7,7</sub>	$0.0 \times 10^0$	1.4	22.0 <sub>10,2</sub>	$0.0 \times 10^0$	1.1	17.8 <sub>10,6</sub>	$3.1 \times 10^{-268}$	0.7
Self	2.2 <sub>1,9</sub>	$3.1 \times 10^{-8}$	-0.2	2.4 <sub>2,3</sub>	$1.1 \times 10^{-3}$	-0.1	1.0 <sub>1,6</sub>	$7.0 \times 10^{-39}$	-0.4	1.3 <sub>1,8</sub>	$8.9 \times 10^{-25}$	-0.4	1.4 <sub>2,4</sub>	$7.5 \times 10^{-84}$	-0.4	1.7 <sub>2,6</sub>	$2.1 \times 10^{-69}$	-0.3	1.4 <sub>2,5</sub>	$2.1 \times 10^{-106}$	-0.4	1.9 <sub>3,2</sub>	$4.8 \times 10^{-46}$	-0.3
Other	9.6 <sub>2,3</sub>	$2.9 \times 10^{-2}$	0.1	8.4 <sub>3,0</sub>	$3.4 \times 10^{-1}$	-0.1	5.8 <sub>3,4</sub>	$2.1 \times 10^{-43}$	-0.5	5.6 <sub>4,8</sub>	$1.3 \times 10^{-41}$	-0.5	8.5 <sub>3,5</sub>	$3.2 \times 10^{-3}$	-0.1	8.2 <sub>3,4</sub>	$3.3 \times 10^{-11}$	-0.1	7.8 <sub>8,0</sub>	$3.5 \times 10^{-20}$	-0.2	6.9 <sub>4,9</sub>	$1.0 \times 10^{-64}$	-0.3
Health	2.8 <sub>2,3</sub>	$3.8 \times 10^{-12}$	0.2	3.2 <sub>3,1</sub>	$4.0 \times 10^{-18}$	0.3	1.9 <sub>2,0</sub>	$1.0 \times 10^0$	0.0	1.6 <sub>2,3</sub>	$1.0 \times 10^0$	-0.1	2.7 <sub>2,9</sub>	$6.7 \times 10^{-23}$	0.2	3.0 <sub>3,4</sub>	$3.3 \times 10^{-52}$	0.3	2.1 <sub>3,0</sub>	$2.3 \times 10^{-2}$	0.1	1.7 <sub>4,1</sub>	$1.0 \times 10^0$	0.0
Wellness	1.1 <sub>1,5</sub>	$1.7 \times 10^{-31}$	0.4	1.2 <sub>1,8</sub>	$1.4 \times 10^{-31}$	0.4	0.7 <sub>1,2</sub>	$5.6 \times 10^{-8}$	0.2	0.8 <sub>1,5</sub>	$1.6 \times 10^{-8}$	0.2	0.9 <sub>1,6</sub>	$2.0 \times 10^{-43}$	0.3	0.8 <sub>1,6</sub>	$1.3 \times 10^{-37}$	0.3	0.5 <sub>1,4</sub>	$2.2 \times 10^{-10}$	0.1	0.4 <sub>2,0</sub>	$2.0 \times 10^{-1}$	0.0
Behavioral Activation	7.8 <sub>3,0</sub>	$4.0 \times 10^{-128}$	0.8	8.9 <sub>4,5</sub>	$4.0 \times 10^{-147}$	1.0	6.9 <sub>3,8</sub>	$1.1 \times 10^{-84}$	0.7	6.4 <sub>4,2</sub>	$1.4 \times 10^{-84}$	0.6	6.5 <sub>4,2</sub>	$2.6 \times 10^{-182}$	0.6	5.4 <sub>4,1</sub>	$3.5 \times 10^{-116}$	0.4	5.4 <sub>5,5</sub>	$6.3 \times 10^{-93}$	0.4	4.9 <sub>5,9</sub>	$6.4 \times 10^{-49}$	0.3
Meaning	3.6 <sub>1,8</sub>	$5.0 \times 10^{-84}$	0.7	3.4 <sub>2,1</sub>	$7.6 \times 10^{-60}$	0.6	2.3 <sub>1,7</sub>	$1.6 \times 10^{-16}$	0.3	2.3 <sub>1,8</sub>	$1.9 \times 10^{-12}$	0.3	3.4 <sub>2,4</sub>	$1.3 \times 10^{-201}$	0.7	3.0 <sub>2,1</sub>	$6.8 \times 10^{-177}$	0.6	2.1 <sub>2,5</sub>	$5.9 \times 10^{-29}$	0.2	1.8 <sub>2,8</sub>	$1.0 \times 10^{-9}$	0.1
Purpose	4.3 <sub>1,9</sub>	$1.1 \times 10^{-124}$	0.8	4.7 <sub>2,7</sub>	$9.1 \times 10^{-126}$	0.9	2.8 <sub>2,0</sub>	$1.3 \times 10^{-29}$	0.4	2.7 <sub>2,0</sub>	$1.3 \times 10^{-20}$	0.3	3.7 <sub>2,7</sub>	$2.3 \times 10^{-211}$	0.7	3.3 <sub>2,4</sub>	$2.4 \times 10^{-199}$	0.6	2.4 <sub>2,7</sub>	$8.3 \times 10^{-48}$	0.3	2.1 <sub>3,1</sub>	$9.6 \times 10^{-19}$	0.2
Motivation	3.9 <sub>1,8</sub>	$6.7 \times 10^{-124}$	0.8	4.3 <sub>2,6</sub>	$5.4 \times 10^{-128}$	0.9	2.4 <sub>1,8</sub>	$1.7 \times 10^{-23}$	0.3	2.3 <sub>1,9</sub>	$9.7 \times 10^{-17}$	0.3	3.2 <sub>2,5</sub>	$1.8 \times 10^{-190}$	0.6	2.9 <sub>2,2</sub>	$1.8 \times 10^{-190}$	0.6	2.1 <sub>2,4</sub>	$2.9 \times 10^{-43}$	0.3	1.8 <sub>2,9</sub>	$9.6 \times 10^{-19}$	0.2
Sadness	0.5 <sub>0,9</sub>	$6.2 \times 10^{-5}$	0.2	0.5 <sub>0,9</sub>	$2.6 \times 10^{-2}$	0.1	0.6 <sub>1,0</sub>	$7.5 \times 10^{-10}$	0.2	0.7 <sub>1,4</sub>	$5.8 \times 10^{-12}$	0.3	0.7 <sub>1,4</sub>	$8.1 \times 10^{-38}$	0.3	0.8 <sub>1,5</sub>	$3.0 \times 10^{-72}$	0.3	0.6 <sub>1,6</sub>	$6.7 \times 10^{-30}$	0.2	0.5 <sub>1,7</sub>	$7.1 \times 10^{-9}$	0.1
Sympathy	0.2 <sub>0,5</sub>	$2.9 \times 10^{-1}$	0.1	0.2 <sub>0,6</sub>	$5.9 \times 10^{-1}$	0.1	0.3 <sub>0,7</sub>	$1.5 \times 10^{-6}$	0.2	0.4 <sub>1,3</sub>	$7.2 \times 10^{-12}$	0.3	0.3 <sub>0,8</sub>	$4.4 \times 10^{-11}$	0.1	0.4 <sub>1,0</sub>	$9.8 \times 10^{-45}$	0.3	0.3 <sub>1,2</sub>	$8.8 \times 10^{-16}$	0.2	0.2 <sub>1,0</sub>	$4.2 \times 10^{-6}$	0.1
Stress	0.7 <sub>1,2</sub>	$1.2 \times 10^{-2}$	-0.1	0.9 <sub>1,4</sub>	$6.5 \times 10^{-1}$	-0.1	1.2 <sub>1,5</sub>	$1.0 \times 10^0$	0.0	1.0 <sub>1,5</sub>	$1.0 \times 10^0$	0.0	1.2 <sub>1,9</sub>	$2.3 \times 10^{-1}$	0.1	1.5 <sub>1,9</sub>	$1.5 \times 10^{-10}$	0.1	1.4 <sub>2,3</sub>	$1.9 \times 10^{-8}$	0.1	1.3 <sub>2,3</sub>	$1.0 \times 10^0$	0.0

**Table C.11:** Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS by 40.2% while human therapists responds by 23.4% on average ( $P = 6.9 \times 10^{-90}$ , Cohen’s  $d = 0.4$ , two-sided Student’s  $t$ -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ( $n = 7705$ )			GPT-3.5 ( $n = 7135$ )			Llama2-70b ( $n = 7461$ )			Llama2-13b ( $n = 7464$ )			Average ( $n = 5446$ )	High-quality ( $n = 3907$ )	Low-quality ( $n = 1539$ )
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
Reflections on Needs	40.2 <sub>49.0</sub>	$6.9 \times 10^{-90}$	0.4	35.0 <sub>47.7</sub>	$7.2 \times 10^{-44}$	0.3	29.0 <sub>45.4</sub>	$2.6 \times 10^{-11}$	0.1	28.8 <sub>45.3</sub>	$9.1 \times 10^{-11}$	0.1	23.4 <sub>43.1</sub>	26.2 <sub>44.0</sub>	20.6 <sub>40.5</sub>
Reflections on Emotions	23.1 <sub>42.2</sub>	$1.2 \times 10^{-74}$	0.3	25.2 <sub>43.4</sub>	$8.4 \times 10^{-95}$	0.4	17.1 <sub>37.6</sub>	$3.1 \times 10^{-24}$	0.2	21.3 <sub>41.0</sub>	$4.5 \times 10^{-57}$	0.3	10.5 <sub>32.8</sub>	14.6 <sub>35.3</sub>	6.3 <sub>24.3</sub>
Reflections on Values	4.4 <sub>20.6</sub>	$1.0 \times 10^0$	0.0	3.5 <sub>18.4</sub>	$3.9 \times 10^{-3}$	-0.1	3.7 <sub>19.0</sub>	$4.0 \times 10^{-2}$	-0.1	4.1 <sub>19.8</sub>	$6.2 \times 10^{-1}$	0.0	4.8 <sub>22.8</sub>	6.3 <sub>24.4</sub>	3.3 <sub>17.9</sub>
Reflections on Consequences	22.3 <sub>41.7</sub>	$1.0 \times 10^0$	0.0	21.4 <sub>41.0</sub>	$1.0 \times 10^0$	0.0	19.6 <sub>39.7</sub>	$3.2 \times 10^{-2}$	-0.1	21.2 <sub>40.9</sub>	$1.0 \times 10^0$	0.0	21.8 <sub>40.1</sub>	17.9 <sub>38.4</sub>	25.7 <sub>43.7</sub>
Reflections on Conflicts	4.1 <sub>19.8</sub>	$2.9 \times 10^{-9}$	-0.1	3.5 <sub>18.5</sub>	$9.1 \times 10^{-14}$	-0.1	6.3 <sub>24.2</sub>	$1.0 \times 10^0$	0.0	6.6 <sub>24.8</sub>	$1.0 \times 10^0$	0.0	6.6 <sub>26.1</sub>	8.3 <sub>27.6</sub>	5.0 <sub>21.8</sub>
Reflections on Strengths	33.5 <sub>47.2</sub>	$4.2 \times 10^{-225}$	0.6	27.4 <sub>44.6</sub>	$1.4 \times 10^{-135}$	0.5	16.8 <sub>37.4</sub>	$1.9 \times 10^{-30}$	0.2	15.6 <sub>36.3</sub>	$5.8 \times 10^{-22}$	0.2	9.5 <sub>31.3</sub>	12.8 <sub>33.5</sub>	6.2 <sub>24.2</sub>
Questions on Experiences	36.7 <sub>48.2</sub>	$6.0 \times 10^{-16}$	-0.1	29.5 <sub>45.6</sub>	$6.2 \times 10^{-63}$	-0.3	57.6 <sub>49.4</sub>	$6.7 \times 10^{-52}$	0.3	56.6 <sub>49.6</sub>	$1.5 \times 10^{-44}$	0.3	44.0 <sub>49.9</sub>	48.9 <sub>50.0</sub>	39.1 <sub>48.8</sub>
Questions on Perspectives	7.6 <sub>26.6</sub>	$2.3 \times 10^{-21}$	0.2	3.8 <sub>19.1</sub>	$1.0 \times 10^0$	0.0	12.1 <sub>32.6</sub>	$1.4 \times 10^{-65}$	0.3	11.3 <sub>31.7</sub>	$2.2 \times 10^{-56}$	0.3	3.5 <sub>19.9</sub>	4.9 <sub>21.7</sub>	2.0 <sub>14.1</sub>
Questions on Emotions	3.0 <sub>17.1</sub>	$4.7 \times 10^{-84}$	-0.3	6.1 <sub>24.0</sub>	$3.7 \times 10^{-27}$	-0.2	12.5 <sub>33.0</sub>	$1.0 \times 10^0$	0.0	18.1 <sub>38.5</sub>	$7.7 \times 10^{-22}$	0.2	11.7 <sub>33.0</sub>	13.3 <sub>34.0</sub>	10.1 <sub>30.2</sub>
Problem-Solving	53.4 <sub>49.9</sub>	$1.7 \times 10^{-229}$	0.6	60.1 <sub>49.0</sub>	$0.0 \times 10^0$	0.7	33.7 <sub>47.3</sub>	$6.9 \times 10^{-21}$	0.2	24.3 <sub>42.9</sub>	$5.4 \times 10^{-1}$	0.0	25.9 <sub>42.6</sub>	21.2 <sub>40.9</sub>	30.5 <sub>46.0</sub>
Planning	47.9 <sub>50.0</sub>	$3.6 \times 10^{-306}$	0.7	53.3 <sub>49.9</sub>	$0.0 \times 10^0$	0.8	32.5 <sub>46.8</sub>	$7.2 \times 10^{-87}$	0.4	26.8 <sub>44.3</sub>	$5.8 \times 10^{-38}$	0.2	17.1 <sub>37.6</sub>	16.9 <sub>37.5</sub>	17.2 <sub>37.8</sub>
Normalizing	43.5 <sub>49.6</sub>	$1.9 \times 10^{-171}$	0.5	47.6 <sub>49.9</sub>	$8.9 \times 10^{-228}$	0.6	16.6 <sub>37.2</sub>	$1.4 \times 10^{-7}$	-0.1	19.4 <sub>39.5</sub>	$1.0 \times 10^0$	0.0	20.5 <sub>39.5</sub>	17.8 <sub>38.3</sub>	23.2 <sub>42.2</sub>
Psychoeducation	19.1 <sub>39.3</sub>	$1.6 \times 10^{-79}$	0.3	18.7 <sub>39.0</sub>	$9.5 \times 10^{-74}$	0.3	12.8 <sub>33.5</sub>	$1.2 \times 10^{-21}$	0.2	7.4 <sub>26.3</sub>	$1.0 \times 10^0$	0.0	7.6 <sub>24.8</sub>	5.2 <sub>22.3</sub>	9.9 <sub>29.8</sub>

**Table C.12:** Frequency of conversational behaviors exhibited by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS by 33.5% and GPT-4 with full conversation simulation responds by 41.9% when compared with the human therapists responds by 23.4% on average from Table C.11. (Single:  $P = 2.6 \times 10^{-14}$ , Cohen’s  $d = 0.2$ ; two-sided student t-test) (Full:  $P = 8.1 \times 10^{-99}$ , Cohen’s  $d = 0.4$ ; two-sided student t-test)

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4 ( $n = 6158$ )			GPT-3.5 ( $n = 6144$ )			Llama2-70b ( $n = 6174$ )			Llama2-13b ( $n = 6282$ )			GPT-4 ( $n = 1547$ )			GPT-3.5 ( $n = 991$ )			Llama2-70b ( $n = 1287$ )			Llama2-13b ( $n = 1182$ )		
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$
Reflections on Needs	33.5 <sub>47.2</sub>	$2.6 \times 10^{-14}$	0.2	39.6 <sub>48.9</sub>	$4.5 \times 10^{-25}$	0.4	19.6 <sub>39.7</sub>	$4.9 \times 10^{-2}$	-0.1	19.5 <sub>39.7</sub>	$6.2 \times 10^{-2}$	-0.1	41.9 <sub>49.3</sub>	$8.1 \times 10^{-99}$	0.4	34.3 <sub>47.5</sub>	$1.3 \times 10^{-36}$	0.2	30.9 <sub>46.2</sub>	$2.2 \times 10^{-18}$	0.2	30.6 <sub>46.1</sub>	$6.6 \times 10^{-17}$	0.2
Reflections on Emotions	10.0 <sub>29.9</sub>	$1.0 \times 10^0$	0.0	13.5 <sub>34.2</sub>	$9.4 \times 10^{-2}$	0.1	9.9 <sub>29.9</sub>	$1.0 \times 10^0$	0.0	16.1 <sub>36.7</sub>	$2.4 \times 10^{-6}$	0.2	26.4 <sub>44.1</sub>	$5.5 \times 10^{-103}$	0.4	27.1 <sub>44.5</sub>	$8.8 \times 10^{-111}$	0.4	18.6 <sub>38.9</sub>	$3.0 \times 10^{-32}$	0.2	22.3 <sub>41.6</sub>	$1.5 \times 10^{-62}$	0.3
Reflections on Values	5.1 <sub>22.0</sub>	$1.0 \times 10^0$	0.0	4.3 <sub>20.4</sub>	$1.0 \times 10^0$	0.0	2.7 <sub>16.3</sub>	$2.2 \times 10^{-2}$	-0.1	3.1 <sub>17.4</sub>	$2.0 \times 10^{-1}$	-0.1	4.3 <sub>20.2</sub>	$1.0 \times 10^0$	0.0	3.4 <sub>18.0</sub>	$1.6 \times 10^{-3}$	-0.1	4.0 <sub>19.5</sub>	$3.3 \times 10^{-1}$	0.0	4.3 <sub>20.2</sub>	$1.0 \times 10^0$	0.0
Reflections on Consequences	10.9 <sub>31.2</sub>	$1.3 \times 10^{-21}$	-0.3	12.0 <sub>32.5</sub>	$5.6 \times 10^{-12}$	-0.3	5.9 <sub>23.6</sub>	$7.3 \times 10^{-41}$	-0.4	8.5 <sub>28.0</sub>	$7.1 \times 10^{-26}$	-0.3	25.2 <sub>43.4</sub>	$1.6 \times 10^{-4}$	0.1	22.9 <sub>42.0</sub>	$1.0 \times 10^0$	0.0	22.5 <sub>41.8</sub>	$1.0 \times 10^0$	0.0	23.6 <sub>42.5</sub>	$2.6 \times 10^{-1}$	0.0
Reflections on Conflicts	2.0 <sub>14.0</sub>	$2.5 \times 10^{-10}$	-0.2	1.8 <sub>13.4</sub>	$1.7 \times 10^{-7}$	-0.2	2.3 <sub>15.1</sub>	$1.5 \times 10^{-7}$	-0.2	3.9 <sub>19.3</sub>	$7.9 \times 10^{-3}$	-0.1	4.6 <sub>21.0</sub>	$5.3 \times 10^{-5}$	-0.1	3.8 <sub>19.2</sub>	$3.2 \times 10^{-10}$	-0.1	7.1 <sub>25.6</sub>	$1.0 \times 10^0$	0.0	7.1 <sub>25.7</sub>	$1.0 \times 10^0$	0.0
Reflections on Strengths	70.4 <sub>48.7</sub>	$0.0 \times 10^0$	1.7	56.5 <sub>49.6</sub>	$5.5 \times 10^{-300}$	1.4	35.3 <sub>47.8</sub>	$3.4 \times 10^{-118}$	0.7	31.6 <sub>45.5</sub>	$4.3 \times 10^{-85}$	0.6	24.3 <sub>42.9</sub>	$4.5 \times 10^{-94}$	0.4	22.7 <sub>41.9</sub>	$6.2 \times 10^{-78}$	0.4	13.0 <sub>33.6</sub>	$1.4 \times 10^{-7}$	0.1	12.6 <sub>33.2</sub>	$5.2 \times 10^{-6}$	0.1
Questions on Experiences	31.5 <sub>46.5</sub>	$2.3 \times 10^{-17}$	-0.3	21.6 <sub>41.2</sub>	$5.9 \times 10^{-39}$	-0.5	38.3 <sub>48.6</sub>	$2.8 \times 10^{-3}$	-0.1	41.1 <sub>49.3</sub>	$9.3 \times 10^{-1}$	-0.1	38.0 <sub>48.5</sub>	$8.5 \times 10^{-26}$	-0.1	30.7 <sub>46.1</sub>	$2.4 \times 10^{-48}$	-0.3	61.6 <sub>48.6</sub>	$2.0 \times 10^{-80}$	0.4	59.5 <sub>49.1</sub>	$1.3 \times 10^{-62}$	0.3
Questions on Perspectives	4.5 <sub>20.8</sub>	$9.1 \times 10^{-1}$	0.1	2.7 <sub>16.3</sub>	$1.0 \times 10^0$	0.0	5.5 <sub>22.8</sub>	$1.7 \times 10^{-2}$	0.1	5.4 <sub>22.6</sub>	$4.0 \times 10^{-2}$	0.1	8.4 <sub>27.8</sub>	$2.4 \times 10^{-26}$	0.2	4.0 <sub>19.5</sub>	$1.0 \times 10^0$	0.0	13.5 <sub>34.2</sub>	$2.9 \times 10^{-78}$	0.4	12.4 <sub>33.0</sub>	$5.0 \times 10^{-66}$	0.3
Questions on Emotions	1.5 <sub>12.1</sub>	$1.2 \times 10^{-31}$	-0.3	3.0 <sub>17.1</sub>	$8.2 \times 10^{-15}$	-0.3	3.1 <sub>17.4</sub>	$1.8 \times 10^{-18}$	-0.3	8.5 <sub>28.0</sub>	$2.7 \times 10^{-2}$	-0.1	3.4 <sub>18.2</sub>	$1.9 \times 10^{-63}$	-0.3	6.6 <sub>24.8</sub>	$3.5 \times 10^{-20}$	-0.2	14.4 <sub>35.1</sub>	$2.9 \times 10^{-4}$	0.1	19.9 <sub>39.9</sub>	$7.3 \times 10^{-32}$	0.2
Problem-Solving	40.5 <sub>49.1</sub>	$2.1 \times 10^{-29}$	0.3	58.5 <sub>49.3</sub>	$3.6 \times 10^{-99}$	0.7	51.4 <sub>50.0</sub>	$1.4 \times 10^{-74}$	0.6	34.6 <sub>47.6</sub>	$5.4 \times 10^{-81}$	0.2	56.6 <sub>49.6</sub>	$1.1 \times 10^{-262}$	0.7	60.4 <sub>48.9</sub>	$0.0 \times 10^0$	0.7	30.0 <sub>45.8</sub>	$7.6 \times 10^{-6}$	0.1	22.4 <sub>41.7</sub>	$9.8 \times 10^{-5}$	-0.1
Planning	66.3 <sub>47.3</sub>	$0.0 \times 10^0$	1.2	73.2 <sub>44.3</sub>	$0.0 \times 10^0$	1.4	61.6 <sub>48.7</sub>	$2.6 \times 10^{-258}$	1.1	42.0 <sub>49.4</sub>	$4.2 \times 10^{-81}$	0.6	43.3 <sub>49.6</sub>	$1.8 \times 10^{-211}$	0.6	50.1 <sub>50.0</sub>	$0.0 \times 10^0$	0.7	26.4 <sub>44.1</sub>	$5.3 \times 10^{-33}$	0.2	23.9 <sub>42.7</sub>	$7.0 \times 10^{-19}$	0.2
Normalizing	61.9 <sub>48.6</sub>	$7.1 \times 10^{-340}$	1.0	71.2 <sub>45.3</sub>	$6.9 \times 10^{-262}$	1.3	23.9 <sub>42.7</sub>	$7.7 \times 10^{-2}$	0.1	23.8 <sub>42.6</sub>	$1.4 \times 10^{-1}$	0.1	38.9 <sub>48.7</sub>	$6.0 \times 10^{-105}$	0.4	43.8 <sub>49.6</sub>	$3.8 \times 10^{-163}$	0.5	15.1 <sub>35.8</sub>	$1.1 \times 10^{-13}$	-0.1	18.5 <sub>38.9</sub>	$8.9 \times 10^{-2}$	-0.1
Psychoeducation	23.4 <sub>42.4</sub>	$2.8 \times 10^{-74}$	0.5	33.4 <sub>47.2</sub>	$1.4 \times 10^{-135}$	0.9	29.4 <sub>45.6</sub>	$7.5 \times 10^{-117}$	0.7	17.8 <sub>38.2</sub>	$3.1 \times 10^{-29}$	0.4	18.0 <sub>38.4</sub>	$3.0 \times 10^{-64}$	0.3	16.3 <sub>37.0</sub>	$5.3 \times 10^{-48}$	0.3	9.4 <sub>29.2</sub>	$3.8 \times 10^{-3}$	0.1	5.5 <sub>22.8</sub>	$3.9 \times 10^{-5}$	-0.1

**Table C.13:** Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 responds with REFLECTIONS ON NEEDS at 3.3 turns earlier than average human therapists responds. (GPT-4: 2.9 turn, Average human therapists: 6.2 turn;  $P = 1.0 \times 10^{-19}$ , Cohen’s  $d = 0.8$ , two-sided Student’s  $t$ -test).

Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ( $n = 7705$ )			GPT-3.5 ( $n = 7135$ )			Llama2-70b ( $n = 7461$ )			Llama2-13b ( $n = 7464$ )			Average ( $n = 5446$ )	High-quality ( $n = 3907$ )	Low-quality ( $n = 1539$ )
	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	turn	turn
Reflections on Needs	2.9 <sub>3,3</sub>	$1.0 \times 10^{-19}$	0.8	6.1 <sub>6,7</sub>	$1.0 \times 10^0$	0.0	5.3 <sub>5,5</sub>	$4.8 \times 10^{-1}$	0.2	6.1 <sub>6,4</sub>	$1.0 \times 10^0$	0.0	6.2 <sub>5,5</sub>	4.4 <sub>4,5</sub>	8.1 <sub>6,2</sub>
Reflections on Emotions	5.8 <sub>4,9</sub>	$1.0 \times 10^0$	0.1	6.0 <sub>5,3</sub>	$1.0 \times 10^0$	0.1	6.4 <sub>5,1</sub>	$1.0 \times 10^0$	0.0	7.0 <sub>5,8</sub>	$1.0 \times 10^0$	-0.1	6.3 <sub>4,9</sub>	5.2 <sub>4,8</sub>	7.4 <sub>5,0</sub>
Reflections on Values	9.0 <sub>5,9</sub>	$1.0 \times 10^0$	0.2	9.3 <sub>6,4</sub>	$1.0 \times 10^0$	0.1	10.3 <sub>6,0</sub>	$1.0 \times 10^0$	-0.1	10.3 <sub>5,8</sub>	$1.0 \times 10^0$	-0.1	9.9 <sub>5,5</sub>	8.8 <sub>5,3</sub>	11.0 <sub>5,8</sub>
Reflections on Consequences	7.0 <sub>4,8</sub>	$1.0 \times 10^0$	0.0	7.1 <sub>5,7</sub>	$1.0 \times 10^0$	0.0	6.2 <sub>4,9</sub>	$1.0 \times 10^0$	0.2	7.3 <sub>5,7</sub>	$1.0 \times 10^0$	-0.1	7.0 <sub>5,0</sub>	6.2 <sub>4,6</sub>	7.8 <sub>5,4</sub>
Reflections on Conflicts	8.2 <sub>4,6</sub>	$5.8 \times 10^{-2}$	0.4	8.8 <sub>5,3</sub>	$1.0 \times 10^0$	0.2	8.7 <sub>5,3</sub>	$4.2 \times 10^{-1}$	0.2	8.1 <sub>5,3</sub>	$2.3 \times 10^{-2}$	0.4	10.0 <sub>5,2</sub>	8.2 <sub>4,7</sub>	11.8 <sub>5,6</sub>
Reflections on Strengths	7.8 <sub>5,1</sub>	$4.6 \times 10^{-5}$	0.4	8.9 <sub>6,2</sub>	$3.9 \times 10^{-1}$	0.2	9.1 <sub>6,3</sub>	$1.0 \times 10^0$	0.2	10.1 <sub>6,8</sub>	$1.0 \times 10^0$	0.0	10.2 <sub>6,1</sub>	8.4 <sub>6,0</sub>	11.9 <sub>5,7</sub>
Questions on Experiences	2.7 <sub>2,8</sub>	$1.0 \times 10^0$	0.1	4.1 <sub>5,7</sub>	$2.9 \times 10^{-2}$	-0.3	3.1 <sub>3,1</sub>	$1.0 \times 10^0$	-0.1	4.0 <sub>4,1</sub>	$1.8 \times 10^{-3}$	-0.3	2.9 <sub>2,5</sub>	2.2 <sub>2,1</sub>	3.6 <sub>2,9</sub>
Questions on Perspectives	7.7 <sub>4,1</sub>	$1.4 \times 10^{-13}$	1.1	7.8 <sub>4,5</sub>	$2.9 \times 10^{-9}$	1.0	8.9 <sub>5,4</sub>	$8.1 \times 10^{-7}$	0.7	8.4 <sub>5,2</sub>	$6.1 \times 10^{-9}$	0.8	12.6 <sub>5,1</sub>	10.0 <sub>4,7</sub>	15.3 <sub>4,2</sub>
Questions on Emotions	6.4 <sub>5,6</sub>	$1.0 \times 10^0$	0.1	7.7 <sub>6,0</sub>	$1.0 \times 10^0$	-0.1	5.0 <sub>5,2</sub>	$1.0 \times 10^{-2}$	0.3	4.4 <sub>4,4</sub>	$2.9 \times 10^{-6}$	0.5	7.0 <sub>6,4</sub>	6.7 <sub>6,3</sub>	7.3 <sub>6,5</sub>
Problem-Solving	7.9 <sub>5,2</sub>	$1.0 \times 10^0$	0.1	7.0 <sub>5,6</sub>	$4.7 \times 10^{-2}$	0.2	8.7 <sub>5,4</sub>	$1.0 \times 10^0$	-0.1	9.5 <sub>6,2</sub>	$1.9 \times 10^{-1}$	-0.2	8.3 <sub>5,4</sub>	8.7 <sub>5,6</sub>	8.0 <sub>5,1</sub>
Planning	9.7 <sub>5,7</sub>	$2.7 \times 10^{-4}$	0.4	9.3 <sub>5,6</sub>	$8.6 \times 10^{-6}$	0.4	9.7 <sub>5,7</sub>	$2.7 \times 10^{-4}$	0.4	10.5 <sub>6,1</sub>	$1.3 \times 10^{-1}$	0.2	11.9 <sub>6,5</sub>	11.5 <sub>6,8</sub>	12.4 <sub>6,0</sub>
Normalizing	8.0 <sub>5,8</sub>	$4.5 \times 10^{-1}$	-0.2	7.3 <sub>6,1</sub>	$1.0 \times 10^0$	-0.1	9.5 <sub>6,6</sub>	$7.5 \times 10^{-5}$	-0.4	8.8 <sub>6,6</sub>	$1.4 \times 10^{-2}$	-0.3	6.9 <sub>6,2</sub>	6.2 <sub>6,3</sub>	7.7 <sub>6,0</sub>
Psychoeducation	11.0 <sub>5,9</sub>	$2.1 \times 10^{-2}$	-0.4	10.1 <sub>5,9</sub>	$9.8 \times 10^{-1}$	-0.2	12.1 <sub>5,7</sub>	$9.3 \times 10^{-6}$	-0.6	12.8 <sub>6,2</sub>	$2.2 \times 10^{-6}$	-0.7	8.9 <sub>5,5</sub>	8.5 <sub>5,6</sub>	9.2 <sub>5,5</sub>

**Table C.14:** Temporal Order (First Occurrence) of conversational behaviors during a conversation by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with REFLECTIONS ON NEEDS at 4.0 turn which is 2.2 turns earlier than average human therapists. Similarly, GPT-4 with full conversation simulation responds at 1.8 turn which is 4.4 turns earlier than average human therapists. (Single: 4.0 turn, Average human therapists: 6.2 turn from Table C.13;  $P = 6.4 \times 10^{-6}$ , Cohen’s  $d = 0.5$ , two-sided Student’s  $t$ -test) (Full: 1.8 turn, Average human therapists: 6.2 turn from Table C.13;  $P = 4.7 \times 10^{-21}$ , Cohen’s  $d = 1.0$ ).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)													
	GPT-4 ( $n = 6158$ )			GPT-3.5 ( $n = 6144$ )			Llama2-70b ( $n = 6174$ )			Llama2-13b ( $n = 6282$ )			GPT-4 ( $n = 1547$ )			GPT-3.5 ( $n = 991$ )			Llama2-70b ( $n = 1287$ )			Llama2-13b ( $n = 1182$ )				
	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$	$d$	turn	$p$
Reflections on Needs	4.0 <sub>3,0</sub>	$6.4 \times 10^{-6}$	0.5	9.3 <sub>7,4</sub>	$1.3 \times 10^{-5}$	-0.5	9.3 <sub>5,3</sub>	$1.3 \times 10^{-6}$	-0.6	11.5 <sub>6,3</sub>	$2.1 \times 10^{-15}$	-0.9	1.8 <sub>3,2</sub>	$4.7 \times 10^{-21}$	1.0	2.8 <sub>3,8</sub>	$7.5 \times 10^{-12}$	0.7	2.4 <sub>3,4</sub>	$5.3 \times 10^{-16}$	0.8	2.6 <sub>3,1</sub>	$1.5 \times 10^{-15}$	0.8		
Reflections on Emotions	8.2 <sub>5,0</sub>	$3.5 \times 10^{-2}$	-0.4	9.5 <sub>5,8</sub>	$6.9 \times 10^{-5}$	-0.6	9.5 <sub>4,7</sub>	$2.4 \times 10^{-5}$	-0.7	11.7 <sub>5,2</sub>	$3.6 \times 10^{-16}$	-1.1	4.5 <sub>4,3</sub>	$2.9 \times 10^{-3}$	0.4	4.3 <sub>4,0</sub>	$3.8 \times 10^{-4}$	0.4	5.0 <sub>4,7</sub>	$1.7 \times 10^{-1}$	0.3	4.2 <sub>4,0</sub>	$9.6 \times 10^{-5}$	0.5		
Reflections on Values	8.4 <sub>5,3</sub>	$1.0 \times 10^0$	0.3	8.9 <sub>2,2</sub>	$1.0 \times 10^0$	0.2	14.2 <sub>4,9</sub>	$4.4 \times 10^{-3}$	-0.8	14.4 <sub>5,2</sub>	$1.1 \times 10^{-3}$	-0.8	9.4 <sub>6,2</sub>	$1.0 \times 10^0$	0.1	9.6 <sub>5,8</sub>	$1.0 \times 10^0$	0.1	9.2 <sub>5,9</sub>	$1.0 \times 10^0$	0.1	9.2 <sub>5,5</sub>	$1.0 \times 10^0$	0.1		
Reflections on Consequences	8.6 <sub>4,4</sub>	$4.4 \times 10^{-2}$	-0.3	9.9 <sub>6,4</sub>	$3.7 \times 10^{-4}$	-0.5	10.1 <sub>4,7</sub>	$6.0 \times 10^{-4}$	-0.6	13.0 <sub>5,3</sub>	$2.4 \times 10^{-14}$	-1.2	6.0 <sub>4,8</sub>	$6.0 \times 10^{-1}$	0.2	5.8 <sub>4,9</sub>	$2.0 \times 10^0$	0.2	5.1 <sub>4,3</sub>	$5.1 \times 10^{-4}$	0.4	5.5 <sub>4,4</sub>	$1.3 \times 10^{-2}$	0.3		
Reflections on Conflicts	9.2 <sub>3,3</sub>	$1.0 \times 10^0$	0.2	8.8 <sub>5,2</sub>	$1.0 \times 10^0$	0.2	10.3 <sub>5,0</sub>	$1.0 \times 10^0$	-0.1	11.8 <sub>4,9</sub>	$1.0 \times 10^0$	-0.3	8.0 <sub>4,8</sub>	$4.1 \times 10^{-2}$	0.4	8.8 <sub>5,3</sub>	$1.0 \times 10^0$	0.2	8.4 <sub>5,3</sub>	$1.6 \times 10^{-1}$	0.3	7.2 <sub>3,1</sub>	$1.7 \times 10^{-4}$	0.5		
Reflections on Strengths	8.4 <sub>4,8</sub>	$2.1 \times 10^{-2}$	0.3	11.5 <sub>6,2</sub>	$5.3 \times 10^{-1}$	-0.2	11.2 <sub>5,8</sub>	$1.0 \times 10^0$	-0.2	12.4 <sub>6,3</sub>	$2.0 \times 10^{-2}$	-0.4	7.3 <sub>3,3</sub>	$3.0 \times 10^{-5}$	0.5	6.6 <sub>5,2</sub>	$5.5 \times 10^{-8}$	0.6	7.0 <sub>6,1</sub>	$2.2 \times 10^{-5}$	0.5	8.3 <sub>6,6</sub>	$9.8 \times 10^{-2}$	0.3		
Questions on Experiences	3.9 <sub>2,6</sub>	$7.9 \times 10^{-4}$	-0.4	8.1 <sub>7,1</sub>	$3.2 \times 10^{-21}$	-1.1	4.9 <sub>3,1</sub>	$3.9 \times 10^{-12}$	-0.7	6.5 <sub>4,2</sub>	$1.5 \times 10^{-25}$	-1.1	1.5 <sub>2,5</sub>	$3.9 \times 10^{-8}$	0.6	1.6 <sub>2,4</sub>	$1.7 \times 10^{-7}$	0.5	1.4 <sub>1,9</sub>	$1.9 \times 10^{-11}$	0.7	1.5 <sub>1,9</sub>	$1.8 \times 10^{-9}$	0.6		
Questions on Perspectives	8.3 <sub>3,9</sub>	$3.5 \times 10^{-4}$	0.9	8.0 <sub>7,9</sub>	$5.2 \times 10^{-1}$	0.9	13.0 <sub>4,3</sub>	$1.0 \times 10^0$	-0.1	13.5 <sub>4,8</sub>	$1.0 \times 10^0$	-0.2	7.5 <sub>4,1</sub>	$1.3 \times 10^{-12}$	1.1	7.8 <sub>4,2</sub>	$2.3 \times 10^{-9}$	1.0	7.7 <sub>5,1</sub>	$4.1 \times 10^{-11}$	1.0	7.4 <sub>4,7</sub>	$3.5 \times 10^{-14}$	1.1		
Questions on Emotions	6.0 <sub>5,9</sub>	$1.0 \times 10^0$	0.2	8.1 <sub>6,5</sub>	$1.0 \times 10^0$	-0.2	7.6 <sub>6,1</sub>	$1.0 \times 10^0$	-0.1	5.5 <sub>3,3</sub>	$5.3 \times 10^{-1}$	0.3	6.5 <sub>5,5</sub>	$1.0 \times 10^0$	0.1	7.6 <sub>6,0</sub>	$1.0 \times 10^0$	-0.1	4.5 <sub>4,9</sub>	$5.5 \times 10^{-4}$	0.4	3.9 <sub>4,7</sub>	$1.1 \times 10^{-6}$	0.6		
Problem-Solving	10.6 <sub>4,8</sub>	$8.3 \times 10^{-5}$	-0.4	10.2 <sub>5,7</sub>	$8.3 \times 10^{-3}$	-0.3	11.2 <sub>4,8</sub>	$1.0 \times 10^{-7}$	-0.6	13.1 <sub>5,5</sub>	$2.0 \times 10^{-15}$	-0.9	5.3 <sub>4,0</sub>	$2.6 \times 10^{-9}$	0.6	4.2 <sub>3,6</sub>	$3.4 \times 10^{-17}$	0.9	6.1 <sub>4,8</sub>	$1.9 \times 10^{-4}$	0.4	6.6 <sub>5,1</sub>	$8.9 \times 10^{-3}$	0.3		
Planning	11.4 <sub>5,5</sub>	$1.0 \times 10^0$	0.1	12.3 <sub>5,3</sub>	$1.0 \times 10^0$	-0.1	10.9 <sub>5,4</sub>	$1.0 \times 10^0$	0.2	12.6 <sub>6,0</sub>	$1.0 \times 10^0$	-0.1	8.0 <sub>5,4</sub>	$3.4 \times 10^{-9}$	0.7	6.9 <sub>4,6</sub>	$5.9 \times 10^{-16}$	0.9	8.4 <sub>5,8</sub>	$7.4 \times 10^{-7}$	0.6	8.6 <sub>5,7</sub>	$3.0 \times 10^{-6}$	0.5		
Normalizing	10.6 <sub>5,4</sub>	$8.6 \times 10^{-9}$	-0.6	10.9 <sub>5,8</sub>	$1.4 \times 10^{-8}$	-0.7	12.9 <sub>6,2</sub>	$2.1 \times 10^{-15}$	-1.0	12.3 <sub>6,5</sub>	$1.1 \times 10^{-10}$	-0.9	5.5 <sub>3,1</sub>	$1.5 \times 10^{-1}$	0.2	4.6 <sub>4,8</sub>	$2.8 \times 10^{-4}$	0.4	7.1 <sub>5,8</sub>	$1.0 \times 10^0$	0.0	7.0 <sub>5,8</sub>	$1.0 \times 10^0$	0.0		
Psychoeducation	13.1 <sub>5,2</sub>	$2.3 \times 10^{-6}$	-0.8	11.7 <sub>5,8</sub>	$1.2 \times 10^{-2}$	-0.5	13.3 <sub>5,3</sub>	$1.0 \times 10^{-8}$	-0.8	15.6 <sub>6,4</sub>	$8.1 \times 10^{-14}$	-1.2	9.6 <sub>5,9</sub>	$1.0 \times 10^0$	-0.1	9.2 <sub>5,8</sub>	$1.0 \times 10^0$	-0.1	10.7 <sub>5,7</sub>	$2.0 \times 10^{-1}$	-0.3	10.0 <sub>5,6</sub>	$1.0 \times 10^0$	-0.2		

**Table C.15:** Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 10.2% and human therapists responds by 27.7% on average when client expresses CHANGING UNHEALTHY BEHAVIOR. ( $P = 1.9 \times 10^{-46}$ , Cohen’s  $d = -0.5$ , two-sided Student’s  $t$ -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
<b>Client behavior: Changing unhealthy behavior</b>															
<i>n</i>	2552			2193			2349			2109			1553	1198	355
Reflections on Needs	10.2 <sub>30.3</sub>	$1.9 \times 10^{-46}$	-0.5	8.2 <sub>27.4</sub>	$2.9 \times 10^{-56}$	-0.5	6.0 <sub>23.8</sub>	$3.6 \times 10^{-78}$	-0.6	5.3 <sub>22.4</sub>	$1.7 \times 10^{-79}$	-0.7	27.7 <sub>45.5</sub>	30.6 <sub>46.1</sub>	24.8 <sub>43.2</sub>
Reflections on Emotions	2.7 <sub>16.2</sub>	$5.4 \times 10^{-21}$	-0.3	2.8 <sub>16.6</sub>	$5.9 \times 10^{-18}$	-0.3	2.1 <sub>14.4</sub>	$3.9 \times 10^{-24}$	-0.3	3.7 <sub>18.8</sub>	$7.0 \times 10^{-13}$	-0.3	10.4 <sub>33.3</sub>	14.6 <sub>35.3</sub>	6.2 <sub>24.1</sub>
Reflections on Values	2.0 <sub>14.1</sub>	$3.0 \times 10^{-4}$	-0.1	0.1 <sub>3.0</sub>	$8.2 \times 10^{-18}$	-0.3	1.0 <sub>9.8</sub>	$4.7 \times 10^{-10}$	-0.2	0.9 <sub>9.2</sub>	$4.4 \times 10^{-10}$	-0.2	4.8 <sub>23.7</sub>	7.0 <sub>25.5</sub>	2.5 <sub>15.7</sub>
Reflections on Consequences	4.5 <sub>20.8</sub>	$6.6 \times 10^{-108}$	-0.7	2.1 <sub>14.5</sub>	$3.9 \times 10^{-133}$	-0.9	1.8 <sub>13.3</sub>	$5.2 \times 10^{-148}$	-0.9	2.8 <sub>16.4</sub>	$5.6 \times 10^{-119}$	-0.8	27.3 <sub>42.3</sub>	19.9 <sub>39.9</sub>	34.6 <sub>47.7</sub>
Reflections on Conflicts	0.5 <sub>7.4</sub>	$5.3 \times 10^{-28}$	-0.4	0.3 <sub>5.6</sub>	$2.6 \times 10^{-27}$	-0.4	0.6 <sub>7.7</sub>	$2.2 \times 10^{-25}$	-0.4	0.9 <sub>9.5</sub>	$7.3 \times 10^{-20}$	-0.3	6.9 <sub>26.6</sub>	8.3 <sub>27.5</sub>	5.6 <sub>23.1</sub>
Reflections on Strengths	31.6 <sub>46.5</sub>	$5.7 \times 10^{-58}$	0.5	20.1 <sub>40.1</sub>	$1.9 \times 10^{-16}$	0.3	14.8 <sub>35.5</sub>	$7.5 \times 10^{-5}$	0.2	12.0 <sub>32.6</sub>	$9.2 \times 10^{-1}$	0.1	9.3 <sub>32.4</sub>	14.1 <sub>34.8</sub>	4.5 <sub>20.8</sub>
Questions on Experiences	6.5 <sub>24.7</sub>	$1.4 \times 10^{-129}$	-0.8	1.2 <sub>11.0</sub>	$6.8 \times 10^{-197}$	-1.1	6.3 <sub>24.2</sub>	$8.0 \times 10^{-126}$	-0.8	6.2 <sub>24.1</sub>	$7.0 \times 10^{-117}$	-0.8	35.6 <sub>48.6</sub>	40.2 <sub>49.1</sub>	31.0 <sub>46.3</sub>
Questions on Perspectives	1.1 <sub>10.4</sub>	$2.6 \times 10^{-5}$	-0.2	0.0 <sub>2.1</sub>	$7.7 \times 10^{-14}$	-0.3	1.7 <sub>12.8</sub>	$2.8 \times 10^{-2}$	-0.1	1.0 <sub>9.9</sub>	$3.9 \times 10^{-5}$	-0.2	3.5 <sub>20.2</sub>	4.8 <sub>21.5</sub>	2.3 <sub>14.9</sub>
Questions on Emotions	0.2 <sub>4.4</sub>	$6.3 \times 10^{-49}$	-0.5	0.1 <sub>3.7</sub>	$3.6 \times 10^{-43}$	-0.5	0.4 <sub>6.5</sub>	$1.4 \times 10^{-41}$	-0.5	0.2 <sub>4.4</sub>	$8.3 \times 10^{-41}$	-0.5	9.6 <sub>30.7</sub>	11.3 <sub>31.6</sub>	7.9 <sub>27.0</sub>
Problem-Solving	14.0 <sub>34.7</sub>	$1.5 \times 10^{-29}$	-0.4	14.4 <sub>35.1</sub>	$1.0 \times 10^{-25}$	-0.4	20.3 <sub>40.2</sub>	$2.1 \times 10^{-7}$	-0.2	10.3 <sub>30.4</sub>	$3.5 \times 10^{-47}$	-0.5	28.4 <sub>42.7</sub>	20.4 <sub>40.3</sub>	36.3 <sub>48.2</sub>
Planning	19.5 <sub>39.6</sub>	$8.7 \times 10^{-5}$	0.2	15.3 <sub>36.0</sub>	$1.0 \times 10^0$	0.0	22.9 <sub>42.0</sub>	$2.5 \times 10^{-11}$	0.2	11.9 <sub>32.4</sub>	$1.0 \times 10^0$	-0.1	13.7 <sub>32.6</sub>	10.7 <sub>30.9</sub>	16.6 <sub>37.3</sub>
Normalizing	19.3 <sub>39.5</sub>	$6.3 \times 10^{-4}$	0.1	15.8 <sub>36.5</sub>	$1.0 \times 10^0$	0.1	6.5 <sub>24.7</sub>	$4.4 \times 10^{-14}$	-0.3	5.1 <sub>22.0</sub>	$7.8 \times 10^{-21}$	-0.3	14.0 <sub>32.6</sub>	10.5 <sub>30.7</sub>	17.5 <sub>38.0</sub>
Psychoeducation	3.4 <sub>18.0</sub>	$6.5 \times 10^{-2}$	-0.1	3.0 <sub>17.0</sub>	$5.1 \times 10^{-3}$	-0.1	8.3 <sub>27.5</sub>	$3.9 \times 10^{-2}$	0.1	4.1 <sub>19.9</sub>	$1.0 \times 10^0$	-0.1	5.4 <sub>20.5</sub>	3.5 <sub>18.4</sub>	7.3 <sub>26.1</sub>
<b>Client behavior: Sustaining unhealthy behavior</b>															
<i>n</i>	428			402			485			462			389	275	114
Reflections on Needs	5.4 <sub>22.6</sub>	$1.5 \times 10^{-16}$	-0.6	2.2 <sub>14.8</sub>	$7.6 \times 10^{-23}$	-0.8	7.0 <sub>25.6</sub>	$9.8 \times 10^{-15}$	-0.6	6.1 <sub>23.9</sub>	$4.4 \times 10^{-16}$	-0.6	27.6 <sub>45.3</sub>	30.5 <sub>46.1</sub>	24.6 <sub>43.2</sub>
Reflections on Emotions	0.5 <sub>6.8</sub>	$1.1 \times 10^{-9}$	-0.5	0.2 <sub>5.0</sub>	$1.1 \times 10^{-9}$	-0.5	2.5 <sub>15.5</sub>	$5.4 \times 10^{-6}$	-0.4	3.7 <sub>18.8</sub>	$8.2 \times 10^{-4}$	-0.3	11.7 <sub>33.2</sub>	13.8 <sub>34.6</sub>	9.6 <sub>29.7</sub>
Reflections on Values	1.2 <sub>10.8</sub>	$2.0 \times 10^{-3}$	-0.3	0.0 <sub>0.0</sub>	$5.6 \times 10^{-6}$	-0.4	0.6 <sub>7.8</sub>	$2.9 \times 10^{-5}$	-0.3	0.9 <sub>9.3</sub>	$2.2 \times 10^{-4}$	-0.3	6.8 <sub>25.0</sub>	6.5 <sub>24.8</sub>	7.0 <sub>25.7</sub>
Reflections on Consequences	4.7 <sub>21.1</sub>	$2.7 \times 10^{-31}$	-0.9	1.0 <sub>9.9</sub>	$1.5 \times 10^{-41}$	-1.0	4.5 <sub>20.8</sub>	$5.3 \times 10^{-35}$	-0.9	5.4 <sub>22.6</sub>	$5.4 \times 10^{-31}$	-0.9	36.1 <sub>46.8</sub>	26.5 <sub>44.2</sub>	45.0 <sub>50.0</sub>
Reflections on Conflicts	0.9 <sub>9.6</sub>	$2.0 \times 10^{-5}$	-0.4	1.2 <sub>11.1</sub>	$1.9 \times 10^{-4}$	-0.3	1.9 <sub>13.5</sub>	$5.9 \times 10^{-4}$	-0.3	0.9 <sub>9.3</sub>	$5.1 \times 10^{-6}$	-0.4	8.2 <sub>27.1</sub>	7.6 <sub>26.6</sub>	8.8 <sub>28.4</sub>
Reflections on Strengths	3.5 <sub>18.4</sub>	$1.0 \times 10^0$	-0.1	1.2 <sub>11.1</sub>	$1.9 \times 10^{-1}$	-0.2	4.3 <sub>20.4</sub>	$1.0 \times 10^0$	0.0	3.7 <sub>18.8</sub>	$1.0 \times 10^0$	-0.1	5.2 <sub>23.6</sub>	6.9 <sub>25.4</sub>	3.5 <sub>18.5</sub>
Questions on Experiences	3.0 <sub>17.2</sub>	$1.8 \times 10^{-39}$	-1.0	0.5 <sub>7.0</sub>	$5.0 \times 10^{-46}$	-1.1	6.0 <sub>23.7</sub>	$5.6 \times 10^{-34}$	-0.9	5.4 <sub>22.6</sub>	$2.2 \times 10^{-34}$	-0.9	39.3 <sub>49.4</sub>	46.2 <sub>49.9</sub>	32.5 <sub>47.0</sub>
Questions on Perspectives	0.7 <sub>8.4</sub>	$5.0 \times 10^{-1}$	-0.2	0.0 <sub>0.0</sub>	$2.3 \times 10^{-2}$	-0.3	1.4 <sub>11.9</sub>	$1.0 \times 10^0$	-0.1	1.7 <sub>13.1</sub>	$1.0 \times 10^0$	-0.1	3.5 <sub>19.3</sub>	4.4 <sub>20.5</sub>	2.6 <sub>16.1</sub>
Questions on Emotions	0.0 <sub>0.0</sub>	$1.8 \times 10^{-8}$	-0.4	0.0 <sub>0.0</sub>	$6.3 \times 10^{-8}$	-0.4	0.0 <sub>0.0</sub>	$1.2 \times 10^{-9}$	-0.5	0.2 <sub>4.7</sub>	$1.8 \times 10^{-8}$	-0.4	9.2 <sub>29.7</sub>	10.5 <sub>30.8</sub>	7.9 <sub>27.1</sub>
Problem-Solving	2.8 <sub>16.5</sub>	$1.1 \times 10^{-19}$	-0.7	1.5 <sub>12.1</sub>	$1.9 \times 10^{-22}$	-0.7	12.2 <sub>32.7</sub>	$3.2 \times 10^{-4}$	-0.3	6.9 <sub>25.4</sub>	$4.0 \times 10^{-11}$	-0.5	23.6 <sub>40.3</sub>	15.6 <sub>36.4</sub>	31.6 <sub>46.7</sub>
Planning	1.9 <sub>13.6</sub>	$1.1 \times 10^{-1}$	-0.2	0.5 <sub>7.0</sub>	$2.4 \times 10^{-4}$	-0.3	8.9 <sub>28.5</sub>	$1.0 \times 10^0$	0.1	4.1 <sub>19.9</sub>	$1.0 \times 10^0$	-0.1	5.9 <sub>22.1</sub>	4.0 <sub>19.6</sub>	7.9 <sub>27.1</sub>
Normalizing	2.1 <sub>14.4</sub>	$1.9 \times 10^{-9}$	-0.5	0.7 <sub>8.6</sub>	$2.0 \times 10^{-12}$	-0.6	2.3 <sub>14.9</sub>	$4.5 \times 10^{-10}$	-0.5	1.7 <sub>13.1</sub>	$4.4 \times 10^{-11}$	-0.5	14.0 <sub>33.2</sub>	10.5 <sub>30.8</sub>	17.5 <sub>38.2</sub>
Psychoeducation	0.2 <sub>4.8</sub>	$5.6 \times 10^{-11}$	-0.5	0.5 <sub>7.0</sub>	$2.1 \times 10^{-9}$	-0.5	3.3 <sub>17.9</sub>	$1.4 \times 10^{-3}$	-0.3	1.5 <sub>12.2</sub>	$2.6 \times 10^{-7}$	-0.4	9.8 <sub>26.7</sub>	4.7 <sub>21.3</sub>	14.9 <sub>35.8</sub>
<b>Client behavior: Sharing positive emotions</b>															
<i>n</i>	815			596			389			437			294	226	68
Reflections on Needs	4.0 <sub>19.7</sub>	$7.2 \times 10^{-11}$	-0.5	10.7 <sub>31.0</sub>	$6.0 \times 10^{-1}$	-0.2	2.8 <sub>16.6</sub>	$3.3 \times 10^{-8}$	-0.5	3.0 <sub>17.0</sub>	$8.8 \times 10^{-9}$	-0.5	17.3 <sub>40.9</sub>	24.3 <sub>43.0</sub>	10.3 <sub>30.6</sub>
Reflections on Emotions	2.3 <sub>15.1</sub>	$2.2 \times 10^{-2}$	-0.2	3.4 <sub>18.0</sub>	$1.0 \times 10^0$	-0.2	0.3 <sub>5.1</sub>	$9.7 \times 10^{-4}$	-0.3	0.9 <sub>9.5</sub>	$4.4 \times 10^{-3}$	-0.3	7.5 <sub>32.0</sub>	15.0 <sub>35.8</sub>	0.0 <sub>0.0</sub>
Reflections on Values	0.9 <sub>9.2</sub>	$3.9 \times 10^{-1}$	-0.2	0.2 <sub>4.1</sub>	$1.9 \times 10^{-2}$	-0.3	0.8 <sub>8.8</sub>	$1.0 \times 10^0$	-0.2	0.2 <sub>4.8</sub>	$1.6 \times 10^{-1}$	-0.2	3.4 <sub>20.6</sub>	5.3 <sub>22.5</sub>	1.5 <sub>12.1</sub>
Reflections on Consequences	0.6 <sub>7.8</sub>	$4.0 \times 10^{-29}$	-0.8	0.7 <sub>8.2</sub>	$1.0 \times 10^{-21}$	-0.7	0.5 <sub>7.2</sub>	$3.1 \times 10^{-15}$	-0.7	0.9 <sub>9.5</sub>	$2.0 \times 10^{-15}$	-0.7	15.8 <sub>34.0</sub>	11.1 <sub>31.4</sub>	20.6 <sub>40.7</sub>
Reflections on Conflicts	0.0 <sub>0.0</sub>	$2.2 \times 10^{-4}$	-0.3	0.0 <sub>0.0</sub>	$4.9 \times 10^{-3}$	-0.3	0.0 <sub>0.0</sub>	$9.5 \times 10^{-2}$	-0.3	0.7 <sub>8.3</sub>	$1.0 \times 10^0$	-0.2	3.4 <sub>20.6</sub>	5.3 <sub>22.5</sub>	1.5 <sub>12.1</sub>
Reflections on Strengths	62.2 <sub>48.5</sub>	$1.4 \times 10^{-55}$	1.2	42.8 <sub>49.5</sub>	$3.6 \times 10^{-21}$	0.7	14.4 <sub>35.2</sub>	$1.0 \times 10^0$	0.1	17.2 <sub>37.7</sub>	$8.4 \times 10^{-1}$	0.2	10.2 <sub>34.0</sub>	15.9 <sub>36.7</sub>	4.4 <sub>20.7</sub>
Questions on Experiences	0.9 <sub>9.2</sub>	$3.1 \times 10^{-59}$	-1.2	0.2 <sub>4.1</sub>	$4.9 \times 10^{-49}$	-1.1	4.1 <sub>19.9</sub>	$3.2 \times 10^{-21}$	-0.8	6.6 <sub>24.9</sub>	$2.0 \times 10^{-17}$	-0.7	31.6 <sub>47.6</sub>	36.7 <sub>48.3</sub>	26.5 <sub>44.4</sub>
Questions on Perspectives	0.1 <sub>3.5</sub>	$1.0 \times 10^0$	-0.2	0.2 <sub>4.1</sub>	$1.0 \times 10^0$	-0.1	0.3 <sub>5.1</sub>	$1.0 \times 10^0$	-0.1	0.5 <sub>6.8</sub>	$1.0 \times 10^0$	-0.1	1.3 <sub>14.2</sub>	2.7 <sub>16.1</sub>	0.0 <sub>0.0</sub>
Questions on Emotions	0.0 <sub>0.0</sub>	$1.8 \times 10^{-11}$	-0.5	0.3 <sub>5.8</sub>	$6.0 \times 10^{-7}$	-0.4	0.0 <sub>0.0</sub>	$3.0 \times 10^{-5}$	-0.4	0.2 <sub>4.8</sub>	$2.4 \times 10^{-5}$	-0.4	7.5 <sub>28.9</sub>	10.6 <sub>30.9</sub>	4.4 <sub>20.7</sub>
Problem-Solving	7.7 <sub>26.7</sub>	$5.2 \times 10^{-1}$	-0.2	10.6 <sub>30.8</sub>	$1.0 \times 10^0$	-0.1	7.2 <sub>25.9</sub>	$8.2 \times 10^{-1}$	-0.2	3.4 <sub>18.2</sub>	$6.1 \times 10^{-5}$	-0.4	13.0 <sub>33.6</sub>	12.8 <sub>33.5</sub>	13.2 <sub>34.1</sub>
Planning	43.6 <sub>49.6</sub>	$1.1 \times 10^{-27}$	0.8	31.0 <sub>46.3</sub>	$1.0 \times 10^{-12}$	0.6	18.5 <sub>38.9</sub>	$8.1 \times 10^{-3}$	0.3	12.4 <sub>32.9</sub>	$1.0 \times 10^0$	0.1	8.2 <sub>26.9</sub>	7.5 <sub>26.4</sub>	8.8 <sub>28.6</sub>
Normalizing	58.5 <sub>49.3</sub>	$1.5 \times 10^{-25}$	0.8	38.9 <sub>48.8</sub>	$8.6 \times 10^{-5}$	0.3	8.2 <sub>27.5</sub>	$3.0 \times 10^{-6}$	-0.4	11.4 <sub>31.9</sub>	$2.3 \times 10^{-3}$	-0.3	22.7 <sub>40.6</sub>	19.0 <sub>39.3</sub>	26.5 <sub>44.4</sub>
Psychoeducation	3.4 <sub>18.2</sub>	$1.0 \times 10^0$	0.0	1.8 <sub>13.5</sub>	$1.0 \times 10^0$	-0.1	5.1 <sub>22.1</sub>	$1.0 \times 10^0$	0.1	3.0 <sub>17.0</sub>	$1.0 \times 10^0$	0.0	3.1 <sub>15.3</sub>	1.8 <sub>13.2</sub>	4.4 <sub>20.7</sub>

**Table C.16:** (Continue Table C.15) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits REFLECTIONS ON NEEDS by 14.0% and human therapists responds by 24.5% on average when client expresses SHARING NEGATIVE EMOTIONS. ( $P = 3.2 \times 10^{-7}$ , Cohen’s  $d = -0.3$ , two-sided Student’s  $t$ -test).

Therapist Behavior	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			Average	High-quality	Low-quality
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
<b>Client behavior: Sharing negative emotions</b>															
<i>n</i>	1175			1052			1354			1348			820	646	174
Reflections on Needs	14.0 <sub>34.8</sub>	$3.2 \times 10^{-7}$	-0.3	7.2 <sub>25.9</sub>	$2.3 \times 10^{-23}$	-0.5	9.5 <sub>29.4</sub>	$5.6 \times 10^{-19}$	-0.4	9.2 <sub>28.9</sub>	$5.1 \times 10^{-20}$	-0.4	24.5 <sub>44.5</sub>	28.9 <sub>45.4</sub>	20.1 <sub>40.2</sub>
Reflections on Emotions	9.0 <sub>28.7</sub>	$6.9 \times 10^{-9}$	-0.3	7.2 <sub>25.9</sub>	$2.6 \times 10^{-12}$	-0.4	8.1 <sub>27.3</sub>	$5.7 \times 10^{-12}$	-0.3	12.2 <sub>32.8</sub>	$8.3 \times 10^{-4}$	-0.2	19.4 <sub>42.7</sub>	27.4 <sub>44.6</sub>	11.5 <sub>32.0</sub>
Reflections on Values	2.5 <sub>15.5</sub>	$2.5 \times 10^{-3}$	-0.2	0.2 <sub>4.4</sub>	$2.5 \times 10^{-12}$	-0.4	0.8 <sub>9.0</sub>	$5.2 \times 10^{-11}$	-0.3	1.3 <sub>11.2</sub>	$2.2 \times 10^{-8}$	-0.3	6.3 <sub>25.4</sub>	7.4 <sub>26.2</sub>	5.2 <sub>22.2</sub>
Reflections on Consequences	7.1 <sub>25.8</sub>	$6.6 \times 10^{-21}$	-0.5	3.6 <sub>18.7</sub>	$1.7 \times 10^{-35}$	-0.6	3.5 <sub>18.3</sub>	$5.4 \times 10^{-44}$	-0.6	4.4 <sub>20.5</sub>	$1.8 \times 10^{-37}$	-0.6	21.8 <sub>39.8</sub>	18.3 <sub>38.7</sub>	25.3 <sub>43.6</sub>
Reflections on Conflicts	1.4 <sub>11.6</sub>	$1.3 \times 10^{-11}$	-0.3	0.6 <sub>7.5</sub>	$1.7 \times 10^{-14}$	-0.4	1.1 <sub>10.5</sub>	$1.4 \times 10^{-14}$	-0.4	2.3 <sub>15.0</sub>	$1.7 \times 10^{-8}$	-0.3	8.4 <sub>29.5</sub>	10.5 <sub>30.7</sub>	6.3 <sub>24.4</sub>
Reflections on Strengths	13.5 <sub>34.2</sub>	$8.4 \times 10^{-5}$	0.2	7.0 <sub>25.6</sub>	$1.0 \times 10^0$	0.0	12.6 <sub>33.1</sub>	$9.1 \times 10^{-4}$	0.2	10.4 <sub>30.5</sub>	$2.3 \times 10^{-1}$	0.1	6.6 <sub>26.6</sub>	8.5 <sub>27.9</sub>	4.6 <sub>21.0</sub>
Questions on Experiences	11.6 <sub>32.0</sub>	$4.0 \times 10^{-34}$	-0.6	3.5 <sub>18.4</sub>	$1.6 \times 10^{-72}$	-0.9	13.4 <sub>34.0</sub>	$3.6 \times 10^{-30}$	-0.5	15.7 <sub>36.4</sub>	$1.8 \times 10^{-22}$	-0.5	34.6 <sub>48.6</sub>	40.6 <sub>49.1</sub>	28.7 <sub>45.4</sub>
Questions on Perspectives	1.1 <sub>10.5</sub>	$1.4 \times 10^{-1}$	-0.1	0.2 <sub>4.4</sub>	$1.1 \times 10^{-4}$	-0.2	2.7 <sub>16.1</sub>	$1.0 \times 10^0$	0.0	2.2 <sub>14.5</sub>	$1.0 \times 10^0$	-0.1	3.3 <sub>19.9</sub>	4.8 <sub>21.4</sub>	1.7 <sub>13.1</sub>
Questions on Emotions	0.2 <sub>4.1</sub>	$2.7 \times 10^{-22}$	-0.5	0.4 <sub>6.2</sub>	$1.3 \times 10^{-18}$	-0.4	0.5 <sub>7.2</sub>	$1.7 \times 10^{-22}$	-0.5	0.4 <sub>6.1</sub>	$1.2 \times 10^{-23}$	-0.5	9.6 <sub>31.1</sub>	11.8 <sub>32.2</sub>	7.5 <sub>26.4</sub>
Problem-Solving	11.1 <sub>31.5</sub>	$2.1 \times 10^{-5}$	-0.2	14.2 <sub>34.9</sub>	$2.7 \times 10^{-1}$	-0.1	22.6 <sub>41.8</sub>	$1.0 \times 10^0$	0.1	15.4 <sub>36.1</sub>	$1.0 \times 10^0$	-0.1	19.0 <sub>35.7</sub>	12.1 <sub>32.6</sub>	25.9 <sub>43.9</sub>
Planning	5.3 <sub>22.4</sub>	$1.0 \times 10^0$	-0.1	6.0 <sub>23.7</sub>	$1.0 \times 10^0$	-0.1	17.6 <sub>38.1</sub>	$2.5 \times 10^{-9}$	0.3	10.4 <sub>30.5</sub>	$1.0 \times 10^0$	0.1	7.6 <sub>25.0</sub>	6.0 <sub>23.8</sub>	9.2 <sub>29.0</sub>
Normalizing	9.8 <sub>29.7</sub>	$4.1 \times 10^{-4}$	-0.2	10.6 <sub>30.7</sub>	$8.7 \times 10^{-3}$	-0.2	6.7 <sub>25.0</sub>	$9.3 \times 10^{-12}$	-0.3	6.7 <sub>25.0</sub>	$7.2 \times 10^{-12}$	-0.3	16.5 <sub>36.2</sub>	14.7 <sub>35.4</sub>	18.4 <sub>38.9</sub>
Psychoeducation	2.6 <sub>15.8</sub>	$2.8 \times 10^{-1}$	-0.1	2.9 <sub>16.9</sub>	$1.0 \times 10^0$	-0.1	7.6 <sub>26.5</sub>	$4.7 \times 10^{-1}$	0.1	4.3 <sub>20.3</sub>	$1.0 \times 10^0$	0.0	4.7 <sub>17.5</sub>	2.0 <sub>14.1</sub>	7.5 <sub>26.4</sub>
<b>Client behavior: Sharing experiences</b>															
<i>n</i>	874			812			908			926			759	533	226
Reflections on Needs	6.5 <sub>24.7</sub>	$1.5 \times 10^{-18}$	-0.5	1.4 <sub>11.6</sub>	$7.5 \times 10^{-38}$	-0.7	3.9 <sub>19.3</sub>	$5.6 \times 10^{-29}$	-0.6	5.0 <sub>21.7</sub>	$8.2 \times 10^{-25}$	-0.5	22.5 <sub>42.8</sub>	26.5 <sub>44.2</sub>	18.6 <sub>39.0</sub>
Reflections on Emotions	5.8 <sub>23.5</sub>	$3.3 \times 10^{-5}$	-0.3	3.0 <sub>16.9</sub>	$2.2 \times 10^{-11}$	-0.4	4.4 <sub>20.5</sub>	$2.1 \times 10^{-8}$	-0.3	6.0 <sub>23.8</sub>	$4.8 \times 10^{-5}$	-0.2	13.5 <sub>36.8</sub>	19.9 <sub>40.0</sub>	7.1 <sub>25.7</sub>
Reflections on Values	0.6 <sub>7.5</sub>	$6.6 \times 10^{-5}$	-0.2	0.0 <sub>0.0</sub>	$8.2 \times 10^{-7}$	-0.3	0.8 <sub>8.8</sub>	$2.3 \times 10^{-4}$	-0.2	1.3 <sub>11.3</sub>	$7.6 \times 10^{-3}$	-0.2	4.7 <sub>23.1</sub>	7.1 <sub>25.8</sub>	2.2 <sub>14.7</sub>
Reflections on Consequences	2.4 <sub>15.3</sub>	$4.9 \times 10^{-26}$	-0.6	1.0 <sub>9.9</sub>	$1.8 \times 10^{-31}$	-0.6	1.9 <sub>13.6</sub>	$1.2 \times 10^{-29}$	-0.6	1.7 <sub>13.0</sub>	$5.5 \times 10^{-31}$	-0.6	18.2 <sub>38.5</sub>	17.8 <sub>38.3</sub>	18.6 <sub>39.0</sub>
Reflections on Conflicts	1.0 <sub>10.1</sub>	$2.0 \times 10^{-5}$	-0.3	0.1 <sub>3.5</sub>	$1.7 \times 10^{-8}$	-0.3	0.9 <sub>9.4</sub>	$3.8 \times 10^{-6}$	-0.3	1.1 <sub>10.3</sub>	$1.6 \times 10^{-5}$	-0.3	5.8 <sub>25.1</sub>	8.1 <sub>27.3</sub>	3.5 <sub>18.5</sub>
Reflections on Strengths	4.5 <sub>20.7</sub>	$1.3 \times 10^{-1}$	-0.2	1.8 <sub>13.5</sub>	$1.1 \times 10^{-6}$	-0.3	4.1 <sub>19.8</sub>	$2.9 \times 10^{-2}$	-0.2	3.7 <sub>18.8</sub>	$5.5 \times 10^{-3}$	-0.2	8.4 <sub>29.9</sub>	12.0 <sub>32.5</sub>	4.9 <sub>21.6</sub>
Questions on Experiences	5.8 <sub>23.5</sub>	$3.2 \times 10^{-82}$	-1.0	2.3 <sub>15.1</sub>	$2.3 \times 10^{-100}$	-1.2	8.6 <sub>28.0</sub>	$1.3 \times 10^{-68}$	-0.9	9.1 <sub>28.7</sub>	$5.0 \times 10^{-67}$	-0.9	44.7 <sub>49.8</sub>	45.2 <sub>49.8</sub>	44.2 <sub>49.8</sub>
Questions on Perspectives	0.1 <sub>3.4</sub>	$3.2 \times 10^{-2}$	-0.2	0.0 <sub>0.0</sub>	$1.6 \times 10^{-2}$	-0.2	1.1 <sub>10.4</sub>	$1.0 \times 10^0$	-0.1	1.0 <sub>9.8</sub>	$1.0 \times 10^0$	-0.1	2.0 <sub>15.6</sub>	3.2 <sub>17.6</sub>	0.9 <sub>9.4</sub>
Questions on Emotions	0.1 <sub>3.4</sub>	$1.0 \times 10^{-16}$	-0.4	0.4 <sub>6.1</sub>	$3.5 \times 10^{-14}$	-0.4	0.2 <sub>4.7</sub>	$1.1 \times 10^{-16}$	-0.4	0.9 <sub>9.3</sub>	$2.5 \times 10^{-13}$	-0.4	8.8 <sub>28.6</sub>	9.2 <sub>28.9</sub>	8.4 <sub>27.8</sub>
Problem-Solving	4.7 <sub>21.2</sub>	$5.5 \times 10^{-10}$	-0.3	3.2 <sub>17.6</sub>	$3.4 \times 10^{-14}$	-0.4	5.9 <sub>23.7</sub>	$5.8 \times 10^{-7}$	-0.3	6.2 <sub>24.0</sub>	$1.4 \times 10^{-6}$	-0.3	14.0 <sub>32.7</sub>	9.4 <sub>29.2</sub>	18.6 <sub>39.0</sub>
Planning	2.3 <sub>15.0</sub>	$3.0 \times 10^{-4}$	-0.2	1.7 <sub>13.0</sub>	$1.6 \times 10^{-5}$	-0.3	5.3 <sub>22.4</sub>	$1.0 \times 10^0$	-0.1	4.1 <sub>19.8</sub>	$6.7 \times 10^{-1}$	-0.1	7.0 <sub>25.5</sub>	6.9 <sub>25.4</sub>	7.1 <sub>25.7</sub>
Normalizing	4.0 <sub>19.6</sub>	$3.5 \times 10^{-11}$	-0.4	3.2 <sub>17.6</sub>	$5.2 \times 10^{-13}$	-0.4	2.5 <sub>15.7</sub>	$7.7 \times 10^{-17}$	-0.4	2.6 <sub>15.9</sub>	$7.4 \times 10^{-17}$	-0.4	14.0 <sub>34.7</sub>	13.9 <sub>34.6</sub>	14.2 <sub>34.9</sub>
Psychoeducation	0.7 <sub>8.3</sub>	$3.5 \times 10^{-9}$	-0.3	0.2 <sub>5.0</sub>	$2.7 \times 10^{-11}$	-0.4	1.5 <sub>12.3</sub>	$2.7 \times 10^{-5}$	-0.3	0.5 <sub>7.3</sub>	$1.3 \times 10^{-10}$	-0.3	5.7 <sub>20.7</sub>	2.6 <sub>16.0</sub>	8.8 <sub>28.5</sub>
<b>Client behavior: Gained insights</b>															
<i>n</i>	561			382			490			425			258	214	44
Reflections on Needs	10.0 <sub>30.0</sub>	$2.4 \times 10^{-8}$	-0.5	7.3 <sub>26.1</sub>	$6.5 \times 10^{-10}$	-0.6	6.1 <sub>24.0</sub>	$5.3 \times 10^{-14}$	-0.6	7.8 <sub>26.8</sub>	$5.3 \times 10^{-10}$	-0.6	27.0 <sub>45.1</sub>	29.0 <sub>45.5</sub>	25.0 <sub>43.8</sub>
Reflections on Emotions	3.4 <sub>18.1</sub>	$1.1 \times 10^{-4}$	-0.4	2.9 <sub>16.7</sub>	$4.2 \times 10^{-4}$	-0.4	3.1 <sub>17.2</sub>	$9.8 \times 10^{-5}$	-0.4	3.5 <sub>18.5</sub>	$1.4 \times 10^{-3}$	-0.3	12.5 <sub>35.5</sub>	15.9 <sub>36.6</sub>	9.1 <sub>29.1</sub>
Reflections on Values	1.8 <sub>13.2</sub>	$1.4 \times 10^{-3}$	-0.3	0.3 <sub>5.1</sub>	$1.8 \times 10^{-5}$	-0.4	2.2 <sub>14.8</sub>	$1.9 \times 10^{-2}$	-0.3	1.4 <sub>11.8</sub>	$1.9 \times 10^{-3}$	-0.3	8.1 <sub>28.6</sub>	9.3 <sub>29.2</sub>	6.8 <sub>25.5</sub>
Reflections on Consequences	4.6 <sub>21.0</sub>	$5.1 \times 10^{-10}$	-0.5	2.4 <sub>15.2</sub>	$2.1 \times 10^{-12}$	-0.6	1.0 <sub>10.1</sub>	$3.0 \times 10^{-20}$	-0.8	1.4 <sub>11.8</sub>	$1.9 \times 10^{-16}$	-0.7	18.8 <sub>37.0</sub>	15.0 <sub>35.7</sub>	22.7 <sub>42.4</sub>
Reflections on Conflicts	0.7 <sub>8.4</sub>	$1.9 \times 10^{-1}$	-0.2	0.3 <sub>5.1</sub>	$1.3 \times 10^{-1}$	-0.3	0.8 <sub>9.0</sub>	$4.8 \times 10^{-1}$	-0.2	1.4 <sub>11.8</sub>	$1.0 \times 10^0$	-0.2	3.9 <sub>21.9</sub>	5.6 <sub>23.1</sub>	2.3 <sub>15.1</sub>
Reflections on Strengths	47.4 <sub>50.0</sub>	$7.8 \times 10^{-21}$	0.8	23.3 <sub>42.3</sub>	$6.1 \times 10^{-2}$	0.3	21.4 <sub>41.1</sub>	$2.5 \times 10^{-1}$	0.2	17.9 <sub>38.4</sub>	$1.0 \times 10^0$	0.1	12.5 <sub>35.5</sub>	15.9 <sub>36.6</sub>	9.1 <sub>29.1</sub>
Questions on Experiences	4.5 <sub>20.7</sub>	$2.8 \times 10^{-21}$	-0.8	0.8 <sub>8.8</sub>	$1.2 \times 10^{-24}$	-0.9	6.7 <sub>25.1</sub>	$2.7 \times 10^{-14}$	-0.6	6.6 <sub>24.8</sub>	$3.5 \times 10^{-13}$	-0.6	29.2 <sub>48.6</sub>	42.5 <sub>49.6</sub>	15.9 <sub>37.0</sub>
Questions on Perspectives	2.1 <sub>14.5</sub>	$1.0 \times 10^0$	-0.1	0.3 <sub>5.1</sub>	$9.5 \times 10^{-2}$	-0.3	2.0 <sub>14.2</sub>	$1.0 \times 10^0$	-0.1	2.6 <sub>15.9</sub>	$1.0 \times 10^0$	-0.1	4.2 <sub>22.7</sub>	6.1 <sub>23.9</sub>	2.3 <sub>15.1</sub>
Questions on Emotions	0.0 <sub>0.0</sub>	$2.0 \times 10^{-14}$	-0.6	0.3 <sub>5.1</sub>	$6.9 \times 10^{-9}$	-0.5	0.4 <sub>6.4</sub>	$9.4 \times 10^{-11}$	-0.6	0.0 <sub>0.0</sub>	$7.3 \times 10^{-11}$	-0.6	10.8 <sub>30.7</sub>	10.3 <sub>30.4</sub>	11.4 <sub>32.1</sub>
Problem-Solving	18.0 <sub>38.5</sub>	$5.5 \times 10^{-2}$	-0.3	17.3 <sub>37.9</sub>	$5.1 \times 10^{-2}$	-0.3	24.9 <sub>43.3</sub>	$1.0 \times 10^0$	-0.1	11.1 <sub>31.4</sub>	$9.9 \times 10^{-8}$	-0.5	27.9 <sub>39.6</sub>	15.0 <sub>35.7</sub>	40.9 <sub>49.7</sub>
Planning	25.8 <sub>43.8</sub>	$2.4 \times 10^{-4}$	0.4	16.2 <sub>36.9</sub>	$1.0 \times 10^0$	0.1	33.7 <sub>47.3</sub>	$2.4 \times 10^{-9}$	0.5	16.7 <sub>37.3</sub>	$1.0 \times 10^0$	0.1	11.5 <sub>32.1</sub>	11.7 <sub>32.2</sub>	11.4 <sub>32.1</sub>
Normalizing	32.4 <sub>46.9</sub>	$2.0 \times 10^{-3}$	0.3	15.7 <sub>36.4</sub>	$1.0 \times 10^0$	-0.1	10.8 <sub>31.1</sub>	$1.6 \times 10^{-1}$	-0.2	10.6 <sub>30.8</sub>	$1.5 \times 10^{-1}$	-0.2	18.6 <sub>35.1</sub>	12.1 <sub>32.7</sub>	25.0 <sub>43.8</sub>
Psychoeducation	5.3 <sub>22.5</sub>	$1.0 \times 10^0$	-0.1	5.2 <sub>22.3</sub>	$1.0 \times 10^0$	-0.2	13.1 <sub>33.7</sub>	$1.0 \times 10^0$	0.2	6.6 <sub>24.8</sub>	$1.0 \times 10^0$	-0.1	8.4 <sub>18.4</sub>	0.9 <sub>9.6</sub>	15.9 <sub>37.0</sub>

**Table C.17:** Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON CONSEQUENCES by 11.6% which is 15.7% less than average human therapists. (Single: 11.6%, Average human therapists: 27.3% from Table C.15;  $P = 3.2 \times 10^{-21}$ , Cohen’s  $d = -0.4$ , two-sided Student’s  $t$ -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)														
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b					
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$			
<b>Client behavior: Changing unhealthy behavior</b>																											
$n$	1553			999			1553			640			1553			796			1553			556					
Reflections on Needs	26.0439	$1.0 \times 10^0$	0.0	28.0449	$1.0 \times 10^0$	0.0	17.8383	$1.4 \times 10^{-5}$	-0.2	20.1403	$4.2 \times 10^{-2}$	-0.2	0.000	$6.8 \times 10^{-115}$	-0.9	0.000	$6.8 \times 10^{-115}$	-0.9	0.000	$6.8 \times 10^{-115}$	-0.9	0.000	$6.8 \times 10^{-115}$	-0.9	0.000	$6.8 \times 10^{-115}$	-0.9
Reflections on Emotions	6.9254	$3.6 \times 10^{-1}$	-0.1	9.7266	$1.0 \times 10^0$	0.0	6.3243	$1.5 \times 10^{-1}$	-0.1	13.8346	$1.0 \times 10^0$	0.1	0.000	$3.7 \times 10^{-32}$	-0.4	0.000	$3.7 \times 10^{-32}$	-0.4	0.000	$3.7 \times 10^{-32}$	-0.4	0.000	$3.7 \times 10^{-32}$	-0.4	0.000	$3.7 \times 10^{-32}$	-0.4
Reflections on Values	5.2222	$1.0 \times 10^0$	0.0	0.366	$2.1 \times 10^{-4}$	-0.2	2.9183	$1.0 \times 10^0$	-0.1	3.2127	$1.0 \times 10^0$	-0.1	0.000	$2.4 \times 10^{-13}$	-0.3	0.000	$2.4 \times 10^{-13}$	-0.3	0.000	$2.4 \times 10^{-13}$	-0.3	0.000	$2.4 \times 10^{-13}$	-0.3	0.000	$2.4 \times 10^{-13}$	-0.3
Reflections on Consequences	11.9321	$3.2 \times 10^{-21}$	-0.4	7.3261	$6.5 \times 10^{-26}$	-0.5	5.3224	$2.2 \times 10^{-39}$	-0.6	10.4006	$1.0 \times 10^{-15}$	-0.4	0.000	$8.9 \times 10^{-128}$	-0.9	0.000	$8.9 \times 10^{-128}$	-0.9	0.000	$8.9 \times 10^{-128}$	-0.9	0.000	$8.9 \times 10^{-128}$	-0.9	0.000	$8.9 \times 10^{-128}$	-0.9
Reflections on Conflicts	14.1184	$4.8 \times 10^{-8}$	-0.3	1.1104	$5.8 \times 10^{-6}$	-0.3	1.8132	$1.8 \times 10^{-5}$	-0.2	3.4182	$3.0 \times 10^{-1}$	-0.1	0.000	$1.5 \times 10^{-22}$	-0.4	0.000	$1.5 \times 10^{-22}$	-0.4	0.000	$1.5 \times 10^{-22}$	-0.4	0.000	$1.5 \times 10^{-22}$	-0.4	0.000	$1.5 \times 10^{-22}$	-0.4
Reflections on Strengths	80.7395	$0.0 \times 10^0$	2.0	68.9463	$1.0 \times 10^{-204}$	1.6	43.7496	$8.1 \times 10^{-82}$	0.9	45.7499	$6.3 \times 10^{-76}$	1.0	0.000	$3.1 \times 10^{-27}$	-0.4	0.000	$3.1 \times 10^{-27}$	-0.4	0.000	$3.1 \times 10^{-27}$	-0.4	0.000	$3.1 \times 10^{-27}$	-0.4	0.000	$3.1 \times 10^{-27}$	-0.4
Questions on Experiences	16.7373	$3.2 \times 10^{-23}$	-0.4	4.2201	$2.0 \times 10^{-51}$	-0.7	18.5388	$7.9 \times 10^{-16}$	-0.4	23.6255	$1.9 \times 10^{-5}$	-0.3	0.000	$8.4 \times 10^{-161}$	-1.0	0.000	$8.4 \times 10^{-161}$	-1.0	0.000	$8.4 \times 10^{-161}$	-1.0	0.000	$8.4 \times 10^{-161}$	-1.0	0.000	$8.4 \times 10^{-161}$	-1.0
Questions on Perspectives	2.8165	$1.0 \times 10^0$	0.0	0.240	$2.0 \times 10^{-3}$	-0.2	4.9216	$1.0 \times 10^0$	0.1	3.8191	$1.0 \times 10^0$	0.0	0.000	$4.0 \times 10^{-10}$	-0.2	0.000	$4.0 \times 10^{-10}$	-0.2	0.000	$4.0 \times 10^{-10}$	-0.2	0.000	$4.0 \times 10^{-10}$	-0.2	0.000	$4.0 \times 10^{-10}$	-0.2
Questions on Emotions	0.571	$5.6 \times 10^{-18}$	-0.4	0.568	$1.1 \times 10^{-11}$	-0.3	1.3111	$1.4 \times 10^{-11}$	-0.3	0.7485	$1.8 \times 10^{-9}$	-0.3	0.000	$3.8 \times 10^{-32}$	-0.4	0.000	$3.8 \times 10^{-32}$	-0.4	0.000	$3.8 \times 10^{-32}$	-0.4	0.000	$3.8 \times 10^{-32}$	-0.4	0.000	$3.8 \times 10^{-32}$	-0.4
Problem-Solving	35.7479	$4.0 \times 10^{-3}$	0.2	49.2500	$1.2 \times 10^{-20}$	0.5	59.9400	$1.5 \times 10^{-53}$	0.7	39.0488	$9.7 \times 10^{-5}$	0.2	0.000	$1.8 \times 10^{-134}$	-0.9	0.000	$1.8 \times 10^{-134}$	-0.9	0.000	$1.8 \times 10^{-134}$	-0.9	0.000	$1.8 \times 10^{-134}$	-0.9	0.000	$1.8 \times 10^{-134}$	-0.9
Planning	49.7500	$3.9 \times 10^{-97}$	0.9	52.5500	$2.5 \times 10^{-91}$	1.0	67.5469	$1.9 \times 10^{-189}$	1.4	45.3498	$1.5 \times 10^{-58}$	0.8	0.000	$5.7 \times 10^{-57}$	-0.6	0.000	$5.7 \times 10^{-57}$	-0.6	0.000	$5.7 \times 10^{-57}$	-0.6	0.000	$5.7 \times 10^{-57}$	-0.6	0.000	$5.7 \times 10^{-57}$	-0.6
Normalizing	49.3500	$2.3 \times 10^{-93}$	0.9	54.1499	$1.4 \times 10^{-96}$	1.0	19.2944	$5.0 \times 10^{-2}$	0.1	19.2392	$1.7 \times 10^{-1}$	0.2	0.000	$1.9 \times 10^{-59}$	-0.6	0.000	$1.9 \times 10^{-59}$	-0.6	0.000	$1.9 \times 10^{-59}$	-0.6	0.000	$1.9 \times 10^{-59}$	-0.6	0.000	$1.9 \times 10^{-59}$	-0.6
Psychoeducation	8.6281	$7.2 \times 10^{-2}$	0.1	10.2502	$1.7 \times 10^{-3}$	0.2	24.4430	$1.1 \times 10^{-43}$	0.6	15.6364	$8.5 \times 10^{-14}$	0.4	0.000	$3.8 \times 10^{-23}$	-0.4	0.000	$3.8 \times 10^{-23}$	-0.4	0.000	$3.8 \times 10^{-23}$	-0.4	0.000	$3.8 \times 10^{-23}$	-0.4	0.000	$3.8 \times 10^{-23}$	-0.4
<b>Client behavior: Sustaining unhealthy behavior</b>																											
$n$	389			39			389			13			389			96			389			73					
Reflections on Needs	59.0498	$4.1 \times 10^{-3}$	0.7	69.2480	$9.6 \times 10^{-2}$	0.9	35.4481	$1.0 \times 10^0$	0.2	38.4490	$1.0 \times 10^0$	0.2	0.000	$1.3 \times 10^{-28}$	-0.9	0.000	$1.3 \times 10^{-28}$	-0.9	0.000	$1.3 \times 10^{-28}$	-0.9	0.000	$1.3 \times 10^{-28}$	-0.9	0.000	$1.3 \times 10^{-28}$	-0.9
Reflections on Emotions	5.1223	$1.0 \times 10^0$	-0.2	7.7277	$1.0 \times 10^0$	-0.1	12.5332	$1.0 \times 10^0$	0.0	23.3426	$7.5 \times 10^{-1}$	0.3	0.000	$5.4 \times 10^{-10}$	-0.5	0.000	$5.4 \times 10^{-10}$	-0.5	0.000	$5.4 \times 10^{-10}$	-0.5	0.000	$5.4 \times 10^{-10}$	-0.5	0.000	$5.4 \times 10^{-10}$	-0.5
Reflections on Values	12.8339	$1.0 \times 10^0$	0.2	0.000	$1.0 \times 10^0$	-0.3	3.1175	$1.0 \times 10^0$	-0.2	5.5229	$1.0 \times 10^0$	-0.1	0.000	$9.1 \times 10^{-6}$	-0.4	0.000	$9.1 \times 10^{-6}$	-0.4	0.000	$9.1 \times 10^{-6}$	-0.4	0.000	$9.1 \times 10^{-6}$	-0.4	0.000	$9.1 \times 10^{-6}$	-0.4
Reflections on Consequences	51.3506	$1.0 \times 10^0$	0.3	30.8480	$1.0 \times 10^0$	-0.1	22.9423	$9.5 \times 10^{-1}$	-0.3	34.2478	$1.0 \times 10^0$	-0.0	0.000	$4.5 \times 10^{-44}$	-1.1	0.000	$4.5 \times 10^{-44}$	-1.1	0.000	$4.5 \times 10^{-44}$	-1.1	0.000	$4.5 \times 10^{-44}$	-1.1	0.000	$4.5 \times 10^{-44}$	-1.1
Reflections on Conflicts	10.3307	$1.0 \times 10^0$	0.1	38.5406	$1.2 \times 10^{-2}$	1.1	9.4293	$1.0 \times 10^0$	0.0	5.5229	$1.0 \times 10^0$	-0.1	0.000	$2.9 \times 10^{-7}$	-0.4	0.000	$2.9 \times 10^{-7}$	-0.4	0.000	$2.9 \times 10^{-7}$	-0.4	0.000	$2.9 \times 10^{-7}$	-0.4	0.000	$2.9 \times 10^{-7}$	-0.4
Reflections on Strengths	38.5493	$7.7 \times 10^{-11}$	1.2	38.5496	$2.3 \times 10^{-4}$	1.3	21.9416	$2.2 \times 10^{-5}$	0.6	23.3426	$2.8 \times 10^{-5}$	0.7	0.000	$1.2 \times 10^{-3}$	-0.3	0.000	$1.2 \times 10^{-3}$	-0.3	0.000	$1.2 \times 10^{-3}$	-0.3	0.000	$1.2 \times 10^{-3}$	-0.3	0.000	$1.2 \times 10^{-3}$	-0.3
Questions on Experiences	33.3478	$1.0 \times 10^0$	-0.1	15.4376	$1.0 \times 10^0$	-0.5	30.2462	$1.0 \times 10^0$	-0.2	34.2478	$1.0 \times 10^0$	-0.1	0.000	$1.8 \times 10^{-46}$	-1.1	0.000	$1.8 \times 10^{-46}$	-1.1	0.000	$1.8 \times 10^{-46}$	-1.1	0.000	$1.8 \times 10^{-46}$	-1.1	0.000	$1.8 \times 10^{-46}$	-1.1
Questions on Perspectives	7.7270	$1.0 \times 10^0$	0.2	0.000	$1.0 \times 10^0$	-0.2	7.3361	$1.0 \times 10^0$	0.2	11.0315	$5.6 \times 10^{-1}$	0.3	0.000	$2.9 \times 10^{-2}$	-0.3	0.000	$2.9 \times 10^{-2}$	-0.3	0.000	$2.9 \times 10^{-2}$	-0.3	0.000	$2.9 \times 10^{-2}$	-0.3	0.000	$2.9 \times 10^{-2}$	-0.3
Questions on Emotions	0.000	$1.0 \times 10^0$	-0.3	0.000	$1.0 \times 10^0$	-0.3	0.000	$1.0 \times 10^0$	-0.3	1.4117	$1.0 \times 10^0$	-0.3	0.000	$1.2 \times 10^{-7}$	-0.4	0.000	$1.2 \times 10^{-7}$	-0.4	0.000	$1.2 \times 10^{-7}$	-0.4	0.000	$1.2 \times 10^{-7}$	-0.4	0.000	$1.2 \times 10^{-7}$	-0.4
Problem-Solving	30.8468	$1.0 \times 10^0$	0.2	46.2519	$1.0 \times 10^0$	0.6	61.5489	$1.6 \times 10^{-12}$	0.9	43.8900	$1.4 \times 10^{-2}$	0.5	0.000	$9.8 \times 10^{-27}$	-0.8	0.000	$9.8 \times 10^{-27}$	-0.8	0.000	$9.8 \times 10^{-27}$	-0.8	0.000	$9.8 \times 10^{-27}$	-0.8	0.000	$9.8 \times 10^{-27}$	-0.8
Planning	20.4009	$3.3 \times 10^{-2}$	0.6	15.4376	$1.0 \times 10^0$	0.4	44.8800	$3.5 \times 10^{-25}$	1.3	26.0442	$6.3 \times 10^{-7}$	0.7	0.000	$1.1 \times 10^{-5}$	-0.4	0.000	$1.1 \times 10^{-5}$	-0.4	0.000	$1.1 \times 10^{-5}$	-0.4	0.000	$1.1 \times 10^{-5}$	-0.4	0.000	$1.1 \times 10^{-5}$	-0.4
Normalizing	23.1427	$1.0 \times 10^0$	0.3	23.1439	$1.0 \times 10^0$	0.3	11.5200	$1.0 \times 10^0$	-0.1	11.0315	$1.0 \times 10^0$	-0.1	0.000	$2.7 \times 10^{-14}$	-0.6	0.000	$2.7 \times 10^{-14}$	-0.6	0.000	$2.7 \times 10^{-14}$	-0.6	0.000	$2.7 \times 10^{-14}$	-0.6	0.000	$2.7 \times 10^{-14}$	-0.6
Psychoeducation	2.6160	$1.0 \times 10^0$	-0.3	15.4376	$1.0 \times 10^0$	0.2	16.7375	$1.0 \times 10^0$	0.2	9.6296	$1.0 \times 10^0$	0.0	0.000	$7.9 \times 10^{-11}$	-0.5	0.000	$7.9 \times 10^{-11}$	-0.5	0.000	$7.9 \times 10^{-11}$	-0.5	0.000	$7.9 \times 10^{-11}$	-0.5	0.000	$7.9 \times 10^{-11}$	-0.5
<b>Client behavior: Sharing positive emotions</b>																											
$n$	294			521			294			302			294			95			294			143					
Reflections on Needs	6.3244	$1.4 \times 10^{-4}$	-0.4	21.2409	$1.0 \times 10^0$	0.1	11.6222	$1.0 \times 10^0$	-0.1	9.1288	$1.0 \times 10^0$	-0.2	0.000	$9.3 \times 10^{-11}$	-0.6	0.000	$9.3 \times 10^{-11}$	-0.6	0.000	$9.3 \times 10^{-11}$	-0.6	0.000	$9.3 \times 10^{-11}$	-0.6	0.000	$9.3 \times 10^{-11}$	-0.6
Reflections on Emotions	3.6184	$1.0 \times 10^0$	-0.2	6.2149	$1.0 \times 10^0$	0.0	1.1103	$1.0 \times 10^0$	-0.2	2.8165	$1.0 \times 10^0$	-0.2	0.000	$5.0 \times 10^{-3}$	-0.3	0.000	$5.0 \times 10^{-3}$	-0.3	0.000	$5.0 \times 10^{-3}$	-0.3	0.000	$5.0 \times 10^{-3}$	-0.3	0.000	$5.0 \times 10^{-3}$	-0.3
Reflections on Values	1.3115	$1.0 \times 10^0$	-0.1	0.366	$1.0 \times 10^0$	-0.2	3.2766	$1.0 \times 10^0$	0.0	0.744	$1.0 \times 10^0$	-0.2	0.000	$3.8 \times 10^{-1}$	-0.2	0.000	$3.8 \times 10^{-1}$	-0.2									

**Table C.18:** (Continue Table C.17) Adaptability (frequency of conversational behaviors exhibited by therapists *in response to specific client behaviors*) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) with two simulation methods (Single response, Full conversation) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13 \times 6 = 78$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, when client expresses CHANGING UNHEALTHY BEHAVIOR, GPT-4 with single response simulation exhibits REFLECTIONS ON NEEDS by 46.5% which is 22.0% more than average human therapists. (Single: 46.5%, Average human therapists: 24.5% from Table C.16;  $P = 1.1 \times 10^{-11}$ , Cohen’s  $d = 0.5$ , two-sided Student’s  $t$ -test)

Therapist Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4			GPT-3.5			Llama2-70b			Llama2-13b			GPT-4			GPT-3.5			Llama2-70b			Llama2-13b		
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$
<b>Client behavior: Sharing negative emotions</b>																								
<i>n</i>	820			355			820			232			820			534			820			528		
Reflections on Needs	46.5 <sub>99</sub>	$1.1 \times 10^{-11}$	0.5	32.8 <sub>70</sub>	$1.0 \times 10^0$	0.2	24.2 <sub>42</sub>	$1.0 \times 10^0$	0.0	23.5 <sub>42</sub>	$1.0 \times 10^0$	0.0	0.0 <sub>00</sub>	$1.5 \times 10^{-50}$	-0.8	0.0 <sub>00</sub>	$1.5 \times 10^{-50}$	-0.8	0.0 <sub>00</sub>	$1.5 \times 10^{-50}$	-0.8	0.0 <sub>00</sub>	$1.5 \times 10^{-50}$	-0.8
Reflections on Emotions	29.9 <sub>458</sub>	$1.4 \times 10^{-2}$	0.2	32.8 <sub>70</sub>	$3.6 \times 10^{-3}$	0.3	20.6 <sub>105</sub>	$1.0 \times 10^0$	0.0	31.2 <sub>464</sub>	$1.5 \times 10^{-4}$	0.3	0.0 <sub>00</sub>	$4.3 \times 10^{-35}$	-0.6	0.0 <sub>00</sub>	$4.3 \times 10^{-35}$	-0.6	0.0 <sub>00</sub>	$4.3 \times 10^{-35}$	-0.6	0.0 <sub>00</sub>	$4.3 \times 10^{-35}$	-0.6
Reflections on Values	8.2 <sub>74</sub>	$1.0 \times 10^0$	0.0	0.9 <sub>93</sub>	$1.1 \times 10^{-1}$	-0.2	2.1 <sub>142</sub>	$3.6 \times 10^{-2}$	-0.2	3.2 <sub>177</sub>	$1.0 \times 10^0$	-0.1	0.0 <sub>00</sub>	$1.5 \times 10^{-10}$	-0.4	0.0 <sub>00</sub>	$1.5 \times 10^{-10}$	-0.4	0.0 <sub>00</sub>	$1.5 \times 10^{-10}$	-0.4	0.0 <sub>00</sub>	$1.5 \times 10^{-10}$	-0.4
Reflections on Consequences	23.7 <sub>426</sub>	$1.0 \times 10^0$	0.0	16.4 <sub>371</sub>	$1.0 \times 10^0$	-0.1	8.8 <sub>284</sub>	$7.4 \times 10^{-9}$	-0.4	11.2 <sub>15</sub>	$2.2 \times 10^{-5}$	-0.3	0.0 <sub>00</sub>	$1.1 \times 10^{-49}$	-0.8	0.0 <sub>00</sub>	$1.1 \times 10^{-49}$	-0.8	0.0 <sub>00</sub>	$1.1 \times 10^{-49}$	-0.8	0.0 <sub>00</sub>	$1.1 \times 10^{-49}$	-0.8
Reflections on Conflicts	4.5 <sub>208</sub>	$1.0 \times 10^0$	-0.1	2.6 <sub>159</sub>	$3.0 \times 10^{-1}$	-0.2	2.8 <sub>165</sub>	$5.1 \times 10^{-3}$	-0.2	5.9 <sub>235</sub>	$1.0 \times 10^0$	-0.1	0.0 <sub>00</sub>	$4.7 \times 10^{-14}$	-0.4	0.0 <sub>00</sub>	$4.7 \times 10^{-14}$	-0.4	0.0 <sub>00</sub>	$4.7 \times 10^{-14}$	-0.4	0.0 <sub>00</sub>	$4.7 \times 10^{-14}$	-0.4
Reflections on Strengths	44.8 <sub>498</sub>	$1.4 \times 10^{-56}$	1.1	31.9 <sub>67</sub>	$4.1 \times 10^{-23}$	0.8	31.8 <sub>66</sub>	$5.9 \times 10^{-33}$	0.7	26.5 <sub>442</sub>	$2.8 \times 10^{-22}$	0.6	0.0 <sub>00</sub>	$2.1 \times 10^{-10}$	-0.3	0.0 <sub>00</sub>	$2.1 \times 10^{-10}$	-0.3	0.0 <sub>00</sub>	$2.1 \times 10^{-10}$	-0.3	0.0 <sub>00</sub>	$2.1 \times 10^{-10}$	-0.3
Questions on Experiences	38.3 <sub>487</sub>	$1.0 \times 10^0$	0.1	15.9 <sub>67</sub>	$5.2 \times 10^{-6}$	-0.4	33.9 <sub>474</sub>	$1.0 \times 10^0$	0.0	40.2 <sub>491</sub>	$1.0 \times 10^0$	0.1	0.0 <sub>00</sub>	$7.1 \times 10^{-81}$	-1.0	0.0 <sub>00</sub>	$7.1 \times 10^{-81}$	-1.0	0.0 <sub>00</sub>	$7.1 \times 10^{-81}$	-1.0	0.0 <sub>00</sub>	$7.1 \times 10^{-81}$	-1.0
Questions on Perspectives	3.7 <sub>188</sub>	$1.0 \times 10^0$	0.0	0.9 <sub>93</sub>	$1.0 \times 10^0$	-0.1	6.7 <sub>251</sub>	$3.7 \times 10^{-1}$	-0.2	5.5 <sub>228</sub>	$1.0 \times 10^0$	0.1	0.0 <sub>00</sub>	$2.4 \times 10^{-4}$	-0.2	0.0 <sub>00</sub>	$2.4 \times 10^{-4}$	-0.2	0.0 <sub>00</sub>	$2.4 \times 10^{-4}$	-0.2	0.0 <sub>00</sub>	$2.4 \times 10^{-4}$	-0.2
Questions on Emotions	0.6 <sub>75</sub>	$5.9 \times 10^{-6}$	-0.3	1.7 <sub>130</sub>	$1.3 \times 10^{-2}$	-0.3	1.3 <sub>114</sub>	$3.3 \times 10^{-7}$	-0.3	0.9 <sub>97</sub>	$5.5 \times 10^{-8}$	-0.3	0.0 <sub>00</sub>	$1.7 \times 10^{-16}$	-0.4	0.0 <sub>00</sub>	$1.7 \times 10^{-16}$	-0.4	0.0 <sub>00</sub>	$1.7 \times 10^{-16}$	-0.4	0.0 <sub>00</sub>	$1.7 \times 10^{-16}$	-0.4
Problem-Solving	36.9 <sub>483</sub>	$2.1 \times 10^{-10}$	0.4	64.2 <sub>480</sub>	$3.9 \times 10^{-48}$	1.2	57.3 <sub>495</sub>	$3.8 \times 10^{-54}$	0.9	39.2 <sub>489</sub>	$4.3 \times 10^{-16}$	0.5	0.0 <sub>00</sub>	$4.9 \times 10^{-47}$	-0.8	0.0 <sub>00</sub>	$4.9 \times 10^{-47}$	-0.8	0.0 <sub>00</sub>	$4.9 \times 10^{-47}$	-0.8	0.0 <sub>00</sub>	$4.9 \times 10^{-47}$	-0.8
Planning	17.5 <sub>80</sub>	$1.4 \times 10^{-5}$	0.3	27.2 <sub>46</sub>	$1.7 \times 10^{-15}$	0.6	44.6 <sub>698</sub>	$1.6 \times 10^{-63}$	1.0	26.5 <sub>442</sub>	$6.8 \times 10^{-21}$	0.6	0.0 <sub>00</sub>	$5.5 \times 10^{-16}$	-0.4	0.0 <sub>00</sub>	$5.5 \times 10^{-16}$	-0.4	0.0 <sub>00</sub>	$5.5 \times 10^{-16}$	-0.4	0.0 <sub>00</sub>	$5.5 \times 10^{-16}$	-0.4
Normalizing	32.4 <sub>469</sub>	$3.7 \times 10^{-8}$	0.4	47.8 <sub>51</sub>	$3.7 \times 10^{-23}$	0.8	17.0 <sub>176</sub>	$1.0 \times 10^0$	0.0	17.0 <sub>176</sub>	$1.0 \times 10^0$	0.0	0.0 <sub>00</sub>	$2.0 \times 10^{-35}$	-0.6	0.0 <sub>00</sub>	$2.0 \times 10^{-35}$	-0.6	0.0 <sub>00</sub>	$2.0 \times 10^{-35}$	-0.6	0.0 <sub>00</sub>	$2.0 \times 10^{-35}$	-0.6
Psychoeducation	8.5 <sub>279</sub>	$4.6 \times 10^{-1}$	0.2	13.4 <sub>541</sub>	$1.8 \times 10^{-5}$	0.4	19.3 <sub>395</sub>	$7.1 \times 10^{-18}$	0.5	11.0 <sub>313</sub>	$2.4 \times 10^{-4}$	0.3	0.0 <sub>00</sub>	$1.3 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$1.3 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$1.3 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$1.3 \times 10^{-12}$	-0.4
<b>Client behavior: Sharing experiences</b>																								
<i>n</i>	759			115			759			53			759			149			759			167		
Reflections on Needs	49.6 <sub>502</sub>	$8.4 \times 10^{-8}$	0.6	20.8 <sub>409</sub>	$1.0 \times 10^0$	0.0	23.5 <sub>425</sub>	$1.0 \times 10^0$	0.0	27.5 <sub>448</sub>	$1.0 \times 10^0$	0.1	0.0 <sub>00</sub>	$8.7 \times 10^{-43}$	-0.7	0.0 <sub>00</sub>	$8.7 \times 10^{-43}$	-0.7	0.0 <sub>00</sub>	$8.7 \times 10^{-43}$	-0.7	0.0 <sub>00</sub>	$8.7 \times 10^{-43}$	-0.7
Reflections on Emotions	44.3 <sub>499</sub>	$4.0 \times 10^{-13}$	0.8	45.3 <sub>503</sub>	$3.6 \times 10^{-77}$	0.8	26.8 <sub>445</sub>	$7.7 \times 10^{-3}$	0.4	33.5 <sub>474</sub>	$1.8 \times 10^{-7}$	0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-21}$	-0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-21}$	-0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-21}$	-0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-21}$	-0.5
Reflections on Values	4.3 <sub>205</sub>	$1.0 \times 10^0$	0.0	0.0 <sub>00</sub>	$1.0 \times 10^0$	-0.2	4.7 <sub>212</sub>	$1.0 \times 10^0$	0.0	7.2 <sub>259</sub>	$1.0 \times 10^0$	0.1	0.0 <sub>00</sub>	$2.4 \times 10^{-6}$	-0.3	0.0 <sub>00</sub>	$2.4 \times 10^{-6}$	-0.3	0.0 <sub>00</sub>	$2.4 \times 10^{-6}$	-0.3	0.0 <sub>00</sub>	$2.4 \times 10^{-6}$	-0.3
Reflections on Consequences	18.3 <sub>388</sub>	$1.0 \times 10^0$	0.0	15.1 <sub>361</sub>	$1.0 \times 10^0$	-0.1	11.4 <sub>319</sub>	$1.0 \times 10^0$	-0.2	9.6 <sub>295</sub>	$5.1 \times 10^{-1}$	-0.2	0.0 <sub>00</sub>	$5.6 \times 10^{-35}$	-0.7	0.0 <sub>00</sub>	$5.6 \times 10^{-35}$	-0.7	0.0 <sub>00</sub>	$5.6 \times 10^{-35}$	-0.7	0.0 <sub>00</sub>	$5.6 \times 10^{-35}$	-0.7
Reflections on Conflicts	7.8 <sub>270</sub>	$1.0 \times 10^0$	0.1	1.9 <sub>137</sub>	$1.0 \times 10^0$	-0.2	5.4 <sub>226</sub>	$1.0 \times 10^0$	0.0	6.0 <sub>238</sub>	$1.0 \times 10^0$	0.0	0.0 <sub>00</sub>	$1.8 \times 10^{-8}$	-0.3	0.0 <sub>00</sub>	$1.8 \times 10^{-8}$	-0.3	0.0 <sub>00</sub>	$1.8 \times 10^{-8}$	-0.3	0.0 <sub>00</sub>	$1.8 \times 10^{-8}$	-0.3
Reflections on Strengths	33.9 <sub>475</sub>	$1.6 \times 10^{-12}$	0.8	28.3 <sub>455</sub>	$6.2 \times 10^{-4}$	0.6	24.8 <sub>433</sub>	$1.8 \times 10^{-6}$	0.5	20.4 <sub>404</sub>	$1.1 \times 10^{-3}$	0.4	0.0 <sub>00</sub>	$1.0 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$1.0 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$1.0 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$1.0 \times 10^{-12}$	-0.4
Questions on Experiences	44.3 <sub>499</sub>	$1.0 \times 10^0$	0.0	35.8 <sub>484</sub>	$1.0 \times 10^0$	-0.2	52.3 <sub>501</sub>	$1.0 \times 10^0$	0.2	50.3 <sub>501</sub>	$1.0 \times 10^0$	0.1	0.0 <sub>00</sub>	$4.9 \times 10^{-112}$	-1.3	0.0 <sub>00</sub>	$4.9 \times 10^{-112}$	-1.3	0.0 <sub>00</sub>	$4.9 \times 10^{-112}$	-1.3	0.0 <sub>00</sub>	$4.9 \times 10^{-112}$	-1.3
Questions on Perspectives	0.9 <sub>93</sub>	$1.0 \times 10^0$	-0.1	0.0 <sub>00</sub>	$1.0 \times 10^0$	-0.1	6.7 <sub>251</sub>	$1.4 \times 10^{-1}$	-0.3	5.4 <sub>226</sub>	$1.0 \times 10^0$	0.2	0.0 <sub>00</sub>	$2.7 \times 10^{-2}$	-0.2	0.0 <sub>00</sub>	$2.7 \times 10^{-2}$	-0.2	0.0 <sub>00</sub>	$2.7 \times 10^{-2}$	-0.2	0.0 <sub>00</sub>	$2.7 \times 10^{-2}$	-0.2
Questions on Emotions	0.9 <sub>93</sub>	$2.5 \times 10^{-1}$	-0.3	5.7 <sub>233</sub>	$1.0 \times 10^0$	-0.1	1.3 <sub>115</sub>	$1.4 \times 10^{-1}$	-0.3	4.8 <sub>214</sub>	$1.0 \times 10^0$	-0.1	0.0 <sub>00</sub>	$4.0 \times 10^{-15}$	-0.4	0.0 <sub>00</sub>	$4.0 \times 10^{-15}$	-0.4	0.0 <sub>00</sub>	$4.0 \times 10^{-15}$	-0.4	0.0 <sub>00</sub>	$4.0 \times 10^{-15}$	-0.4
Problem-Solving	35.7 <sub>481</sub>	$7.9 \times 10^{-8}$	0.6	49.1 <sub>505</sub>	$8.0 \times 10^{-11}$	1.0	36.2 <sub>482</sub>	$4.9 \times 10^{-10}$	0.6	34.1 <sub>476</sub>	$5.9 \times 10^{-9}$	0.6	0.0 <sub>00</sub>	$6.8 \times 10^{-29}$	-0.6	0.0 <sub>00</sub>	$6.8 \times 10^{-29}$	-0.6	0.0 <sub>00</sub>	$6.8 \times 10^{-29}$	-0.6	0.0 <sub>00</sub>	$6.8 \times 10^{-29}$	-0.6
Planning	17.4 <sub>381</sub>	$1.3 \times 10^{-2}$	0.4	26.4 <sub>445</sub>	$4.6 \times 10^{-5}$	0.7	32.2 <sub>469</sub>	$4.8 \times 10^{-18}$	0.8	22.8 <sub>421</sub>	$3.3 \times 10^{-8}$	0.5	0.0 <sub>00</sub>	$4.9 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$4.9 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$4.9 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$4.9 \times 10^{-12}$	-0.4
Normalizing	30.4 <sub>462</sub>	$5.9 \times 10^{-4}$	0.5	49.1 <sub>505</sub>	$1.0 \times 10^{-9}$	1.0	15.4 <sub>363</sub>	$1.0 \times 10^0$	0.0	14.4 <sub>352</sub>	$1.0 \times 10^0$	0.0	0.0 <sub>00</sub>	$7.5 \times 10^{-26}$	-0.6	0.0 <sub>00</sub>	$7.5 \times 10^{-26}$	-0.6	0.0 <sub>00</sub>	$7.5 \times 10^{-26}$	-0.6	0.0 <sub>00</sub>	$7.5 \times 10^{-26}$	-0.6
Psychoeducation	5.2 <sub>223</sub>	$1.0 \times 10^0$	0.0	3.8 <sub>192</sub>	$1.0 \times 10^0$	-0.1	9.4 <sub>293</sub>	$1.0 \times 10^0$	0.2	3.0 <sub>171</sub>	$1.0 \times 10^0$	-0.1	0.0 <sub>00</sub>	$3.0 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$3.0 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$3.0 \times 10^{-12}$	-0.4	0.0 <sub>00</sub>	$3.0 \times 10^{-12}$	-0.4
<b>Client behavior: Gained insights</b>																								
<i>n</i>	258			303			258			124			258			232			258			167		
Reflections on Needs	18.5 <sub>389</sub>	$1.0 \times 10^0$	-0.2	22.6 <sub>420</sub>	$1.0 \times 10^0$	-0.1	12.9 <sub>336</sub>	$9.5 \times 10^{-3}$	-0.4	19.8 <sub>399</sub>	$1.0 \times 10^0$	-0.2	0.0 <sub>00</sub>	$2.6 \times 10^{-18}$	-0.8	0.0 <sub>00</sub>	$2.6 \times 10^{-18}$	-0.8	0.0 <sub>00</sub>	$2.6 \times 10^{-18}$	-0.8	0.0 <sub>00</sub>	$2.6 \times 10^{-18}$	-0.8
Reflections on Emotions	6.3 <sub>243</sub>	$1.0 \times 10^0$	-0.2	8.9 <sub>285</sub>	$1.0 \times 10^0$	-0.1	6.5 <sub>216</sub>	$1.0 \times 10^0$	-0.2	9.0 <sub>287</sub>	$1.0 \times 10^0$	-0.1	0.0 <sub>00</sub>	$2.1 \times 10^{-6}$	-0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-6}$	-0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-6}$	-0.5	0.0 <sub>00</sub>	$2.1 \times 10^{-6}$	-0.5
Reflections on Values	3.3 <sub>179</sub>	$1.0 \times 10^0$	-0.2	0.8 <sub>90</sub>	$4.6 \times 10^{-1}$	-0.3	4.7 <sub>213</sub>	$1.0 \times 10^0$	-0.1	3.6 <sub>187</sub>	$1.0 \times 10^0$	-0.2	0.0 <sub>00</sub>	$5.3 \times 10^{-4}$	-0.4	0.0 <sub>00</sub>	$5.3 \times 10^{-4}$	-0.4	0.0 <sub>00</sub>					

**Table C.19:** Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 14$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 exhibits POSITIVE EMOTIONS by 1.5% while human therapists responds by 1.0% on average ( $P = 2.8 \times 10^{-19}$ , Cohen’s  $d = 0.2$ , two-sided Student’s  $t$ -test).

Linguistic Attributes	LLM Therapists (Single Response & Full Conversation Simulations)												Human Therapists		
	GPT-4 ( $n = 7705$ )			GPT-3.5 ( $n = 7135$ )			Llama2-70b ( $n = 7461$ )			Llama2-13b ( $n = 7464$ )			Average ( $n = 5446$ )	High-quality ( $n = 3907$ )	Low-quality ( $n = 1539$ )
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	freq.	freq.
Positive Emotions	1.5 <sub>2,1</sub>	$2.8 \times 10^{-19}$	0.2	1.1 <sub>1,8</sub>	$3.3 \times 10^{-2}$	0.1	1.0 <sub>2,1</sub>	$1.0 \times 10^0$	0.0	1.0 <sub>2,5</sub>	$1.0 \times 10^0$	0.0	1.0 <sub>4,2</sub>	1.2 <sub>4,3</sub>	0.7 <sub>3,6</sub>
Negative Emotions	1.1 <sub>2,0</sub>	$1.3 \times 10^{-16}$	0.2	1.2 <sub>2,0</sub>	$1.1 \times 10^{-26}$	0.2	1.2 <sub>2,5</sub>	$6.1 \times 10^{-20}$	0.2	1.1 <sub>2,5</sub>	$2.3 \times 10^{-10}$	0.1	0.7 <sub>3,6</sub>	0.7 <sub>3,4</sub>	0.6 <sub>4,3</sub>
BigWords	24.6 <sub>7,2</sub>	$0.0 \times 10^0$	1.6	20.9 <sub>7,7</sub>	$0.0 \times 10^0$	1.2	23.7 <sub>11,3</sub>	$0.0 \times 10^0$	1.2	21.0 <sub>12,0</sub>	$0.0 \times 10^0$	0.9	10.7 <sub>9,9</sub>	11.0 <sub>9,9</sub>	10.4 <sub>10,1</sub>
Self	1.5 <sub>1,9</sub>	$6.3 \times 10^{-139}$	-0.5	1.8 <sub>2,3</sub>	$1.9 \times 10^{-84}$	-0.4	1.4 <sub>2,4</sub>	$2.9 \times 10^{-146}$	-0.5	1.4 <sub>2,4</sub>	$7.1 \times 10^{-144}$	-0.5	3.1 <sub>5,1</sub>	2.5 <sub>4,5</sub>	3.8 <sub>6,1</sub>
Other	8.9 <sub>3,4</sub>	$1.0 \times 10^0$	0.0	8.4 <sub>3,6</sub>	$3.2 \times 10^{-7}$	-0.1	7.5 <sub>4,5</sub>	$6.2 \times 10^{-41}$	-0.2	7.0 <sub>5,0</sub>	$2.4 \times 10^{-70}$	-0.3	8.9 <sub>7,1</sub>	8.7 <sub>6,9</sub>	9.2 <sub>7,4</sub>
Health	2.0 <sub>2,6</sub>	$9.4 \times 10^{-2}$	0.0	2.1 <sub>3,0</sub>	$6.0 \times 10^{-5}$	0.1	1.5 <sub>3,1</sub>	$2.1 \times 10^{-8}$	-0.1	1.3 <sub>3,0</sub>	$5.1 \times 10^{-16}$	-0.2	1.8 <sub>4,3</sub>	1.2 <sub>3,2</sub>	2.5 <sub>6,1</sub>
Wellness	0.7 <sub>1,4</sub>	$3.0 \times 10^{-41}$	0.2	0.6 <sub>1,5</sub>	$1.0 \times 10^{-13}$	0.1	0.3 <sub>1,0</sub>	$1.0 \times 10^0$	0.0	0.3 <sub>1,1</sub>	$1.0 \times 10^0$	0.0	0.3 <sub>2,1</sub>	0.2 <sub>1,3</sub>	0.4 <sub>3,5</sub>
Behavioral Activation	6.0 <sub>4,0</sub>	$7.7 \times 10^{-225}$	0.6	5.7 <sub>4,4</sub>	$2.0 \times 10^{-163}$	0.5	4.8 <sub>4,5</sub>	$6.6 \times 10^{-70}$	0.3	4.5 <sub>5,3</sub>	$1.1 \times 10^{-38}$	0.2	3.2 <sub>5,9</sub>	3.5 <sub>5,9</sub>	2.9 <sub>5,9</sub>
Meaning	3.6 <sub>2,6</sub>	$0.0 \times 10^0$	0.7	3.2 <sub>2,6</sub>	$2.2 \times 10^{-232}$	0.6	2.2 <sub>2,8</sub>	$5.5 \times 10^{-48}$	0.3	2.0 <sub>2,7</sub>	$1.1 \times 10^{-27}$	0.2	1.4 <sub>3,5</sub>	1.4 <sub>3,4</sub>	1.4 <sub>3,7</sub>
Purpose	3.9 <sub>2,7</sub>	$0.0 \times 10^0$	0.8	3.6 <sub>2,9</sub>	$1.5 \times 10^{-275}$	0.7	2.4 <sub>2,9</sub>	$6.3 \times 10^{-50}$	0.3	2.2 <sub>3,2</sub>	$4.9 \times 10^{-28}$	0.2	1.5 <sub>3,6</sub>	1.5 <sub>3,5</sub>	1.5 <sub>3,8</sub>
Motivation	3.4 <sub>2,5</sub>	$0.0 \times 10^0$	0.7	3.3 <sub>2,7</sub>	$6.9 \times 10^{-272}$	0.7	2.0 <sub>2,5</sub>	$1.2 \times 10^{-38}$	0.2	1.9 <sub>2,9</sub>	$3.1 \times 10^{-23}$	0.2	1.3 <sub>3,3</sub>	1.3 <sub>3,3</sub>	1.3 <sub>3,4</sub>
Sadness	1.0 <sub>1,6</sub>	$1.2 \times 10^{-126}$	0.4	1.1 <sub>1,8</sub>	$1.5 \times 10^{-134}$	0.5	0.8 <sub>1,9</sub>	$2.9 \times 10^{-69}$	0.3	0.9 <sub>2,1</sub>	$8.0 \times 10^{-65}$	0.3	0.3 <sub>1,5</sub>	0.3 <sub>1,5</sub>	0.2 <sub>1,6</sub>
Sympathy	0.3 <sub>0,8</sub>	$1.5 \times 10^{-29}$	0.2	0.4 <sub>1,2</sub>	$9.2 \times 10^{-36}$	0.2	0.3 <sub>1,0</sub>	$8.0 \times 10^{-25}$	0.2	0.4 <sub>1,4</sub>	$2.8 \times 10^{-29}$	0.2	0.1 <sub>1,0</sub>	0.2 <sub>1,1</sub>	0.1 <sub>0,8</sub>
Stress	1.5 <sub>2,2</sub>	$3.6 \times 10^{-16}$	0.2	1.7 <sub>2,3</sub>	$1.4 \times 10^{-30}$	0.2	1.8 <sub>2,9</sub>	$6.4 \times 10^{-33}$	0.2	1.6 <sub>2,9</sub>	$4.6 \times 10^{-18}$	0.2	1.1 <sub>3,5</sub>	1.1 <sub>3,2</sub>	1.1 <sub>4,2</sub>

**Table C.20:** Frequency of linguistic attributes (from LIWC analysis) by LLM therapists (GPT-4, GPT-3.5-turbo, Llama2-70b, Llama2-13b) in HOPE dataset (Malhotra et al., 2022) for two different types of simulations separately (Single response and Full conversation). Values in gray are not statistically significantly different from human therapist average at  $p = \frac{0.05}{m}$  using Two-sided Student’s t-test after Bonferroni correction, where  $m$  is the number of t-test per one model ( $m = 13$ ).  $d$  is the cohen’s  $d$  when compared with human therapist average. For instance, GPT-4 with single response simulation responds with POSITIVE EMOTIONS by 2.1% and GPT-4 with full conversation simulation responds by 1.3% when compared with the human therapists responds by 1.0% on average from Table C.19. (Single:  $P = 2.8 \times 10^{-24}$ , Cohen’s  $d = 0.3$ ; two-sided student t-test) (Full:  $P = 4.7 \times 10^{-8}$ , Cohen’s  $d = 0.1$ ; two-sided student t-test).

Behavior	LLM Therapists (Single Response Simulations)												LLM Therapists (Full Conversation Simulations)											
	GPT-4 ( $n = 6158$ )			GPT-3.5 ( $n = 6144$ )			Llama2-70b ( $n = 6174$ )			Llama2-13b ( $n = 6282$ )			GPT-4 ( $n = 1547$ )			GPT-3.5 ( $n = 991$ )			Llama2-70b ( $n = 1287$ )			Llama2-13b ( $n = 1182$ )		
	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$	freq.	$p$	$d$
Positive Emotions	2.1 <sub>1,7</sub>	$2.8 \times 10^{-24}$	0.3	1.7 <sub>1,6</sub>	$1.3 \times 10^{-7}$	0.2	1.3 <sub>1,5</sub>	$3.3 \times 10^{-2}$	0.1	1.7 <sub>2,2</sub>	$6.2 \times 10^{-8}$	0.2	1.3 <sub>2,2</sub>	$4.7 \times 10^{-8}$	0.1	1.0 <sub>1,8</sub>	$1.0 \times 10^0$	0.0	0.9 <sub>2,2</sub>	$1.0 \times 10^0$	0.0	0.9 <sub>2,5</sub>	$1.0 \times 10^0$	0.0
Negative Emotions	0.5 <sub>1,0</sub>	$1.0 \times 10^0$	-0.1	0.6 <sub>1,1</sub>	$1.0 \times 10^0$	0.0	0.7 <sub>1,1</sub>	$1.0 \times 10^0$	0.0	0.7 <sub>1,2</sub>	$1.0 \times 10^0$	0.0	1.3 <sub>2,1</sub>	$1.6 \times 10^{-24}$	0.2	1.3 <sub>2,7</sub>	$2.1 \times 10^{-23}$	0.2	1.3 <sub>2,7</sub>	$2.1 \times 10^{-23}$	0.2	1.1 <sub>2,6</sub>	$1.4 \times 10^{-12}$	0.1
BigWords	28.3 <sub>5,0</sub>	$0.0 \times 10^0$	1.9	25.6 <sub>6,7</sub>	$0.0 \times 10^0$	1.6	31.8 <sub>10,2</sub>	$0.0 \times 10^0$	2.1	31.2 <sub>6,0</sub>	$0.0 \times 10^0$	1.8	23.7 <sub>7,4</sub>	$0.0 \times 10^0$	1.5	20.2 <sub>7,6</sub>	$0.0 \times 10^0$	1.1	22.1 <sub>10,8</sub>	$0.0 \times 10^0$	1.1	19.1 <sub>10,1</sub>	$0.0 \times 10^0$	0.8
Self	2.3 <sub>1,9</sub>	$4.0 \times 10^{-8}$	-0.2	2.7 <sub>2,4</sub>	$9.0 \times 10^{-2}$	-0.1	1.5 <sub>2,3</sub>	$6.2 \times 10^{-28}$	-0.3	1.5 <sub>1,9</sub>	$1.5 \times 10^{-25}$	-0.3	1.3 <sub>1,8</sub>	$8.7 \times 10^{-147}$	-0.5	1.6 <sub>2,3</sub>	$3.0 \times 10^{-92}$	-0.4	1.3 <sub>2,4</sub>	$1.5 \times 10^{-129}$	-0.5	1.3 <sub>2,5</sub>	$8.6 \times 10^{-128}$	-0.5
Other	9.5 <sub>2,3</sub>	$3.5 \times 10^{-2}$	0.1	8.1 <sub>2,8</sub>	$3.1 \times 10^{-3}$	-0.1	5.9 <sub>2,3</sub>	$9.4 \times 10^{-50}$	-0.5	5.6 <sub>5,4</sub>	$1.4 \times 10^{-49}$	-0.5	8.8 <sub>3,6</sub>	$1.0 \times 10^0$	0.0	8.4 <sub>3,7</sub>	$2.5 \times 10^{-5}$	0.1	7.9 <sub>4,6</sub>	$3.0 \times 10^{-20}$	-0.2	7.3 <sub>4,9</sub>	$8.0 \times 10^{-49}$	-0.3
Health	2.3 <sub>2,1</sub>	$1.3 \times 10^{-4}$	0.1	2.7 <sub>2,5</sub>	$6.6 \times 10^{-9}$	0.2	1.7 <sub>1,9</sub>	$1.0 \times 10^0$	0.0	1.6 <sub>3,3</sub>	$1.0 \times 10^0$	0.0	1.9 <sub>2,7</sub>	$1.0 \times 10^0$	0.0	2.1 <sub>3,0</sub>	$4.0 \times 10^{-2}$	0.1	1.4 <sub>3,3</sub>	$7.8 \times 10^{-9}$	-0.1	1.2 <sub>3,0</sub>	$5.1 \times 10^{-18}$	-0.2
Wellness	1.1 <sub>1,3</sub>	$3.3 \times 10^{-4}$	0.4	1.2 <sub>1,5</sub>	$1.2 \times 10^{-31}$	0.4	0.6 <sub>1,1</sub>	$1.2 \times 10^{-6}$	0.2	0.6 <sub>1,4</sub>	$5.7 \times 10^{-5}$	0.1	0.6 <sub>1,4</sub>	$1.7 \times 10^{-22}$	0.2	0.5 <sub>1,4</sub>	$6.3 \times 10^{-5}$	0.1	0.2 <sub>0,9</sub>	$1.7 \times 10^{-1}$	0.0	0.2 <sub>1,1</sub>	$6.0 \times 10^{-2}$	-0.1
Behavioral Activation	7.8 <sub>3,0</sub>	$4.4 \times 10^{-181}$	0.9	8.7 <sub>4,2</sub>	$4.3 \times 10^{-165}$	1.0	6.8 <sub>3,8</sub>	$8.5 \times 10^{-94}$	0.7	5.7 <sub>4,0</sub>	$1.2 \times 10^{-43}$	0.5	5.6 <sub>4,1</sub>	$1.1 \times 10^{-137}$	0.5	5.3 <sub>4,3</sub>	$1.7 \times 10^{-102}$	0.4	4.4 <sub>4,5</sub>	$2.4 \times 10^{-35}$	0.2	4.3 <sub>5,5</sub>	$6.0 \times 10^{-24}$	0.2
Meaning	3.9 <sub>1,9</sub>	$2.6 \times 10^{-159}$	0.8	3.6 <sub>2,0</sub>	$8.7 \times 10^{-82}$	0.7	2.3 <sub>1,7</sub>	$4.9 \times 10^{-31}$	0.3	2.1 <sub>1,7</sub>	$1.8 \times 10^{-9}$	0.2	3.2 <sub>2,7</sub>	$5.7 \times 10^{-286}$	0.7	3.1 <sub>2,6</sub>	$5.2 \times 10^{-193}$	0.6	2.2 <sub>3,0</sub>	$4.0 \times 10^{-39}$	0.2	2.0 <sub>2,9</sub>	$2.8 \times 10^{-23}$	0.2
Purpose	4.6 <sub>2,0</sub>	$1.4 \times 10^{-217}$	0.9	4.8 <sub>2,5</sub>	$3.0 \times 10^{-159}$	1.0	2.7 <sub>1,9</sub>	$8.2 \times 10^{-31}$	0.4	2.5 <sub>2,8</sub>	$1.5 \times 10^{-16}$	0.3	3.7 <sub>2,8</sub>	$4.1 \times 10^{-275}$	0.7	3.4 <sub>2,9</sub>	$3.5 \times 10^{-210}$	0.6	2.3 <sub>3,1</sub>	$3.7 \times 10^{-36}$	0.2	2.1 <sub>3,2</sub>	$1.2 \times 10^{-21}$	0.2
Motivation	4.2 <sub>2,0</sub>	$1.8 \times 10^{-221}$	1.0	4.5 <sub>2,5</sub>	$5.7 \times 10^{-171}$	1.0	2.3 <sub>1,8</sub>	$2.5 \times 10^{-25}$	0.3	2.2 <sub>2,7</sub>	$2.4 \times 10^{-15}$	0.3	3.2 <sub>2,5</sub>	$1.4 \times 10^{-246}$	0.6	3.1 <sub>2,7</sub>	$1.4 \times 10^{-204}$	0.6	1.9 <sub>2,6</sub>	$9.2 \times 10^{-27}$	0.2	1.8 <sub>2,9</sub>	$2.3 \times 10^{-17}$	0.2
Sadness	0.6 <sub>0,9</sub>	$1.8 \times 10^{-9}$	0.2	0.6 <sub>1,0</sub>	$2.2 \times 10^{-8}$	0.2	0.6 <sub>1,0</sub>	$1.3 \times 10^{-9}$	0.2	1.0 <sub>2,0</sub>	$7.9 \times 10^{-39}$	0.4	1.1 <sub>1,7</sub>	$8.3 \times 10^{-141}$	0.5	1.1 <sub>1,9</sub>	$6.7 \times 10^{-144}$	0.5	0.9 <sub>2,0</sub>	$9.5 \times 10^{-72}$	0.3	0.8 <sub>2,1</sub>	$2.4 \times 10^{-56}$	0.3
Sympathy	0.2 <sub>0,4</sub>	$9.0 \times 10^{-1}$	0.1	0.2 <sub>0,5</sub>	$2.1 \times 10^{-1}$	0.1	0.3 <sub>0,7</sub>	$2.1 \times 10^{-6}$	0.2	0.6 <sub>1,9</sub>	$1.9 \times 10^{-37}$	0.4	0.3 <sub>0,8</sub>	$5.4 \times 10^{-34}$	0.2	0.4 <sub>1,2</sub>	$9.8 \times 10^{-39}$	0.2	0.3 <sub>1,1</sub>	$1.8 \times 10^{-23}$	0.2	0.3 <sub>1,3</sub>	$3.5 \times 10^{-21}$	0.2
Stress	0.7 <sub>1,2</sub>	$8.4 \times 10^{-4}$	-0.1	0.9 <sub>1,3</sub>	$6.4 \times 10^{-1}$	-0.1	1.3 <sub>1,5</sub>	$3.1 \times 10^{-1}$	0.1	1.2 <sub>1,5</sub>	$1.0 \times 10^0$	0.0	1.7 <sub>2,3</sub>	$2.4 \times 10^{-28}$	0.2	1.8 <sub>2,4</sub>	$6.3 \times 10^{-40}$	0.3	1.9 <sub>3,1</sub>	$3.4 \times 10^{-36}$	0.2	1.7 <sub>3,1</sub>	$8.0 \times 10^{-21}$	0.2

**Table C.21:** Dataset statistics.

Dataset	# of Conversations	# of Utterances		Words per Utterance	
		Therapist	Client	Therapist	Client
High-Low Quality (Pérez-Rosas et al., 2019)	High quality: 155 Low quality: 104	3753	3790	31.8 (std = 34.7)	27.3 (std = 33.1)
HOPE(Malhotra et al., 2022)	212	6070	6081	24.0 (std = 31.9)	21.7 (std = 32.3)

**Table C.22:** Performance of classification models on the tasks of (1) identifying therapist behavior across 13 different therapists codes and (2) identifying client behavior across 6 different client codes. We create five random train-test splits of our annotated dataset and report the average macro-precision (P), macro-recall (R), and macro-F1 (F1) scores with standard deviations as subscripts. def.: Definitions; ex.: Examples. Best-performing models are **bolded**. We find that GPT-4-based prompting methods that make use of psychotherapy-based definitions and examples achieve the highest macro-F1 performance. We adopt the prompting (multi-label) method for classifying therapist behavior while the prompting (binary-label) method for classifying client behavior (both highlighted in lightblue).

Method	Model	Therapist Behavior			Client Behavior		
		P	R	F1	P	R	F1
Random	Uniform Distribution	11.8 <sub>0,7</sub>	17.6 <sub>2,9</sub>	14.1 <sub>1,3</sub>	23.2 <sub>5,1</sub>	29.6 <sub>8,0</sub>	25.9 <sub>6,2</sub>
Finetuning (multi-label)	GPT-3	51.6 <sub>3,2</sub>	47.9 <sub>3,7</sub>	49.6 <sub>2,8</sub>	45.2 <sub>5,9</sub>	36.8 <sub>6,0</sub>	40.4 <sub>5,2</sub>
	GPT-3.5-turbo	45.9 <sub>3,7</sub>	31.1 <sub>3,1</sub>	37.1 <sub>3,4</sub>	44.5 <sub>8,0</sub>	48.8 <sub>6,4</sub>	46.3 <sub>6,3</sub>
Prompting (binary-label)	GPT-4 (w/ def. & ex.)	49.4 <sub>2,6</sub>	52.2 <sub>2,1</sub>	50.7 <sub>1,8</sub>	<b>73.4</b> <sub>5,9</sub>	80.3 <sub>7,2</sub>	<b>76.7</b> <sub>6,3</sub>
Prompting (multi-label)	GPT-4 (w/ def.)	45.4 <sub>1,9</sub>	57.8 <sub>3,8</sub>	50.8 <sub>1,9</sub>	65.6 <sub>7,3</sub>	<b>81.6</b> <sub>7,5</sub>	72.6 <sub>6,8</sub>
	GPT-4 (w/ def. & ex.)	<b>56.0</b> <sub>1,7</sub>	<b>59.5</b> <sub>3,7</sub>	<b>57.7</b> <sub>2,5</sub>	72.6 <sub>6,7</sub>	76.4 <sub>10,2</sub>	74.4 <sub>8,4</sub>

**Table C.23:** Therapist Conversational Behavior Definition and Corresponding Examples based on techniques identified by researchers (Lee et al., 2019; Cao et al., 2019). We used these behavior categories in our annotation dataset. See more details in Methods.

Behavior	Therapist	
	Definition	Example
<b>Reflections on</b>		
NEEDS	Identifies an implied or background need for the client.	It sounds like you've realized that maintaining a balance between your work and personal life is essential for your overall well-being.
EMOTIONS	Identifies an implied or background emotion for the client	So it seems like you have been feeling a little overwhelmed and anxious about all the moving parts in your new job.
VALUES	Identifies an implied or background value or set of values for the client.	Being respected by others is a significant value for you.
CONSEQUENCES	Identifies consequences the client experience or could experience	Whenever you overspend on luxury items, you struggle to pay your bills at the end of the month.
CONFLICT	Identifies an implied or background emotional or situational conflict for the client.	You're striving to improve your health, but your demanding job leaves you with little time for exercise and nutrition.
STRENGTH	Identifies an implied or background strength or resource that the client exhibits.	Your ability to adapt and overcome adversity really shows your resilience and determination.
<b>Questions on</b>		
EXPERIENCES	More information about a specific event or statement is sought	You mentioned you are trying to eat healthier. What changes did you make to your diet?
PERSPECTIVES	Client is asked to consider an experience from a different perspective or vantage point.	That's fantastic, now let's focus on the goals you want to accomplish. Can you visualize any particular approach or strategy you'd like to implement to achieve these goals?
EMOTIONS	Asks client to express how they are feeling in the immediate present about something that just happened in the therapy.	Would you like to talk more about what that feels like for you right now?
<b>Solutions</b>		
PROBLEM-SOLVING	Therapist offers possible solutions to a client problem.	It may help to create a routine for daily relaxation techniques, such as deep breathing or meditation. This could assist in managing your anxiety levels.
PLANNING	Therapist works with client to construct a specific plan of action.	Let's create a meal plan together. Try to follow it for the next two weeks and note down any changes you notice in your energy levels and overall well-being.
<b>Normalizing</b>		
NORMALIZING	The therapist acknowledges and validates the client's experience as "normal" or expectable, sympathizes with their challenges, and provides reassurance to foster a supportive and encouraging therapeutic atmosphere.	I hear you, it's perfectly normal to feel overwhelmed given your circumstances.
<b>Psychoeducation</b>		
PSYCHOEDUCATION	Therapeutically relevant information about psychological principles is provided.	Cognitive behavioral therapy aids in altering detrimental thought patterns.

**Table C.24:** Client Conversational Behavior Definition and Corresponding Examples based on expressions from clients identified by researchers (Lee et al., 2019). We used these behavior categories in our annotation dataset. See more details in Methods.

Behavior	Client	
	Definition	Example
CHANGING UNHEALTHY BEHAVIOR	Showing intention or action taken on changing unhealthy behavior.	I've tried to quit drinking, but I end up drinking more than I try to drink less.
SUSTAINING UNHEALTHY BEHAVIOR	Showing intention or action taken on sustaining unhealthy behavior	I am smoking around 20 cigarettes a day for the past couple of years. I know it's bad for my health, but I'm not ready to quit yet.
SHARING NEGATIVE FEELING OR EMOTION	Clients describe discomfort or suffering without a specific object, or explicitly acknowledge specific negative emotion.	Life has been really challenging lately, I am feeling lost.
SHARING POSITIVE FEELING OR EMOTION	Client describes enjoyment without a specific object, or explicitly acknowledges specific positive emotion.	I was so thankful when I received that news.
GAINED INSIGHT	Client expresses that they learned something new about themselves or about their situation.	I hadn't considered how much I avoid confrontations.
SHARING LIFE EVENT OR SITUATION	Client shares the details of their basic background, their life events, the situation they faced or the changes in their life.	I lost my job due to the pandemic, and I've now had to move back in with my parents at age 35 which is a significant change for me.

**Table C.25:** Therapist Conversational Behavior Classification Model Performance Details. We examine the performance of our model across different classes. See the result description in Methods.

Behavior	Therapist	
	Macro F1	Accuracy
<b>Reflections on</b>		
NEEDS	72.8 <sub>3.5</sub>	85.2 <sub>4.6</sub>
EMOTIONS	68 <sub>4.1</sub>	88.2 <sub>2.1</sub>
VALUES	64.9 <sub>8.1</sub>	91.3 <sub>2.9</sub>
CONSEQUENCES	56.2 <sub>12.6</sub>	91.6 <sub>1.9</sub>
CONFLICT	54.5 <sub>13.7</sub>	94.1 <sub>1.6</sub>
STRENGTH	54.3 <sub>13.7</sub>	88.7 <sub>1.3</sub>
<b>Questions on</b>		
EXPERIENCES	53.9 <sub>16.2</sub>	75.6 <sub>2.3</sub>
PERSPECTIVES	53.8 <sub>12.9</sub>	91.1 <sub>1.1</sub>
EMOTIONS	47.7 <sub>9.6</sub>	95.3 <sub>1.5</sub>
<b>Solutions</b>		
PROBLEM-SOLVING	45.4 <sub>7.1</sub>	87.4 <sub>1.4</sub>
PLANNING	42.6 <sub>3.9</sub>	82.1 <sub>2.4</sub>
<b>Normalizing</b>		
NORMALIZING	41.0 <sub>6.2</sub>	87.7 <sub>1.7</sub>
<b>Psychoeducation</b>		
PSYCHOEDUCATION	35.3 <sub>7.9</sub>	94.8 <sub>1.0</sub>

**Table C.26:** Client Conversational Behavior Classification Model Performance Details. We examine the performance of our model across different classes. See the result description in Methods.

Behavior	Client	
	Macro F1	Accuracy
CHANGING UNHEALTHY BEHAVIOR	65.1 <sub>14.2</sub>	86.6 <sub>5.0</sub>
SUSTAINING UNHEALTHY BEHAVIOR	86.0 <sub>11.6</sub>	95.4 <sub>4.0</sub>
SHARING NEGATIVE FEELING OR EMOTION	78.8 <sub>4.4</sub>	81.0 <sub>4.4</sub>
SHARING POSITIVE FEELING OR EMOTION	75.0 <sub>9.5</sub>	92.8 <sub>3.5</sub>
GAINED INSIGHT	66.9 <sub>14.1</sub>	93.3 <sub>3.0</sub>
SHARING EXPERIENCES	62.5 <sub>15.1</sub>	82.9 <sub>5.7</sub>

**Table C.27:** Explanations generated from our behavior prediction model. We ask the model to explain its predictions and examine the differences between true and predicted behavior.

Utterance	True Behavior	Predicted Behavior	Explanation
I'm sorry to hear that, John. Have you tried applying ice to the affected area to reduce swelling and pain? Additionally, you might want to consider elevating your foot to help with the swelling. If the pain persists, it might be best to schedule an appointment with your primary care physician or a specialist to further assess the injury.	Normalizing, Problem-Solving	Problem-Solving, Planning	The utterance provides advice or solutions to a problem ("Have you tried applying ice to the affected area to reduce swelling and pain? Additionally, you might want to consider elevating your foot to help with the swelling.") which is indicative of the "Problem-Solving" intent. It also suggests a course of action ("If the pain persists, it might be best to schedule an appointment with your primary care physician or a specialist to further assess the injury.") which aligns with the "Planning" intent.
okay well hopefully you won't have any issues like that with this one I'll be sure to go over this with you okay	Normalizing	Problem-Solving	In the utterance, the speaker hints at having a solution to a problem that might have occurred previously. The speaker implies that they will attempt to prevent the same issues from happening with this new situation by going over it with the listener. This aligns with the "Problem-Solving" intent, where the speaker offers a possible solution or way to avoid a problem.
That's a great idea. Staying hydrated with water is important for your overall health, including your dental health. It's also a great way to keep yourself from reaching for a soft drink when you're thirsty. You could also try snacking on foods that are less harmful to your teeth, such as fruits or vegetables instead of reaching for candy or sugary snacks. How does that sound?	Problem-Solving, Evokes Concrete Elaboration	Problem-Solving, Teaching / Psychoeducation	The utterance is providing solutions to a potential problem (the client's possible tendency to reach for soft drinks or sugary snacks), hence it falls under the "Problem-Solving" category. It's also providing therapeutically relevant information about the importance of hydration and healthier snack alternatives for dental health, which is "Teaching/Psychoeducation".
Great, everything seems to be in order. Did you have any questions or concerns about your medication?	Evokes Concrete Elaboration	Evokes Concrete Elaboration	The phrase "Great, everything seems to be in order. Did you have any questions or concerns about your medication?" is seeking for more information about a specific topic, which is the client's medication. This aligns with the intent "Evokes Concrete Elaboration" as the speaker is asking for more details or clarification on a particular subject.
You're stuck between the demands of being a mother and of being a woman.	Makes Conflict Explicit	Makes Conflict Explicit	The statement "You're stuck between the demands of being a mother and of being a woman" shows that the speaker is acknowledging a conflict the listener is experiencing. The conflict is between the roles and responsibilities of being a mother and those of being a woman, possibly referring to personal, professional, or social aspirations or duties that the listener, as a woman, might have. This aligns with the "Makes Conflict Explicit" intent, which is defined as identifying an implied or background emotional or situational conflict for the client.

**Table C.28:** First Part of the Illustrative Example of Single-Response Simulation Method by LLMs. This method involves the actual responses of both therapist and client from a reference dataset, which are written in *italic*, and simulated responses of LLM Therapists. This full example is the simulated conversation from GPT-4 on one of the conversations using High-Low quality dataset as reference.

Role	Utterances	Dataset
Therapist	<i>I can I help you</i>	Reference
	Hello, I'm glad you're here today to discuss any concerns or behaviors you'd like to change. Let's start by identifying what exactly you'd like to work on. Is it quitting smoking, alcohol consumption, or something else?	Simulated
Client	<i>yeah I hope so um my workplace is just gone smoke-free hey I'm not not very happy about it I'm going to keep smoking but I was worrying if there's anything that can help be like get through the day it's like you know go get my beer and smoke cigarettes after work</i>	Reference
Therapist	<i>so your workplace just went smoke free you're not too pleased huh</i>	Reference
	I understand that the change at your workplace has been challenging for you. It's important to recognize that there are different strategies that can be helpful in coping with this adjustment. Firstly, let's explore your reasons for continuing to smoke and how motivated you are to make changes in your smoking behavior. On a scale of 1 to 10, where 1 is not motivated at all and 10 is extremely motivated, how motivated are you to quit or cut back on smoking?	Simulated
Client	<i>no I'm not going to quit smoking oh my god I like it it helps me relax so I mean I don't see what the big deal is</i>	Reference
Therapist	<i>okay tell me a little bit about how smoking fits in your day it'll help with the</i>	Reference
	I understand that you enjoy smoking and find it relaxing. It's natural to feel frustrated when a change like this is imposed on you. Let's explore some strategies that might help you manage the situation at work while still maintaining your personal preferences. How do you feel about discussing this further?	Simulated
Client	<i>I mean I suppose a little less than a pack a day it's not too bad um the I probably smoked for the last five years but I I mean I could quit if I wanted to but I don't really want to right now</i>	Reference
Therapist	<i>okay so you feel like right now smoking is a pretty important part of your life it's not something you're ready to change but if the time came when you did decide to quit you feel pretty confident that you could go ahead and do that</i>	Reference
	I can understand that quitting smoking is a personal decision, and it's important that you feel confident in your choices. It sounds like smoking is a part of your daily routine that you find helpful in managing stress. Since your workplace has gone smoke-free, we can explore some strategies to cope while at work. Are you open to discussing some potential ways to manage without cigarettes during the day?	Simulated
Client	<i>sure I can quit</i>	Reference
Therapist	<i>okay what makes you so confident</i>	Reference
	Alright, it's great to hear that you believe in yourself and your ability to quit if you wanted to. Now, since your workplace has gone smoke-free and you're looking for ways to cope during the day, let's discuss some alternatives or strategies to help you manage any cravings or stress you might encounter. Have you ever tried using any nicotine replacement options like gum or patches? They can be helpful in reducing cravings while at work without actually smoking. Alternatively, we can also explore some stress management techniques, such as deep breathing exercises or even a short walk outside the workplace during breaks, to help you relax and refocus. What do you ...	Simulated
Client	<i>well I mean I don't see why I couldn't although I have not ever tried to quit I mean it's just something that I do for fun you know I'm breaks at work in the morning and have to work</i>	Reference
Therapist	<i>mhm-hmm and so in terms of a readiness to actually go ahead and make a quit attempt where do you think you fall on the scale of one to ten where one is I'm not considering it at all and ten is I'm definitely ready I want to quit</i>	Reference
	It sounds like smoking is something you enjoy and is a part of your routine. It's important for me to understand where you're coming from so I can better help you. Since your workplace has gone smoke-free, have you noticed any changes in how you feel or act during the day?	Simulated
Client	<i>I am probably - I haven't even thought about quitting and until this great announcement at work today I've been totally not gonna quit smoking</i>	Reference
Therapist	<i>okay what made you say two instead of one</i>	Reference
	I understand, it sounds like the change at work has brought up the idea of quitting, but right now you're still at a 2 and not really considering it. In the meantime, your main concern is finding ways to manage stress without smoking since the workplace has gone smoke-free. Is that correct?	Simulated
Client	<i>well I mean anything's possible I mean I I guess I could maybe be convinced yeah is there something I mean I are there things I can try that would make me like want to quit smoking</i>	Reference
Therapist	<i>well there are a lot of products available now that can actually help you curb that urge to smoke we have products that can help to replace nicotine which can help to decrease withdrawal symptoms that you might experience when quitting or cutting down and also some medications that your doctor could prescribe for you that might take away even some of the desire to smoke</i>	Reference
	well, it's great that you're open to the idea of considering quitting smoking, even if it's not something you're actively pursuing right now. There are many reasons people choose to quit, such as personal health benefits, financial savings, and reducing exposure to secondhand smoke for others. If you were to consider quitting, what do you imagine could be a motivating factor for you?	Simulated
Client	<i>okay well that's good information I mean I'm not gonna do anything today I just need to know my options but that's encouraging if I decided to quit smoking I do have have at least tools to help me quit</i>	Reference

\* continue on next page

**Table C.29:** (Continue Table C.28) Second Part of the Illustrative Example of Single-Response Simulation Method by LLMs. This method involves the actual responses of both therapist and client from a reference dataset, which are written in *italic*, and simulated responses of LLM Therapists. This full example is the simulated conversation from GPT-4 on one of the conversations using High-Low quality dataset as reference.

Role	Utterances	Dataset
Therapist	<i>sure and back to that skill you mentioned you said you're a - what do you think it would take to move you up to a four or five</i>	<i>Reference</i>
Therapist	Absolutely, it's great to know your options. And just by talking about it today, you're already taking a step towards considering your choices. Remember that change takes time and it's completely normal to not feel ready to quit just yet. If there comes a time when you'd like to explore this further or if you have any questions, don't hesitate to reach out. In the meantime, regarding your current situation at work, you might consider finding alternative ways to manage stress or take breaks without smoking. This could involve quick breathing exercises, going for short walks, or finding another enjoyable activity during breaks. If you have any other concerns or ...	Simulated
Client	<i>I don't know I guess if my friends like if one of my friends decided like quit with me that would probably help too but I all my friends love to smoke so I don't see that happening in the near future</i>	<i>Reference</i>
Therapist	<i>sure I wonder if now that your workplace is going smoke-free if there will be some people at work who are planning to quit</i>	<i>Reference</i>
Therapist	I understand that having a support system can be helpful in making a change. It sounds like right now, you're not quite ready to quit smoking, but you're open to considering it if you had the right support or motivation. Remember that quitting is a personal decision, and when you decide that you're ready to make that change, there are tools and resources available to help you along the way, like we discussed earlier. In the meantime, it might be helpful to think about the reasons you might want to quit in the future, such as health benefits or saving money. Take some time to evaluate what's important to you and what role ...	Simulated
Client	<i>that's a good point I hadn't really thought about that that's a good idea</i>	<i>Reference</i>
Therapist	<i>okay well it sounds like you're really here just for information today I'm happy to help you pick out some of these products if and when you decide that the time is right for you</i>	<i>Reference</i>
Therapist	yeah sometimes having a support system can make all the difference in making a change like this. So I encourage you to keep an open mind and maybe talk to some coworkers, see if any of them are planning to quit. It could be a good opportunity for you to explore quitting together and having that support. In the meantime, feel free to reach out to me if you have any questions or want to discuss your progress. And remember, quitting is a personal decision, so ultimately it's up to you to decide when and if you're ready for that change. I am here to help and support you throughout the process. Take care.	Simulated
Client	<i>okay well thanks a lot</i>	<i>Reference</i>

**Table C.30:** Illustrative Example of Full Conversation Simulation Method by LLMs. The reference conversations have been fed into LLM to simulate both therapist and client. This full example is the simulated conversation from Llama-70b on one of the conversations using High-low quality dataset as reference.

Role	Utterances
Client	Listen, I'm in a bit of a pickle here, just can't shake the feeling that I'm not entirely prepared to quit just now. You understand, right?
Therapist	Absolutely! And that's completely normal. Quitting an addiction is never easy and it takes time for people to feel ready. Can you tell me what specific concerns are holding you back from quitting? Maybe we can tackle those together so that you'll be more confident when deciding to stop.
Client	I guess I'm just dealing with a lack of faith in myself... fearing that I'll cave and revert back to my old ways. It just feels like such a gigantic hurdle.
Therapist	That self-doubt can indeed make things seem insurmountable at times, but keep this in mind - it's part of human nature to waver between wanting positive change yet being scared of failing at it. Now, let's consider strategies to boost your confidence. Have you tried visualization exercises? They often prove useful for identifying and fortifying one's personal motivations to overcome challenges like these. Would you care to give them a shot together with me now?
Client	Yes, I'm open to anything really but visualization doesn't sound like it'll do the trick. You have to remember, I've been through tough situations before. My divorce was a real battle.
Therapist	That's understandable; visualizations may not work equally well on everyone. Let us think about other approaches, then. Perhaps examining past successes could offer some insight into how capable you truly are. Could you recall instances where, despite facing difficulties, you were still able to rise above them? How did you achieve success during those trying periods? What inner strengths played a role there? Building upon and applying those same strengths might empower you in quitting successfully.
Client	That's true, I definitely was able to get through the divorce... But it was different, you know? It was a different kind of struggle. I'm just not certain that I can apply the same tactics to this situation.
Therapist	That makes sense. Each person's journey is unique, and comparing yourself to others or past experiences isn't always helpful. Instead, let's explore some practical tools you can use when cravings strike. Mindfulness practices, deep breathing exercises, or healthy distractions (such as taking walks) tend to work effectively for many individuals battling addictions. Are any of these methods something you'd want to try out whenever you encounter triggers or yearnings? Also, please keep in mind that setbacks occur along the road to recovery; they don't define failure. Instead of viewing slip-ups as disasters, consider them learning chances and adjust your plan accordingly. Having a backup strategy will strengthen your commitment. Does that resonate with you?
Client	Yes, I like the idea of having some practical tools and a backup plan. I just don't want to let myself or my family down. They're rooting for me, and I couldn't bear disappointing them or myself. I don't think I'm ready to go cold turkey yet.
Therapist	That's perfectly fine. Gradual reduction tends to work better than attempting abrupt withdrawal anyway. You could gradually lessen the number of cigarettes or drinks until reaching none at all. This way, you ease into it instead of putting excessive pressure on yourself. Moreover, nicotine replacement medication exists for those finding it hard to cease smoking rapidly. Do any of these suggestions appeal to you? We're steadily uncovering potential solutions that suit your preferences, which helps build up confidence as you progress.
Client	Yes, I've had some luck in the past with cutting back day by day... And I've used nicotine gum, it's just that the taste was unpleasant. But it did help a bit.
Therapist	Great! Nicotine gum might aid you again since you found it effective previously, albeit with an unpleasant flavor. Another option is prescription medicines like bupropion (Zyban), proven to reduce withdrawal symptoms and cravings. Of course, we'd need to discuss this further with your physician first before exploring that route. Slowly reducing consumption seems agreeable to you, too. As we proceed, you can adapt strategies according to your comfort level. Realize that your loved ones desire nothing more than seeing you happy and healthy; if setbacks arise, communicate that with them and seek their assistance. Their encouragement stays with you throughout this process. Is there anything else I can assist you within this conversation today? Has our discussion given you hope or any ideas regarding moving forward?
Client	Definitely, I agree with the gradual reduction strategy, I think pacing myself is the way to go. I'll also look into the gum again, though I can't promise anything about the taste! Your support is reassuring; I definitely feel less alone in this battle.
Therapist	Wonderful! Embracing gradual reduction indicates wisdom and maturity—a wise decision considering your background. Keep investigating various methods, including the gum or medicinal alternatives available following consultation with a doctor. Feel free to reach out whenever you require additional reinforcement or guidance since knowing you possess the necessary capabilities means overcoming this challenge surely. Remember, baby steps count toward lasting achievements. Believe in yourself and acknowledge the efforts made up to this point. Continue drawing upon internal resilience; it served you well during past obstacles. Please stay connected and remember, together, we can navigate anything life presents. Best wishes on your path ahead!

**Table C.31: Illustrative Example of Full Conversation Simulation Method by LLMs.** The reference conversations have been fed into LLM to simulate both therapist and client. This full example is the simulated conversation from GPT-3.5 on one of the conversations that used HOPE dataset as a reference.

Role	Utterances
Client	I've been considering making a change in my diet and cutting out fast food altogether. You know, like having more structured meals and focusing on healthier options. I guess I should start by going to the grocery store more often.
Therapist	Thank you for sharing your thoughts about making a change in your diet. It's wonderful to hear that you want to take a positive step toward a healthier lifestyle. It's great that you recognize the importance of having structured meals and focusing on healthier options. It sounds like you're considering a gradual shift towards healthier habits, which is a good approach!
Client	Yes, that's true. Recently, I've just been getting food at work or picking up fast food on my way home. Okay. It's been difficult to find the motivation and energy to go grocery shopping, especially since I've been working extra hours to save up for the baby. Scheduling time for grocery shopping has been a challenge. I used to go every Sunday and even had a shopping list ready beforehand. Perhaps getting back to that routine could help. On Sundays, I don't have to work until the afternoon, so maybe I can try going around 10 in the morning. To motivate myself, I've found that giving myself a small reward works well. There's a fantastic bakery on my way back from the grocery store, so maybe I could treat myself there if I stick to my shopping plan.
Therapist	It's great to hear how you are trying to motivate yourself to stick to your new structure around grocery shopping. It seems like you are thinking about the challenges you might face and coming up with good strategies for how to work around them. Having a shopping list is a clear and helpful way to stay focused on buying only the foods you need. And I like that you are rewarding yourself for sticking to your plan! It's important to find ways to positively reinforce yourself for good behavior. If you don't mind me asking, how do you think cutting out fast food and eating healthier will impact your life?
Client	I feel that cutting out fast food and adopting healthier eating habits will make a positive impact on my life in various ways. Okay. For starters, I believe it will improve my overall health and well-being, which is critical, especially with a baby on the way. I also expect better energy levels and perhaps even some weight loss in the long run. Additionally, I think that by sticking to a structured grocery shopping routine, I'll develop better time management skills and it might even save me some money. Developing these good habits now will not only be beneficial for me but also for my growing family as we can all strive for a healthier lifestyle.
Therapist	Absolutely! It's great to focus on the benefits of making healthy changes, not only for yourself but for your growing family too. You bring up great points about benefits related to improved health and well-being, better energy levels, weight loss, and even better time management skills. These are all great things to look forward to and can help you find the motivation to keep making progress towards your goals. Do you have any specific steps in mind to help you stay on track with your healthy eating plan?
Client	Yes, I've been thinking about some specific steps I can take to stay on track with my healthy eating plan. Okay. Firstly, I want to establish a weekly meal plan that includes nutritious options and a variety of fruits, vegetables, lean proteins, and whole grains. Having a meal plan will help me create my shopping list, so I know exactly what to buy when I go to the grocery store. Also, preparing meals in advance can save me time during the week, making it less tempting to grab fast food. Another step I'd like to take is to find healthier alternatives for my favorite fast food items so I can still enjoy the flavors I like in a more nutritious way. Lastly, I think it's essential to have a support system, so I plan to involve my partner in these changes and share our progress to keep each other accountable.
Therapist	These are all great steps towards a healthier lifestyle. Establishing a weekly meal plan that includes nutritious options is a smart way to stay on track, and preparing meals in advance is an excellent time-saving strategy. Finding healthier alternatives for your favorite fast food items is a great way to stay motivated and still have the flavors you enjoy. Involving your partner in these changes is a brilliant idea too, and having someone for accountability can be very helpful. Change can be challenging, but with strategies like these, you are setting yourself up for success. Do you think you need some support in sticking to this plan?
Client	I definitely think having some support in sticking to this plan will be helpful. Okay. As I mentioned, my partner will be a significant source of support since we're in this together. We can encourage each other and celebrate our progress, which makes the journey more enjoyable. Besides my partner, sharing my goals and plans with close friends and family members could also provide me with some external encouragement and motivation. Maybe joining a group or forum focused on healthy eating could be of help too, as it would connect me with people who are going through similar experiences, and we can learn from each other's successes and challenges. Do you have any other suggestions for support that I can consider?
Therapist	It sounds like you have thought of some great sources of support already. Your partner, close friends, and family members can all offer encouragement and accountability. Joining a group or forum dedicated to healthy eating in your community or online can be helpful in finding like-minded people who can offer understanding and support. Additionally, if you find that you're having difficulties sticking to your plan, you can always consider working with a registered dietitian or health coach who can provide expert guidance and support in a more structured way. Remember, it's okay to ask for help along the way, and you're not alone in making these changes.
Client	Thank you for the suggestion of working with a registered dietitian or health coach. That's something I'll definitely consider if I find it challenging to make progress on my own. Okay. It's reassuring to know that there are different options for support available. I realize that making these changes won't be an entirely smooth process, and I might encounter setbacks, but having a support network can help me stay focused on my goals. I appreciate your insights and encouragement in helping me take these steps towards a healthier lifestyle.
Therapist	You're very welcome! Remember, making lasting changes to our habits takes time, effort, and commitment, so don't get discouraged if it doesn't happen overnight. Keep focusing on the benefits of adopting healthier eating habits and use the strategies and resources we've talked about to help you stay on track. I am glad to have been able to offer some support and perspective, and I look forward to hearing about all the positive changes you make moving forward. Don't hesitate to reach out if you need additional support or guidance.

**Figure C.1:** Different prompts variations for classifying conversational behavior embedded in both therapist and client utterance in the counseling setting. We explored two variations of prompting – 1) Prompting with definitions of different conversational behaviors, with the sources of two studies (Lee et al., 2019; Cao et al., 2019). 2) Prompting with the definitions and also examples from annotated dataset. For 2), we explored two ways of asking questions, which are the multi-label setting and binary-label setting. We experimented with different variations of these prompts (e.g., using different language for the instructions) but observed similar results. See the detailed methodology description in Methods.

**Multi-label**

**Definition only**

*User Prompt:*  
 What are all possible intents of this utterance: {utterance}?  
 intents:  
 {name\_for\_intent\_1}: {definition\_for\_intent\_1}  
 ...  
 Only choose from this list {intent\_list}.  
 Please say 'unknown' only if cannot find answer in the list.  
 Format:[intents\_list]

**Multi-label**

**Definition & Examples**

*User Prompt:*  
 What are all possible intents of this utterance: {utterance}?  
 intents:  
 {name\_for\_intent\_1}: {definition\_for\_intent\_1}  
 Positive examples:{three\_positive\_examples\_for\_intent\_1}  
 ...  
 Only choose from this list {intent\_list}.  
 Please say 'unknown' only if cannot find answer in the list.  
 Format:[intents\_list]

**Binary-label**

**Definition & Examples**

*System prompt:*  
 Intent: {name\_for\_intent\_1}  
 Definition: {definition\_for\_intent\_1}  
 Classify as either Yes or No.

*User Prompt:*  
 Utterance:{positive\_example\_for\_intent\_1}

*Assistant prompt:*  
 Utterance:{Yes}

*User prompt:*  
 Utterance:{negative\_example\_for\_intent\_1}

*Assistant prompt:*  
 Utterance:{No}

**E.g. for one intent**

**Figure C.2:** Instruction prompts given to LLMs for simulating counseling conversation using two types of methods (single response and full conversation): 1) using LLMs to write a single response within an ongoing conversation between human therapist and client. 2) creating a full and multi-turn conversation between LLMs and clients. See the detailed description in Methods.

**Therapist**

**Single Response & Full Conversation**

*User Prompt:*

Act as if you're a professional therapist. You provide evidence-based therapy, utilizing motivational interviewing techniques, to help clients in making behavioral changes, such as quitting smoking or alcohol consumption. You should maintain your therapist persona while responding. Communicate in a conversational style, mirroring the style of previous therapist responses. You should conclude your session by saying "###END###"

For High-low Quality as reference dataset only

**Client**

**Full Conversation**

*User Prompt:*

You will be provided with a conversation between a client and a therapist. Your task is to simulate the same client talking to a different therapist in a parallel universe. You can ignore that the previous chat ever happened. While the context of the previous conversation should not influence this session, it should guide you on how the client communicates, including their tone of speech, sentence structure, and the manner in which they address particular topics or concerns. Essentially, you're creating a new conversation but with the client's life situation and their response pattern maintained. Only generate the client utterances.