

©Copyright 2025

Daniil D Filienko

From Therapy to Treatment:  
Transforming Healthcare Support  
with Large Language Models

Daniil D Filienko

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

Master of Science

University of Washington

2025

Committee:

Martine De Cock

Weichao Yuwen

Program Authorized to Offer Degree:  
Computer Science and Systems

University of Washington

**Abstract**

From Therapy to Treatment:  
Transforming Healthcare Support  
with Large Language Models

Daniil D Filienko

Chair of the Supervisory Committee:  
Martine De Cock  
School of Engineering and Technology

The rapid advancement of Large Language Models (LLMs) has opened new avenues for AI-assisted healthcare, particularly in chronic disease management. This study explores the application of in-context learning methods to enhance LLMs' ability to deliver Problem-Solving Therapy (PST) and support tuberculosis (TB) treatment adherence. We investigate how LLMs can improve the quality and empathy of AI-driven therapy sessions. Additionally, we propose the integration of LLMs into digital adherence technologies to facilitate interactive patient-provider communication during TB treatment. We leverage prompt engineering, Retrieval Augmented Generation (RAG), and multi-agent systems. Our evaluation across both projects employs both automatic metrics and expert human assessment to analyze the effectiveness of these AI-driven interventions. Findings indicate that while LLMs provide a promising tool for enabling better ongoing care for people with chronic disease across different fields, challenges remain in maintaining privacy, safety, and ethical considerations. This research contributes to the growing field of AI-enhanced healthcare, highlighting the potential and limitations of LLMs in bridging mental health and infectious disease treatment gaps.

## TABLE OF CONTENTS

	Page
Chapter 1: Introduction . . . . .	1
1.1 Motivation . . . . .	1
1.2 Overview . . . . .	2
1.3 Contributions . . . . .	4
1.4 Publications with Results from the Thesis . . . . .	6
1.5 References . . . . .	7
Chapter 2: Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy . . . . .	10
Chapter 3: Large Language Model-Powered Conversational Agent Delivering Problem-Solving Therapy (PST) for Family Caregivers: Enhancing Empathy and Therapeutic Alliance Using In-Context Learning . . . . .	21
Chapter 4: Transforming Tuberculosis Care: Optimizing Large Language Models for Enhanced Clinician-Patient Communication . . . . .	36
Chapter 5: Conclusion and Discussion . . . . .	48

## ACKNOWLEDGMENTS

First, I would like to acknowledge the collaborators who made these research projects possible. I have received a lot of help from the researchers mentioned here and many others, which makes my heart overflow with gratitude. There are too many people to name, but my thanks go to the research groups supervised by Prof. Weichao Yuwen and Prof. Sarah Iribarren. Prof. Martine De Cock has a unique ability to gather people across disciplines, fields, and continents, different in specialties, but united by the shared purpose. The projects I have worked with all involved groups from different backgrounds, showing me the indispensable value of interdisciplinary research.

My research group deserves an acknowledgment for teaching me to ask right questions and be critical to my work. Professor Martine De Cock deserves all of the credit for her unwavering support, encouragement, and immense patience.

The financial support I've received as an NIH AIM-AHEAD Research Fellow and as a Carwein-Andrews Distinguished Fellow, along with support from the UWT Founders Endowment and the UW Population Health Initiative, enabled me to finish this research.

Finally, my parents made all this possible. They supported me, despite being first-generation immigrants in this country themselves. I am indebted to them.

## Chapter 1

# INTRODUCTION

### *1.1 Motivation*

Chronic disease can truly be called a scourge of our time. Speaking from personal experience, I know firsthand how many of my friends and family have been severely impacted by some form of chronic illness. Both of my parents have either battled or are currently battling cancer, so I have seen firsthand the toll chronic disease take both on the patient and on the family members caring for them. I know that I am not alone in that situation. According to the CDC (Centers for Disease Control and Prevention), 6 out of 10 Americans have at least one chronic disease [21], so chances are high that you or your loved ones are affected too, whether directly or indirectly. And while cancer may be a chronic disease which more people have heard about, there are other chronic diseases that cause sufferings to millions of people per year world-wide despite being treatable and curable.

A recurring challenge in chronic disease management are barriers to finding care. The demand for healthcare services is very high, and it continues to increase as the population ages and medicine advances. The supply of healthcare providers is not sufficient to meet the existing and rising demand [24]. Shortages in physicians and nurses make it hard to receive care in a timely manner, a phenomenon that is especially pronounced in low- and middle-income regions. Innovative AI solutions hold potential to reduce the burden on caregivers in the future. In this thesis, I investigate the possibilities of Large Language Models (LLMs) to expand access to care for patients with a chronic diseases as well as for their family caregivers.

## 1.2 Overview

There are many people with chronic diseases in the US. In addition to getting treatment from healthcare professionals, individuals with chronic diseases rely on family caregivers, i.e. family or friends who help with all kinds of daily activities and provide mental support. Some studies suggest that as many as 18.2% of American adults provide some form of unpaid care for an adult family member [6]. Caregivers experience physical, mental, and financial burdens, which lead to burnout. They prioritize the loved one under their care and often do not have enough time or resources to schedule treatment for themselves. Hence it comes as no surprise that, in the family caregiving community, there is a large interest in developing self-care skills [3], demonstrating a great place to implement a high-quality AI based tools for psychological support. Such self-care and support are necessary to mitigate negative effects of caregiving, such as additional stress or depression [16].

In my thesis, I propose a conversational AI agent to support family caregivers in their self-care. The agent is based on Problem Solving Therapy (PST), a proven psychotherapy method [13]. Unlike an existing rule-based system to provide dialogue-based PST care to family caregivers [10], my conversational agent builds on recently developed Large Language Models (LLMs) to guide the user through a PST therapy session. The questions and responses of the LLM-based conversational agent flow more naturally than those of rule-based systems. This matters, as it is known that more successful human counselors also tend to be more varied in their responses and use less templated answers [1]. Benefiting from LLMs being easy accessible online, the agent can be used by family caregivers even in remote areas with limited ability to see human counselors or otherwise lacking necessary psychological resources.

In Chapter 2, I present my findings on how prompt engineering can be used to create such a conversational AI agent. I improve upon existing work [4] by incorporating different prompt engineering methods, such as few-shot learning [2], using a state-of-the-art LLM model, namely GPT-4, and perform a comprehensive evaluation with clinical experts and relevant

empathy assessment algorithms [17]. Such an evaluation is necessary before deploying the agent to real caregivers, in order to assess the agent’s effectiveness and ability to adhere to PST guidance.

In Chapter 3, I extend the work from Chapter 2 by applying my methods to an open-source LLM, namely Llama3.1-70b. I furthermore incorporate Retrieval Augmented Generation (RAG) [11] for more accurate psychotherapy guidance following, and expand the evaluation to real caregivers. Such an evaluation is critical to assess whether the tool can address real needs of real people. The collected results are promising, showing that caregivers generally find the model both engaging and empathetic.

In Chapter 4, I investigate how the results and insights obtained for English speaking family caregivers in Chapter 2 and 3, transfer to another language (Spanish) in a very different healthcare setting. The problem that I address in Chapter 4 is treatment non-adherence in tuberculosis (TB) patients. TB continues to be the world’s deadliest infectious disease despite being preventable and curable [23]. The WHO’s 2030 targets in alleviating its burden have not been met [7], leading to more than a million deaths in 2023 alone [23]. The death rate is especially high in low- and middle-income countries with limited access to necessary resources. Even when provided with the medications, the 6 months treatment course remains challenging to complete, due to limited social support, caused by TB stigmatization, painful side-effects, and inconsistent communication between patients and treatment supporters [19]. Digital Adherence Technologies (DATs), which include tools such as phone-based chats enabling faster information exchange between healthcare provider and patients, and ingestible sensors enabling consistent monitoring of the patient’s status, became a promising direction to improve treatment adherence [18]. However, many of the current DATs put additional strain on the treatment providers. With the use of LLMs, I propose to improve the communication process during treatment by generating high-quality response suggestions for the provider. Suggested responses directly respond to the patients’ needs and take into account both the previous interactions between the patient and the provider, local guidance, and AI model’s own knowledge, leading to faster communication

with less mental strain for the treatment provider.

Both of these problem domains, i.e. supporting family caregivers and supporting individuals with TB, require a solution that is both fast and human-like. Large Language Models present a viable option, combining fast computations with human-like speech generation and empathy capabilities. I improve upon the off-the-shelf capabilities of generalist models, such as GPT-4 and Llama-3, by incorporating well-designed in-context learning techniques, leading to promising results. My work includes a comprehensive evaluation by both clinical experts (Chapters 2 and 4) and the intended users of the LLM (Chapter 3). It demonstrates the impressive ability of LLMs to be fitted to different medical contexts successfully.

### **1.3 Contributions**

Large Language Models (LLMs) – such as the GPT series of models that power the popular chatbot ChatGPT – are a milestone in the history of AI. The rise in capabilities and popularity of LLMs within less than a decade since their underlying transformer architecture was introduced [22] is truly remarkable. Whether trained on masking, such as BERT [5] or on next token prediction, such as the GPT series [15], transformers have spurred breakthroughs and a lot of excitement across academia and industry, a trend which is expected to continue in the foreseeable future.

A lot of the improvement in transformers during the past few years has been due to increasing size and number of parameters, allowing them to capture and generalize over greater amounts of data, rather than fundamentally new architectures. While modern models are good generalists [14], additional injections of domain specific knowledge help LLMs to perform better in specialized domains, such as healthcare. *In this thesis, I investigate the potential and pitfalls of domain knowledge injection through in-context learning for the development of conversational agents to support medical professionals and provide broader access to care.*

My work focuses on tailoring LLMs specifically for two real-world use cases, (1) providing Problem Solving Therapy (PST) to English speaking family caregivers (see Chapters

2 and 3) and (2) supporting tuberculosis (TB) patients in Spanish speaking Argentina to adhere to their treatment (see Chapter 4). Previous and concurrent work on Conversational AI in healthcare tends to either focus on developing highly empathetic psychological care bots [8, 9] or highly clinically accurate diagnosis and treatment models [12, 20]. Research in psychological care focuses on improving the empathetic capabilities of the model, making it more human-like while strictly adhering to psychological therapy guidelines. Studies in clinical diagnosis and treatment focus on ensuring that the AI provides proper diagnosis and treatment options for the given symptom or question [12, 20]. Empathy focused efforts improve the stylistic abilities of the model to make it appear more human-like while clinical accuracy focused ones improve the factuality of responses to ensure model’s robustness. *My work spans both domains, bridging that important gap, akin to how a good physician would be trained to be both relatable and accurate.*

## Primary Contributions

The primary contributions made in this thesis are the following:

- Adapting different in-context learning techniques to tailor general purpose LLMs to important healthcare applications, namely delivering Problem Solving Therapy (PST) to family caregivers and supporting tuberculosis (TB) patients in adhering to their treatment.
- Designing an evaluation framework necessary to empirically compare different in-context learning techniques to adapt LLMs to deliver PST therapy and TB treatment support.
- Extending existing line of work in in-context learning [14] to a non-English language setting.

#### 1.4 Publications with Results from the Thesis

Chapter 2 and 4 of this thesis have been published in venues related to AI and healthcare, and Chapter 3 is currently under review as a paper submission.

- I presented Chapter 2 at AMIA 2024:

**D. Filienko**, Y. Wang, C. El Jazmi, S. Xie, T. Cohen, M. De Cock, W. Yuwen. Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. *American Medical Informatics Association (AMIA) Annual Symposium*, 2024

- Chapter 3 is under review as a conference submission:

L. Wang, D. Carrington, **D. Filienko**, C. El Jazmi, S. Xie, M. De Cock, S. Iribarren, W. Yuwen. Large Language Model-Powered Conversational Agent Delivering Problem-Solving Therapy (PST) for Family Caregivers: Enhancing Empathy and Therapeutic Alliance Using In-Context Learning. *Under review*, 2025

- I presented Chapter 4 at the AAAI-25 GenAI4Health workshop:

**D. Filienko**, M. Nizar, J. Roberti, D. Galdamez, H. Jakher, S. Iribarren, W. Yuwen, M. De Cock. Transforming Tuberculosis Care: Optimizing Large Language Models for Enhanced Clinician-Patient Communication. *Workshop on Large Language Models and Generative AI for Health at 2025 AAAI Conference on Artificial Intelligence*, 2025

## 1.5 References

- [1] Tim Althoff, Kevin Clark, and Jure Leskovec. Natural language processing for mental health: Large scale discourse analysis of counseling conversations. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
- [3] Qianqian Chen, Huimin Zhang, Suwei Yuan, Wenwei Liu, and Tongzhou Lyu. Perceived training needs of the informal caregivers of older adults: A cross-sectional study. In *Healthcare*, volume 12, page 2369. MDPI, 2024.
- [4] Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *CoRR*, abs/2305.13614, 2023.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.
- [6] National Alliance for Caregiving. Caregiving in the US 2015. *NAC and the AARP Public Institute*, 2015. Greenwald & Associates, Washington, DC.
- [7] Rena Fukunaga, Philippe Glaziou, Jennifer B. Harris, Anand Date, Katherine Floyd, and Tereza Kasaeva. Epidemiology of tuberculosis and progress toward meeting global targets – worldwide, 2019. *Morbidity and Mortality Weekly Report (MMWR)*, 70(12):427–430, 2021.
- [8] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems*, 2023.

- [9] Cheng Kang, Daniel Novak, Katerina Urbanova, Yuqing Cheng, and Yong Hu. Domain-specific improvement on psychotherapy chatbot using assistant. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 351–355, 2024.
- [10] William R Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [12] Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, et al. Polaris: A safety-focused llm constellation architecture for healthcare. *arXiv preprint arXiv:2403.13313*, 2024.
- [13] Arthur M Nezu, Christine M Nezu, and Thomas J D’Zurilla. *Problem-solving therapy: A treatment manual*. Springer Publishing Company, 2012.
- [14] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolò Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *CoRR*, abs/2311.16452, 2023.
- [15] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. <https://api.semanticscholar.org/CorpusID:49313245>, 2018.
- [16] Richard Schulz and Jill Eden. Committee on family caregiving for older adults board on health care services health and medicine division. *Families caring for an aging America*, pages 1–366, 2016.
- [17] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, 2020.

- [18] Ramnath Subbaraman, Laura de Mondesert, Angella Musiimenta, Madhukar Pai, Kenneth H Mayer, Beena E Thomas, and Jessica Haberer. Digital adherence technologies for the management of tuberculosis therapy: mapping the landscape and research priorities. *BMJ Global Health*, 3(5):e001018, 2018.
- [19] Habteyes Tola, Azar Tol, Davoud Shojaeizadeh, and Gholamreza Garmaroudi. Tuberculosis treatment non-adherence and lost to follow up among TB patients with or without HIV in developing countries: A systematic review. *Iranian Journal of Public Health*, 44(1):1–11, 06 2015.
- [20] Tao Tu, Anil Palepu, Mike Schaeckermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*, 2024.
- [21] U.S. Centers for Disease Control and Prevention (CDC). About chronic diseases. <https://www.cdc.gov/chronic-disease/about/>, accessed on Mar 19, 2025.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [23] World Health Organization (WHO). Global tuberculosis report. <https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023>, 2023.
- [24] Xiaoming Zhang, Daniel Lin, Hugh Pforsich, and Vernon W Lin. Physician workforce in the united states of america: forecasting nationwide shortages. *Human resources for health*, 18:1–9, 2020.

Chapter 2

**TOWARD LARGE LANGUAGE MODELS AS A  
THERAPEUTIC TOOL:  
COMPARING PROMPTING TECHNIQUES TO IMPROVE  
GPT-DELIVERED PROBLEM-SOLVING THERAPY**

## Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy

Daniil Filienko, BS,<sup>1</sup> Yinzhou Wang, MS,<sup>3</sup> Caroline El Jazmi, BS,<sup>1</sup> Serena Xie, MS,<sup>2</sup>, Trevor Cohen, MBChB, Ph.D.,<sup>2</sup>, Martine De Cock, Ph.D.<sup>1</sup>, Weichao Yuwen, Ph.D., RN<sup>1</sup>

<sup>1</sup>University of Washington, Tacoma, WA; <sup>2</sup>University of Washington, Seattle, WA;

<sup>3</sup>Dartmouth College, Hanover, NH

### Abstract

*While Large Language Models (LLMs) are being quickly adapted to many domains, including healthcare, their strengths and pitfalls remain under-explored. In our study, we examine the effects of prompt engineering to guide Large Language Models (LLMs) in delivering parts of a Problem-Solving Therapy (PST) session via text, particularly during the symptom identification and assessment phase for personalized goal setting. We present evaluation results of the models' performances by automatic metrics and experienced medical professionals. We demonstrate that the models' capability to deliver protocolized therapy can be improved with the proper use of prompt engineering methods, albeit with limitations. To our knowledge, this study is among the first to assess the effects of various prompting techniques in enhancing a generalist model's ability to deliver psychotherapy, focusing on overall quality, consistency, and empathy. Exploring LLMs' potential in delivering psychotherapy holds promise with the current shortage of mental health professionals amid significant needs, enhancing the potential utility of AI-based and AI-enhanced care services.*

### Introduction

Numerous studies have demonstrated the potential of Large Language Model (LLM) usage in medical applications, ranging from question-answering services to producing medical notes [1, 2]. Deployment of LLMs in AI-supported care services holds the potential to mitigate healthcare costs and broaden access to care. Companies are already offering products that attempt to undertake the roles of administrative and medical clinicians through the use of LLMs.<sup>1</sup> One of the fields that could greatly benefit from additional resources is psychotherapy, with as many as 20% of people worldwide needing mental health care [3] in the context of a global scarcity of mental health professionals [4].

Previously, studies have demonstrated the advantages of using relevant and empathetic responses in mental health dialogues [5, 6]. Althoff et al. [5], in particular, showed that more successful human counselors use fewer templated replies and produce varied responses to similar questions. LLMs' ability to generate coherent and contextually appropriate responses may provide an ideal tool for simulating such behavior, though concerns remain about their safety and utility as patient-facing tools. In this paper, we describe a study of LLMs' ability to provide such relevant and empathetic responses in real-time psychotherapy dialogues.

In this study, we explore the ability of off-the-shelf LLMs to deliver Problem-Solving Therapy (PST) for family caregivers, targeting common caregiving symptoms, such as fatigue and anxiety, through a dialogue system that provides caregivers the tools to self-monitor symptoms, problem-solve, and take appropriate actions. PST is an effective form of cognitive behavioral therapy with a specific protocol, making it a good test case for LLM performance, and providing the model with specific guidelines to follow.

We demonstrate that non-medical LLMs can display a reasonable performance in PST due to their remarkable ability to use the information provided in the prompt to enhance their performance [7]. We adapt an LLM flow that already showed good performance for medical Q&A [1] to medical dialogue generation. We explore the extent to which a general-purpose LLM such as GPT-4 can be improved by methods not involving modifying the model's weights and compare it with a previously developed human-curated rule-based system [8] in the context of PST for family caregivers with standardized patients portrayed by actors. Actors were provided sample responses and narratives describing the personas that they were playing in the interactions with the model to ensure consistent behavior. To compare the LLM performance on this domain-specific task meaningfully, we recruited multiple clinicians to evaluate the dialogues without knowing how the dialogues were created. Hence, our work has two primary objectives. The first is to investigate the use of prompt engineering methods to improve a general-purpose LLM's ability to deliver steps of

<sup>1</sup>see e.g. <https://www.hippocraticai.com/>

PST. The second is to provide comprehensive automated and human evaluations of the empathy and overall quality of the models’ output.

### **Related Work**

Existing work to improve the accuracy of LLMs for medical applications leverages (1) *fine-tuning* the model’s weights through further training on application-specific data and/or (2) developing novel ways to query the model to trigger a high-quality response by altering the prompt or including relevant information or examples in the prompt, a method called *prompt engineering*. Since for fine-tuning, the end results are a reflection of the quality of the fine-tuning corpora used, effective fine-tuning typically necessitates expensive manual review and in-domain expert curation to produce meaningful improvements in the model’s responses [9], in addition to highly expensive hardware necessary to run fine-tuning algorithms. In the present study, we focus on prompt engineering, a promising option that has been shown to produce comparable improvements without the resource-intensive fine-tuning [1]. To our knowledge, this is the first study on improving the accuracy of LLM delivering PST through prompt engineering [10]

**Usage of Large Language Models as Therapeutic Chatbots.** Literature on the usage of LLMs in psychotherapy is nascent. A few published studies have explored LLMs’ ability to lead a therapeutic conversation with a user [11, 12, 10]. Fu et al. described utilizing an LLM to augment a human therapist, providing helpful suggestions rather than singularly leading the conversation, i.e., keeping a human in the loop who can detect hallucinations generated by the LLM and increase the controllability of the system [12]. While such a study provides a setup well-suited for current LLMs, our work explores the current limits of the models’ performances in fully autonomous settings to explore the potential for full deployment. Wang et al. performed fine-tuning of a GPT-2 model to generate suitable PST responses [10]. However, their work did not explore prompt engineering methods that recently arose to prominence due to their effectiveness with larger models [1], producing results comparable to those of fine-tuned models [13]. A recent study [11] explored zero-shot prompt engineering to facilitate a diagnostic conversation based on the Diagnostic and Statistical Manual of Mental Disorders-5 standard [14]. They explored similar questions to ours, achieving an increase in the model’s empathy in responses via the use of zero-shot prompting. In this study, we expand on prompt engineering methods by using recently emerged techniques including Zero-Shot [15], Few-Shot [7], and Zero-Shot Chain-of-Thought [16], which we will refer to as Chain-of-Thought (COT) hereafter.

**Zero-shot prompting:** in this setup, a model receives only a natural language instruction to perform a task, without prior demonstrations. Relying on the model’s pre-trained knowledge, this approach maximizes convenience and potential for robustness, albeit typically being the most ineffective due to the absence of examples that could clarify the task’s format or expected output [15]. **Few-shot prompting:** in few-shot, a model is given a small number of input-output pairs as a guide. This approach enables LLMs to continue generating appropriate output for similar inputs without fine-tuning the model’s weights [7]. Optimal example selection is key in few-shot prompting. Brown et al. [7] demonstrated that by incorporating up to eight high-quality examples in prompts, GPT-3’s performance on various natural language benchmarks significantly improved [7]. **Chain-of-Thought (CoT):** CoT leads LLMs through a step-by-step reasoning process, effectively encouraging them to “think out loud”. Although the specific style can vary without drastically affecting performance, its presence is crucial for improving problem-solving capabilities [16]. This technique encourages the model to consider intermediate tokens, leading to more robust responses. CoT has been shown to significantly boost the problem-solving performance of LLMs across complex tasks, from arithmetic to commonsense challenges [16].

### **Methods**

We used various prompt engineering techniques and their combinations to improve LLM performance when delivering a PST session as a therapist bot. We were guided by findings from Nori et al. who showed that by utilizing various standard prompt engineering techniques, a generalist model, in particular state-of-the-art GPT-4, can respond to medical questions with accuracy on par with models specifically fine-tuned for this task [1]. We utilize the same methods and examine whether their insights can be applied to psychotherapy. While in this study we used GPT-4 through Microsoft Azure’s Application Programming Interface (API), at the time of writing, GPT-4 is also available to a wide audience as the backend of OpenAI’s ChatGPT conversational agent.<sup>2</sup>

In preliminary experiments, our team evaluated multiple prompting strategies in delivering PST and found large inconsistencies in expected model behavior. Models produced at times low quality or generally incoherent responses

<sup>2</sup><https://chat.openai.com/>

when performing full PST. To mitigate this, we decided to focus on the two PST steps that were the hardest to control and were most prone to producing unexpected and potentially harmful dialogue outputs. The two steps are *problem selection and identification* (in our case scenario for family caregivers, we named it *caregiving symptom assessment*) and *goal setting*. By merging our technical insights in prompt engineering with clinicians’ knowledge of PST, we designed prompts that are specifically relevant to these two steps of PST. As Van Veen et al. [2] have previously shown, human assessments of LLM performance do not always closely correlate with automated metrics, especially in domains requiring expert knowledge, such as healthcare. Hence, in addition to using an automated empathy evaluation method developed by Sharma et al. [6] to evaluate the models’ responses, we had a group of clinicians familiar with PST evaluate our dialogues without knowledge of whether the dialogue was LLM-derived. We provide more details about the prompt design and the evaluation process below.

### *Designing the prompts*

We started with a naive prompt shown in Table 1 and used it as a baseline. We then included the three prompt engineering techniques. We also experimented with combinations of these techniques for better downstream performance.

Type of Prompt	Content of the Prompt
Baseline	Your responsibility is to guide the conversation with a caregiver through the principles of PST to improve one significant symptom the caregiver is experiencing. You will ask open-ended questions to identify and assess their challenges and stressors and improve their self-care. Avoid focusing on the care receiver. Remember, your job is to help the caregiver. When the caregiver asks for goal suggestions by saying ‘Can you suggest some goals for me?’, take this as your cue to thoroughly review the conversation you’ve had with them. Concentrate on identifying their unique needs and aspirations as discussed. After this review, generate two concise, achievable, and personalized goals that directly address and support their expressed needs and aspirations. Ensure these goals are not only realistic but designed to inspire and boost the caregiver’s motivation.
Zero-Shot	In the process of identifying and assessing the caregiver’s symptoms, assess all five aspects if it has not been mentioned in the conversation: 1. Symptom Identification (What are the caregiver’s symptoms?) 2. Symptom Frequency (How often do the caregiver’s symptoms occur?) 3. Symptom Context (Where and when do the caregiver’s symptoms occur? Are there specific people present, or are certain activities involved?) 4. Symptom Severity (On a scale of 1 to 5, how severe are the caregiver’s symptoms?) 5. Previous Measures (What has the caregiver already tried to alleviate the symptoms?)
Few-Shot	Below are ideal dialogue examples illustrating how you, the assistant, should evaluate and address the challenges and stressors of caregivers, referred to here as the user, during conversations. **Attached would be 9 excerpts from multi-turn PST therapy with a human therapist**
Chain-of-Thought	Think about the user’s input step by step.

Table 1: Our prompts categorized by various types of prompting techniques used

**Baseline:** Leveraging insights from the RoleLLM framework [17], which highlights substantial improvements in language model performance through role-conditioned instruction tuning, we started by implementing a role-playing approach to augment our chatbot’s functionality as a PST assistant. This foundational prompt within our system prompt architecture, shown in Table 1, is composed of two parts which serve two distinct functions. First, it defines the chatbot’s role as an assistant to caregivers, delineating its core functions, tasks, and desired behaviors to facilitate conversations aligned with PST principles, aiming at the identification and analysis of caregivers’ symptoms. Second, it ensures proper behavior of the model during the goal setting stage, encouraging it to provide proper goals. **Zero-Shot:** To enhance the chatbot’s ability to generate structured dialogue, following the PST protocol, we built upon the baseline prompt with the instructions shown in the second row of Table 1, reflecting the sub-stages necessary for accomplishing symptom identification and assessment, as well as the goal suggestion steps of PST. Central to our strategy is including precise language explaining five predefined criteria necessary for the successful identification and assessment of caregiver symptoms: Symptom Identification, Symptom Frequency, Symptom Context, Symptom

Severity, and Previous Measures [8]. This approach is an attempt to enhance the depth and relevance of the chatbot’s interactions, grounding its dialogue framework in these five essential aspects of the symptom identification and analysis step in PST. **Few-Shot:** Our implementation of the few-shot prompting method incorporates nine high-quality examples curated by clinicians as in-context learning examples to capture the complexities of symptom identification and analysis within the PST framework [18]. These examples aim to demonstrate to the model a set of ideal responses to realistic user inputs, providing comprehensive coverage of the symptom identification and analysis phase. The first six examples present single-turn dialogues addressing the 5 sub-steps of PST that are listed in the zero-shot prompt (see the second row in Table 1), spanning symptom identification and assessment, including frequency, context, severity, and prior mitigation efforts. Additional examples featuring multi-turn dialogues are designed to enhance the chatbot’s capability to conduct coherent and contextually relevant conversations within PST. The multi-turn dialogue examples are essential for demonstrating to the chatbot the intricacies of ongoing therapeutic dialogue, including the need for follow-up questions and the reiteration of the points made to the user during the dialogue. **Chain-of-Thought:** Our implementation of CoT is an attempt to systematically improve the chatbot’s analytical and problem-solving capabilities by forcing it to be more ‘thoughtful’ in its style of output, which can lead to better outcomes as demonstrated in previous studies [1, 15].

Model	Prompt Structure	Preliminary Assessment	Evaluation
0	Baseline	Pass	Proceed
1	Baseline + Zero-Shot	Fail	N/A
2	Baseline + Few-Shot	Pass	Proceed
3	Baseline + Zero-Shot + Few-Shot	Pass	Proceed
4	Baseline + Zero-Shot + Few-Shot + CoT	Pass	Proceed
5	Baseline + Few-Shot + CoT	Pass	Proceed
6	Baseline + Zero-Shot + CoT	Fail	N/A
7	Human-curated rule-based system	Pass	Proceed

Table 2: Overview of Model Structures

### Model Development

We developed the models gradually, each step aimed at mitigating the faults noticed at the previous stage (Table 2). We produced models of various complexity, which we then evaluated for their ability to achieve the intended objectives.

**Model 0:** A baseline model, referred to as Model 0, introduces the chatbot in its role of guiding caregivers through step 3, “Identify and Assess Symptom,” and step 4, “Goal Setting,” following the principles of PST. **Model 1 and Model 2:** Expanding on Model 0, Models 1 and 2 improve the chatbot’s ability to follow PST structure by introducing key prompt engineering techniques. Model 1 builds upon Model 0 by integrating a set of structured guidelines into the prompt via zero-shot prompting. This approach systematically guides the chatbot into assessing caregiver’s challenges and stressors in a specific order. Model 2 further advances Model 0 through the integration of few-shot learning, which leverages selected single-turn and multi-turn example dialogues from previously recorded sessions to set guidelines for the conversation flow. **Model 3 through 6:** In our further explorations, inspired by the methodology of Nori et al. [1], we utilize various combinations of prompt engineering techniques across Models 3 to 6. Model 3 integrates zero-shot’s structured symptom assessment with few-shot learning to produce more robust output that more closely follows the intended response style. Model 4 incorporates CoT prompting to improve the models’ problem-solving results. Models 5 and 6 adjust these components to explore their combined effects on the chatbot’s PST performance, with a focus on more effective goal setting. **Model 7:** While not a model per se, we also curated four dialogues from the rule-based chatbot baseline to compare against our methods following our previous work [8].

### Generating Dialogues for Evaluation

To achieve a higher level of consistency and protect caregivers’ privacy, we used personas to generate dialogues. We first crafted four caregiver personas with exemplary replies based on our prior work with family caregivers. We recruited three research team members with direct professional experiences interacting with family caregivers. They were asked to portray one or two of the personas and converse with each bot therapist (Model 0 through 6 in Table 2). They were instructed to use consistent and exemplary replies as much as possible. We collected 28 PST dialogues (4 personas \* 7 models). We randomized the sequence of the models and did not disclose which specific models they

Dimensions	Statements or Questions	Measurements
Symptom assessment	The therapist successfully assessed the five aspects of the caregiver's symptoms.	5-point Likert scale (1 = Completely disagree; 5 = Completely agree)
Goal setting	The goal suggested by the therapist is appropriate for the caregiver.	
Emotional Reactions	The therapist expressed emotions such as warmth, compassion, concern, or similar feelings towards the caregivers.	3-point Likert scale (0 - therapist not expressing them at all; 1 - therapist expressing them to some weak degree; 2 - therapist expressing them strongly.) We used the framework from Sharma et al. [6] to evaluate peer-to-peer dialogues for both algorithmic and human evaluation.
Interpretations	The therapist communicated an understanding of feelings and experiences inferred from the caregiver's responses.	
Explorations	The therapist improved their understanding of the caregiver by exploring the feelings and experiences not stated in the caregiver's response.	
Overall	Overall, how was this therapy session?	5-point Likert scale (1 - very bad; 5 - very good)

Table 3: Evaluation components

interacted with. After we collected all the dialogues, the research team examined the conversations and found that Models 1 and 6 had at least one dialogue in which the bot therapist asked all five symptom assessment questions at once. When the caregiver responded to one of the questions, the bot therapist could not recognize that one question had been answered and repeated all the questions. This behavior was undesired and would automatically receive low ratings in the human evaluation. Thus, we eliminated them in this preliminary step prior to human evaluation. The models that proceeded to human evaluation were Models 0, 2, 3, 4, and 5. In addition, the research team curated a set of four dialogues with human-delivered PST sessions. These dialogues were curated based on the four personas and the research team's prior rule-based bot-delivered PST protocol [8]. In total, we had 24 dialogues (6 models \* 4 personas) that proceeded to human evaluation.

#### ***Evaluating Dialogues - Human Evaluation***

We recruited seven clinicians, namely four nurses and three clinical psychologists, to evaluate the quality of the therapy responses. The evaluators were unaware of the models and evaluated the dialogues in random sequence. Evaluators evaluated each dialogue on two aspects, conversational quality and perception of relational capacity, which were adopted from the chatbot evaluation mechanisms by Zhang et al. [19]. For conversational quality, since the generated dialogues specifically focused on the "symptom assessment" and "goal setting" steps of PST, the evaluators were asked to evaluate the quality of these two steps. To evaluate the symptom assessment step, we asked the evaluators to consider if the therapist assessed all five aspects of symptoms. To evaluate the goal-setting step, evaluators were asked to assess the appropriateness of the goals suggested by the therapist in the dialogue. For the perception of relational capacity, we focused on therapist empathy, which was shown to be a predictor for therapy outcome [20]. We employed the three communication mechanisms to measure empathy developed by Sharma et al. [6]: Emotional Reactions (ER), Interpretations (IN), and Explorations (EX). Strong empathetic communication expresses emotions reacting to what the user said (Emotional Reaction), communicates an understanding of the user's feeling or experience (Interpretation), and explores the user's feelings and experiences that are not stated (Exploration) [6]. Evaluators evaluated all three aspects. Additionally, we asked evaluators to provide a rating based on the overall impression of the therapy session. Details about the evaluation components and questions are included in Table 3. Moreover, to gain a deeper understanding of the ratings and what factors contributed to the ratings, we also asked the evaluators to provide a brief rationale for each rating.

#### ***Evaluating Dialogues - Automatic Evaluation***

In addition to human evaluation, we adapted the algorithm from Sharma et al. [6] to rate the conversations' empathy.

We used the default implementation of the algorithm provided on the authors’ GitHub repository, a base RoBERTa classifier trained on Reddit dialogue data labeled by the authors. The classifier quantifies the model’s Emotional Reactions (ER), measuring its ability to express positive emotions when responding to a user’s post; Interpretations (IP), evaluating the model’s ability to produce a relevant shared experience; and Explorations (EX), which captures a level of active interest and engagement with the user’s post.

### Data Analysis

Each of the six models that generated dialogues received 28 sets of ratings (4 persona-based dialogues \* 7 evaluators). We computed the averages and standard deviations of the ratings for each model. We report on the models with the highest and lowest ratings on each evaluation component. Three team members conducted a rapid deductive analysis of the qualitative rationales provided by the evaluators to identify specific factors contributing to the ratings.

### Results

Model	Symptom Assessment (1-5)	Goal Setting (1-5)	Emotional Reactions (0-2)		Interpretations (0-2)		Explorations (0-2)		Overall (1-5)
	Human	Human	Human	Algorithm	Human	Algorithm	Human	Algorithm	Human
0	3.15 (1.35)	4.00 (1.09)	1.56 (0.58)	0.83 (0.17)	1.59 (0.64)	0.04 (0.06)	1.50 (0.69)	0.79 (0.21)	3.68 (1.06)
2	4.68 (0.55)	4.74 (0.45)	1.71 (0.46)	0.93 (0.08)	1.64 (0.49)	0.06 (0.10)	1.39 (0.74)	1.22 (0.08)	4.36 (0.73)
3	4.75 (0.44)	4.46 (0.74)	1.57 (0.57)	0.92 (0.19)	1.68 (0.55)	0.11 (0.02)	1.21 (0.74)	1.29 (0.27)	3.82 (0.90)
4	4.80 (0.41)	4.25 (0.93)	1.68 (0.48)	0.91 (0.18)	1.75 (0.44)	0.02 (0.03)	1.32 (0.61)	1.31 (0.16)	3.82 (0.79)
5	3.70 (1.35)	3.86 (1.38)	1.64 (0.49)	0.93 (0.25)	1.71 (0.46)	0 (0)	1.35 (0.69)	0.76 (0.18)	3.61 (1.13)
7	4.62 (0.5)	3.59 (1.01)	1.22 (0.80)	0.89 (0.14)	1.22 (0.89)	0.04 (0.06)	0.77 (0.65)	1.31 (0.13)	3.19 (1.18)

Table 4: Evaluation results. The results are reported in Mean (Standard Deviation) format.

Table 4 shows the mean scores with standard deviations of both the human and automated evaluations. In general, Models 2, 3, and 4 were rated higher compared to the human-curated rule-based system (Model 7) in all evaluated areas and delivered the two steps of PST successfully. Overall, evaluators rated Model 2 the best, which is also the highest-rated model for goal-setting and the exploration dimension of empathy. The human-curated rule-based model was rated the lowest across all evaluated aspects except for symptom assessment.

**Symptom assessment.** For symptom assessment, the improved models (Models 2-5) were all able to assess a caregiver’s symptom in at least four turns of dialogue. When the “caregiver” shared a challenge such as “*I’m feeling overwhelmed caring for my father with cancer and looking for support,*” the bot therapist would first provide an empathetic response, followed by a therapeutic component that continued the PST process. For example, one response by a bot to this specific “caregiver” was, “*Hello, I’m really glad you reached out for support. It’s incredibly important to take care of yourself, especially when you’re in a caregiving role. It sounds like you’re going through a lot right now. Can you tell me more about what specifically is making you feel overwhelmed?*” The average rating for symptom assessment across the four models was 4.5 out of 5, indicating that “the therapist successfully assessed the five aspects of the caregiver’s symptoms”. Model 4, with all prompt engineering techniques, scored the highest for symptom assessment (mean = 4.80). The majority of the evaluators thought Model 4 successfully assessed all aspects of the caregiver’s symptoms and had good symptom identification. Model 0 using only baseline prompting was rated the lowest. Evaluators reported that this bot missed assessing certain aspects of caregivers’ symptoms, such as severity and frequency, and had poor follow-up questions on context. Moreover, several evaluators mentioned that the bot asked multiple questions at a time and made it overwhelming. Tables 5 and 6 show example bot responses from the models with the highest and lowest rating for each criterion, including symptom assessment.

**Goal setting.** When suggesting health goals for the caregivers to improve an identified symptom, Model 2 was rated the best (mean = 4.74). Evaluators’ feedback on Model 2 included keywords “personalized”, “collaborative”, “realistic”, “appropriate”, “specific”, “straightforward”, “reasonable”, “relevant”, “well-structured”, and “detailed”. Model 7 (i.e. human-curated rule-based model) was rated the lowest (mean=3.59). Evaluators pointed out a few drawbacks including being “not clear”, “not specific”, “superficial”, and “providing only one suggestion”. Sample

Evaluation Aspect	Human Rating	Model	Example Chatbot Response
Symptom Assessment	Highest (mean=4.80)	4	“It sounds like you’re dealing with some intense emotions. I want to acknowledge how challenging it can be to manage fear and sadness, especially in a caregiving role. To better understand and help you with these feelings, could you tell me how often you experience these feelings of fear and sadness?”
	Lowest (mean=3.15)	0	“It sounds like you’re carrying a heavy load both emotionally and practically. Let’s address these one at a time. Firstly, regarding emotional support, can you tell me about your current support system? Do you have friends, family, or a community you can talk to about what you’re going through? Secondly, for the practical help with caregiving tasks, have you been able to identify specific tasks that are particularly challenging or time-consuming? Are there routine aspects of care that you think someone else could assist you with?”
Goal Setting	Highest (mean=4.74)	2	“Excellent. Since you’re open to online support groups and you’re also looking to enhance your emotional support, here are two personalized goals that might be beneficial for you: 1. <b>**Join an Online Support Group**</b> : Within the next week, aim to identify and join at least one online support group for caregivers. Spend a few minutes each day, or as your schedule allows, to engage with the group. This can be a place where you can share your experiences, receive advice, and connect with others who understand what you’re going through. 2. <b>**Schedule Regular Check-ins**</b> : Set up a weekly check-in with a friend or family member who can offer emotional support. This could be a phone call, video chat, or even a text exchange, where you can share your feelings and challenges. Having a set time each week for emotional check-ins can provide you with something to look forward to and ensure you have a consistent outlet for your emotions. How do these goals sound to you? Are they something you feel you can commit to?”
	Lowest (mean=3.59)	7	“Thanks for being open to it. Which of the two would you like to try for this week? 1. Journal writing. 2. Practice the 4-7-8 deep breathing exercise technique .”

Table 5: Symptom Assessment and Goal Setting Examples

chatbot responses from Model 2 and Model 7 for goal setting are included in Table 5. In the same persona mentioned above, one bot therapist made the following goal recommendation: *“Let’s work on setting some goals that can help you manage these overwhelming feelings more effectively. Based on what you’ve shared, here are two personalized goals that you might consider: 1. Establish a Daily Unwinding Ritual: Given that your feelings of fear and sadness intensify during the evenings and nights, creating a calming evening routine could be beneficial. This might include activities like meditation, listening to soothing music, or practicing gentle yoga. The goal would be to engage in this ritual every evening for at least 15 minutes to help transition into a more relaxed state of mind before bedtime. 2. Seek Emotional Support: It’s important to have a support system in place. You could set a goal to reach out to a friend, family member, or a support group for caregivers at least once a week. Sharing your experiences and feelings with others who understand can provide comfort and may offer new perspectives or coping strategies. How do these goals sound to you?”*

**Empathy.** For the three dimensions of empathy, the human-curated rule-based model was rated consistently as the lowest across all three dimensions, with other models being rated similarly. Model 2 scored the highest in ER (mean = 1.71), followed by Model 4 (mean = 1.68). One evaluator expressed approval of the emotional reactions exhibited by the therapist bot powered by Model 2, stating *“I think the empathy came across better in this one than some others because it made a statement and then asked a question. It didn’t pile questions on top of questions like in some of the prompts... it makes the conversation seem less empathetic and two sided.”* Model 4 was rated the highest in the IP (mean = 1.75). The majority of the evaluators expressed that there was good use of empathetic statements and reflections. One evaluator provided an example of *“an excellent follow-up on feeling/experiences and insights into*

Evaluation Aspect	Human Rating	Model	Example Chatbot Response (Algorithm rating on a scale of 0-2)
Emotional Reactions	Highest (mean=1.71)	2	“That must be quite exhausting for you. Not getting enough sleep can certainly exacerbate stress and pain. To better understand your situation, can you tell me more about the context of your insomnia? For example, are there specific thoughts or worries that keep you awake, or is it the pain that’s more disruptive?” (Algorithm rating: 1)
	Lowest (mean=1.22)	7	“I can see that. I have one more question to ask about your overwhelmed feeling. Have you done anything to help you with your overwhelmed feeling?” (Algorithm rating: 1)
Interpretations	Highest (mean=1.75)	4	“That sounds really tough, being at home and feeling that heightened awareness of your daughter’s anxiety and the challenges in connecting with her.” (Algorithm rating: 0)
	Lowest (mean=1.22)	7	“I hear you. Let’s use the Problem Solving Therapy (PST) process that I mentioned to you to explore how we may address this.” (Algorithm rating: 0)
Explorations	Highest (mean=1.50)	0	“That sounds quite challenging and must be affecting your daily functioning. Not being able to sleep well for 4-6 nights a week is significant. Can you share more about the context of these sleepless nights? For example, are there specific thoughts related to your daughter’s illness or your pain that tend to keep you awake?” (Algorithm rating: 2)
	Lowest (mean=0.77)	7	“Great. Let’s use the Problem Solving Therapy (PST) process that I mentioned to you to explore how we may address this.” (Algorithm rating: 2)

Table 6: Emotional Reactions, Interpretations, and Explorations Examples

the feelings” in which the bot responded “*It can be harder at night when it’s quiet.*” For the Exploration dimension of empathy, evaluators rated Model 0 the highest (mean = 1.50). One evaluator expressed that “*In this session, the therapist does more open ended questions that get to more of the root of the problem.*” For the lowest rated human-curated rule-based model across all three dimensions of empathy, evaluators reported that the bot said things like “*got it*” that demonstrated poor empathy and regurgitated sentences like “*I am so sorry to hear that ...*”, which made it feel robotic. Several evaluators also reported that the human-curated rule-based model displayed “zero or very limited explorations of deeper thoughts and feelings.” Moreover, the bot was perceived as “fake and too cheerful.” One evaluator said that “*It used the word ‘glad’ very often and also thanked the caregiver quite often. It’s important for the bot to in some ways mirror the caregiver, meet them where they are at. And someone worrying 4 out of 5 may not want it to seem overly cheerful.*” According to the automated evaluation on empathy, all models performed very similarly in terms of ER, with an average close to 1, meaning that the therapist expressed emotional reactions to some weak degree. Rarely were our models nor the human dialogue able to demonstrate Interpretations as all models were scored close to 0 by the automated evaluation algorithm. The majority of the models, except for Model 0 and Model 5, demonstrated moderate empathy in Explorations, with Models 4 and the human-curated rule-based model scored the highest (mean = 1.31) by the automated evaluation.

## Discussion

In this study, we used various prompting techniques to improve GPT-delivered PST for family caregivers and used both human and automatic algorithms to evaluate the therapy dialogues. We found that by using prompt engineering techniques, we were able to improve the quality of the therapy conversations beyond the baseline prompt, but with considerable limitations. Empathy evaluated by both human and the algorithm did not vary significantly across models, despite improving over the baseline in the emotional reaction and interpretation dimensions. Below, we discussed specific prompt engineering techniques and their performances.

Zero-shot learning did not perform as well as the other techniques that we evaluated. Our finding that explicit directions may not be sufficient to adapt a model to a domain-specific task is in line with the literature [1, 7]. In our scenario, the

difference in performance between zero-shot learning and the other techniques was more pronounced because of the nature of the downstream task, i.e. PST being a protocolized therapy. PST requires a specific way of conversing with the user. Zero-shot prompting focuses on explicitly defining tasks for the model to follow, but many aspects of what constitutes good therapy are implicit (e.g. more actionable advice being preferred over more generic, overly optimistic comments being preferred less [6]). The possible approaches to achieve the explicit goals are almost endless. Explicitly demonstrating examples of high-quality PST responses as part of few-shot prompt led to better model performance, showing that providing the model with data is still necessary to produce coherent dialogues. Our findings align with previous results from the literature, that show that providing few-shot examples typically noticeably increases the models’ performance compared to zero-shot baselines on tasks involving the generation of coherent word sequences, akin to how PST demands a coherent dialogue [7].

Adding CoT resulted in better empathy, especially exploration. However, it reduced the quality of symptom identification and goal setting. The original CoT work by Wei et al. [16] did not explore dialogue-generating tasks, and it is possible that we could experiment with alternative CoT prompts that would improve empathy while maintaining task quality. However, given that LLMs are fundamentally next-word predictors trained on large corpora of text, it is possible that in the pre-training data of the model, chunks of text beginning with phrases similar to our prompt happened to be more exploratory and less actionable. We find that explanation reasonable because, in human writings, texts nudging someone to think tend to be more contemplative than action-driven in nature.

In line with previous work [2], we found low agreement between the human and automatic evaluations. This may be partially due to the model by Sharma et al. [6] trained on single-turn Reddit posts, presenting a different setting than PST. Future studies should examine the generalizability of the automatic evaluation algorithms with domain shifts.

Although the LLM-delivered therapy dialogues received higher scores than the human-curated baseline, it is worth noting that the human baseline is not actual therapist-delivered PST dialogues. We adapted dialogues the team created for developing a rule-based chatbot to deliver PST to family caregivers. Some responses were designed to be generic such as using *“I am sorry to hear that”* instead of a more empathetic response using psychotherapeutic techniques such as normalizing (e.g., *“I am sorry to hear that. Many caregivers feel isolated while caring for an ill child.”*). In our study, to protect real caregivers’ privacy, we used a persona-based approach to collect dialogues. It would be preferable to engage with actual caregivers. In the future, we plan to develop privacy-preserving technologies in order to guarantee the privacy of the user’s information when training or interacting with LLMs, so that our in-context learning processes do not expose users’ private information to commercially-hosted LLMs. This study intended to explore the ability of off-the-shelf LLMs with prompt engineering to deliver part of PST and achieved promising results. In the future, we will continue to improve the models to deliver full PST by performing fine-tuning and Retrieval Augmented Generation (RAG) techniques. In the aforementioned reading comprehension task [7], we can see that fine-tuned models still can outperform generalists with few shot examples. However, recent studies [1, 13] showed that it is not always true, possibly due to the increased knowledge contained in the larger models, allowing them to perform on-par with or better than fine-tuned models. We will continue to explore the best techniques or combinations of techniques that enable an automated chatbot to deliver a full PST session.

## Conclusion

In this paper, we adapt multiple novel prompt engineering approaches to improve an LLM’s ability to deliver part of a psychotherapy session. Consistent with previous findings [11], we demonstrate that the model’s capability to deliver protocolized therapy can be improved with the proper use of prompt engineering methods, albeit with limitations. Through both automatic and human evaluation, we show an improvement over the baseline model after applying our methods to PST, demonstrating that some prompt engineering techniques are better at improving the performance of the models than others. Hence, while the current models cannot be deployed directly in psychotherapy settings without human oversight, this work contributes to the effort in exploring the potential of LLMs as a therapeutic tool. As such, this work represents an important step toward using LLMs to address the limited availability of human therapists in the context of an escalating need for mental health services.

**Acknowledgments.** The authors would like to thank Microsoft for the UW Azure Cloud Computing Credits for Research program. Daniil Filienko is a Carwein-Andrews Distinguished Fellow. This research was, in part, funded by the National Institutes of Health (NIH) Agreement No.1OT2OD032581 and R21NR020634, and the Rita and Alex Hilman Foundation Emergent Innovation Grant. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

## References

1. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *CoRR*. 2023;abs/2311.16452.
2. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. *Nature Medicine*. 2024;(4):1134-42.
3. Holmes E, Ghaderi A, Harmer C, Ramchandani P, Cuijpers P, Morrison A, et al. The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet*. 2018 Mar;5(3):237-86.
4. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. *The Lancet*. 2018;392(10157):1553-98.
5. Althoff T, Clark K, Leskovec J. Natural Language Processing for Mental Health: Large Scale Discourse Analysis of Counseling Conversations. *Transactions of the Association for Computational Linguistics*. 2016 05;4.
6. Sharma A, Miner A, Atkins D, Althoff T. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In: Webber B, Cohn T, He Y, Liu Y, editors. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics; 2020. p. 5263-76.
7. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*. vol. 33; 2020. p. 1877-901.
8. Kearns WR, Kaura N, Divina M, Vo C, Si D, Ward T, et al. A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health Counseling Dialog. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. Association for Computing Machinery; 2020. p. 1-9.
9. Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, et al. LIMA: Less Is More for Alignment. In: *Advances in Neural Information Processing Systems*. vol. 36; 2023. p. 55006-21.
10. Wang L, Mujib MI, Williams J, Demiris G, Huh-Yoo J. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *ArXiv*. 2021;abs/2107.13115.
11. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation. *CoRR*. 2023;abs/2305.13614.
12. Fu G, Zhao Q, Li J, Luo D, Song C, Zhai W, et al. Enhancing Psychological Counseling with Large Language Model: A Multifaceted Decision-Support System for Non-Professionals. *arXiv preprint arXiv:230815192*. 2023.
13. Ovadia O, Brief M, Mishaeli M, Elisha O. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. *ArXiv*. 2023;abs/2312.05934.
14. *Diagnostic and Statistical Manual of mental disorders : DSM-5™*. 5th ed. Washington, DC: American Psychiatric Publishing, a Division of American Psychiatric Association; 2013.
15. Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, et al.. Finetuned Language Models are Zero-Shot Learners; 2022. 10th International Conference on Learning Representations (ICLR).
16. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. vol. 35; 2022. p. 22199-213.
17. Wang ZM, Peng Z, Que H, Liu J, Zhou W, Wu Y, et al. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In: *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics; 2024. p. 14743-77.
18. Yuwen W, Chang J, Divina M, Fan X, Kearns W, Peredo A, et al. Comparing Caregiving Needs in Asian and White Family Caregivers through a Journaling Exercise Delivered by a Conversational Agent. *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*. 2022;2022:1208-16.
19. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint. *Journal of Medical Internet Research*. 2020 Sep;22(9):e22845.
20. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*. 2018 Dec;55(4):399-410.

Chapter 3

**LARGE LANGUAGE MODEL-POWERED  
CONVERSATIONAL AGENT DELIVERING  
PROBLEM-SOLVING THERAPY (PST)  
FOR FAMILY CAREGIVERS:  
ENHANCING EMPATHY AND THERAPEUTIC ALLIANCE  
USING IN-CONTEXT LEARNING**

## **Large Language Model-Powered Conversational Agent Delivering Problem-Solving Therapy (PST) for Family Caregivers: Enhancing Empathy and Therapeutic Alliance Using In-Context Learning**

### **Liying Wang**

Institute on Digital Health and Innovation, College of Nursing, Florida State University, Tallahassee, USA, lw24y@fsu.edu

Center of Population Sciences for Health Equity, College of Nursing, Florida State University, Tallahassee, USA, lw24y@fsu.edu

### **Daffodil Carrington**

School of Nursing, University of Washington, Seattle, WA, USA, dalinson@uw.edu

### **Daniil Filienko**

School of Engineering and Technology, University of Washington, Tacoma, WA, USA, daniilf@uw.edu

### **Caroline El Jazmi**

Department of Computer Science, University of Texas at Austin, Austin, TX, USA, eljazmi@utexas.edu

### **Serena Jinchen Xie**

School of Medicine, University of Washington, Seattle, WA, USA, jinchenx@uw.edu

### **Martine De Cock**

School of Engineering and Technology, University of Washington, Tacoma, WA, USA, mdecock@uw.edu

### **Sarah Iribarren**

Department of Biobehavioral Nursing and Health Informatics, University of Washington, Seattle, WA, USA, sjiribar@uw.edu

### **Weichao Yuwen**

School of Nursing & Healthcare Leadership, University of Washington Tacoma, Tacoma, WA, USA, wyuwen@uw.edu

Family caregivers often face substantial mental health challenges due to their multifaceted roles and limited resources. This study explored the potential of a large language model (LLM)-powered conversational agent to deliver evidence-based mental health support for caregivers, specifically Problem-Solving Therapy (PST) integrated with Motivational Interviewing (MI) and Behavioral Chain Analysis (BCA). A within-subject experiment was conducted with 28 caregivers interacting with four LLM configurations to evaluate empathy and therapeutic alliance. The best-performing models incorporated Few-Shot and Retrieval-Augmented Generation (RAG) prompting techniques, alongside clinician-curated examples. The models showed improved contextual understanding and personalized support, as reflected by qualitative responses and quantitative ratings on perceived empathy and therapeutic alliances. Participants valued the model's ability to validate emotions, explore unexpressed feelings, and provide actionable strategies. However, balancing thorough assessment with efficient advice delivery remains a challenge. This work highlights the potential of LLMs in delivering empathetic and tailored support for family caregivers.

CCS CONCEPTS • Human-centered computing • Human computer interaction (HCI) • Interaction paradigms • Natural language interfaces

**Additional Keywords and Phrases:** Caregivers, Mental Health, Conversational Agent, Problem Solving Therapy, Behavioral Chain Analysis, Few-Shot Learning, Retrieval Augmented Generation

### **Reference style**

- ACM Reference Formatting : ACM SIGCHI Proceedings

## 1 INTRODUCTION

Family caregivers in the US reached 53 million in 2024 [14]. Due to the complex medical and psychosocial needs of the patients, family caregivers often struggle to fulfill the demands from multiple facets of their identity of being a caregiver, a parent, a daughter, a son, a spouse, and an employee. The multi-faceted stress from the role conflicts as a caregiver, coupled with the lack of resources to effectively cope with such stress, is often associated with adverse mental health outcomes for caregivers themselves. The resources may include time, financial support, social support, mental health literacy, and existing adaptive coping strategies. According to the Family Caregiver Annual Report, 43% of family caregivers report having sleep difficulties, and about 36% experience symptoms of depression [14]. During the COVID-19 pandemic, 57.2% of caregivers screened positive for symptoms of an anxiety or depressive disorder which was three times the prevalence compared to non-caregivers [11].

Mental health professionals are in shortage in the US and around the world. In the US, there are only 30 clinical psychologists and 16.6 psychiatrists for every 100,000 people, and in rural areas, these numbers are lower [27]. More than one-third of the US population (122 million) live in mental health professional shortage areas [29]. Due to the severe shortage, receiving mental health care often takes months of waiting and the cost sometimes is not fully covered by insurance plans. Family caregivers, in addition to the waiting time and high cost, face additional challenges including: logistical challenges in managing personal work schedule, caregiving schedule, and other family responsibilities such as parenting; low priority of self-care in the face of competing demands; feeling guilty or shame when taking time for oneself.

Studies have reported people tend to use large language models (LLMs) as mental health resources precisely due to the on-demand nature and the flexibility it offers, in contrast to the weekly hourly session that needs to be pre-scheduled with a mental health professional [20]. LLMs, with their language ability and context awareness, have been found to be empathetic and capable of providing advice and suggestions that are context-appropriate [2, 9]. However, LLMs have risks of hallucinations, generating responses that are potentially harmful, and lack of probing and assessment prior to providing advice, which then leads to responses that are not customized to the user, reducing engagement and effectiveness [22].

Contextual understanding is especially crucial to provide tailored support to caregivers with complex needs. In clinical practice, a thorough assessment is the foundation of appropriate interventions and can deepen provider understanding of patients' symptoms (severity, duration, frequency), trajectories, triggers, and self-management strategies. For example, a thorough understanding of a caregiver's day-to-day schedule can enhance the feasibility of the strategies by taking into account the amount of time it takes to implement the recommendations. Indeed, previous research found that caregivers found a chatbot delivering problem-solving therapy (PST) to be empathetic and overall helpful, but lacked contextual understanding and tended to jump to advice-giving [16]. This study, building upon previous work, aims to continue to improve the chatbot's contextual understanding of caregivers before providing advice, which would contribute to perceived empathy, therapeutic alliance, and tailored advice-giving. A within-subject experiment was conducted with 28 participants, where each participant tested 4 LLM models and evaluated the model performance on empathy and therapeutic alliance.

## 2 RELATED WORK

### 2.1 Digital interventions for mental health among family caregivers

Digital mental health tools developed for supporting caregivers include mobile apps, web-based mental health tools, virtual reality, and telephone-delivered interventions [47]. The majority of the digital tools focus on providing coping skills training, emotion regulation skills, and educational materials around mental health. Studies support the potential of digital tools in increasing caregiver coping skills, emotion regulation, motivation to self-care, and decreasing stress and burden [36]. There are several limitations in the existing digital mental health tools for caregivers. First, there is little room for customization and adjusting to the complexity and dynamic nature of the caregiving role. For example, a caregiver of a

cancer patient who is at the beginning of their cancer treatment versus at the end of their life may experience different stages of their role accordingly [17]. Second, most existing digital mental health tools heavily rely on Cognitive Behavioral Therapy (CBT). While CBT is an evidence-based therapy for multiple mental health conditions, it may not be the most effective therapeutic framework for caregivers [34]. CBT typically heavily emphasizes cognitive strategies that take time and practice, which may be particularly challenging for caregivers who are already overwhelmed with their caregiving responsibilities and have limited time and mental bandwidth [19, 21, 31]. The cognitive restructuring exercises in CBT, while valuable, may feel like an additional burden rather than immediate relief for caregivers dealing with practical day-to-day challenges. In contrast, Problem-Solving Therapy (PST) offers a more straightforward, action-oriented approach that aligns better with caregivers' needs [15]. PST's focus on identifying concrete problems and developing practical solutions makes it more accessible and immediately applicable to caregivers' situations [13]. The skills learned through PST are also easier to generalize across different caregiving challenges, providing caregivers with a flexible toolkit that can be adapted as their responsibilities evolve.

## **2.2 Applications of LLMs in mental health**

LLM-based chatbots, trained on large amounts of text data, have demonstrated the capability of generating human-like language and contextual understanding. These chatbots have expanded the possibilities of delivering mental health support, compared to previous rule-based and learning-based chatbots [7]. There is an increasing number of studies on the application of LLMs in mental health. LLMs demonstrated capabilities in assisting in the diagnosis of mental health issues, depression management, suicide risk evaluation, supporting psychiatric education, and providing accessible, de-stigmatized mental health support [20, 33]. Empathy is a key clinical element that facilitates the therapeutic effects of mental health interventions and strengthens therapeutic alliance (i.e., a collaborative and trusting relationship between a client and a therapist that enables effective communication and goal-setting during the therapeutic process) [32]. LLMs exhibited the core elements of cognitive empathy, including emotion recognition and providing emotionally validating responses aligned with the contexts [42]. LLMs also demonstrated higher emotional awareness and provided more empathetic responses, compared to human benchmarks [8, 45]. The majority of LLM-based chatbots developed for caregivers focused on developing information support for family caregivers. For example, studies reported that LLMs such as GPT can provide readable and reliable information that enhanced caregivers' knowledge of Alzheimer's disease-related dementias, pediatric oncology, and rehabilitation concerns and education needs of stroke survivors and caregivers. [30, 35, 39]. One LLM-based chatbot was fine-tuned with PST session scripts and was tested on a group of family caregivers, but the study was more focused on the sentiment of chatbot responses than its ability to effectively deliver mental health support [44].

## **2.3 Integrating PST and Motivational Interviewing (MI) in LLM-delivered mental health support**

PST has shown effectiveness in many mental health interventions, including for depression, anxiety disorder, PTSD, and substance use [10, 43]. In addition, PST interventions were found effective in releasing caregiver burden and enhancing coping, self-efficacy, depression, and quality of life among family caregivers [18, 25]. Kern et al.'s [47] work on designing tools for caregivers began with a rule-based PST therapy chatbot, which would give a response from a set of human-curated options based on the user's demonstrated need [48]. Filienko et al.'s [16] work extended this rule-based algorithm to utilize transformer-based LLMs. This was motivated by the findings that variety and creativity in human response were associated with more successful human counselors [49]. Therefore, LLMs which have demonstrated promising results in their ability to adapt to the provided information and lead empathetic dialogues [5, 16], provide a powerful new technology for generating more engaging and effective psychotherapy dialogues. The step-by-step approach of PST can steer LLMs to generate structured guidance, while the flexibility of LLMs allows for the incorporation of other therapeutic frameworks alongside PST. In the work by Filienko et al. [16], qualitative feedback highlighted two key gaps in LLM-based therapy chatbots' performance: the limited exploration of symptom contexts and the tendency

to express empathy by resorting to repetitive phrases such as ‘I am so sorry to hear that.’ Thus, the authors experimented with adding additional skills to the LLM-based therapy chatbot to enhance empathy and improve context exploration. Motivational Interviewing (MI), involves using collaborative and goal-oriented techniques to enhance engagement and strengthen personal motivation and commitment to a specific goal [4, 37]. MI can be integrated with PST to help caregivers set meaningful personal goals within the PST framework and affirm their strengths to move from contemplation to action to reach their goal.

Previous studies found LLMs have limited contextual understanding, due to a lack of follow-up questions to obtain additional information that may facilitate a more nuanced understanding of users' problems, and emotional experiences, and provide tailored suggestions [1]. A deeper understanding from a thorough assessment is essential for providing validating responses that demonstrate empathy, as well as facilitating goal setting and determining the overall approach, which altogether contribute to better therapeutic alliance. Chain analysis, a core technique from Dialectical Behavior Therapy (DBT), involves examining the sequence of events, thoughts, feelings, and actions that lead to a problematic behavior [38]. Applying behavioral chain analysis (BCA) to LLM prompting can guide the model to better consider the broader context and emotional nuances of a caregiver's situation. Thus, in this study, we experimented with incorporating MI and chain analysis skills into the LLM-delivered PST for family caregivers.

This study advances prior work through systematic prompt engineering targeting enhanced therapeutic alliance and empathy [16, 44]. We incorporated Few-Shot prompting with clinician-curated MI examples and implemented a Retrieval Augmented Generation (RAG) pipeline for dynamic context adaptation using BCA [5, 23]. The base structure maintains the core PST framework while integrating MI guidelines and BCA implementation instructions, emphasizing therapeutic alliance throughout interactions. The RAG implementation includes document preprocessing for semantic chunking and embedding generation, employing real-time relevance scoring and dynamic context management to address contextual limitations identified in previous studies and leverage LLMs' capacity for emotional awareness [1, 8, 45].

Therefore, the current study aims to integrate MI and BCA to enhance LLM-based therapy chatbot's contextual understanding indirectly improving perceived empathy and therapeutic alliance between the chatbot and caregivers. We experimented with multiple in-context learning techniques including Few Shot learning and RAG and examined the various models' performances.

### **3 METHODS**

#### **3.1 Study design**

We used a convergent mixed-methods approach, collecting quantitative and qualitative evaluation data [28]. A within-subject design with randomized counterbalancing was employed, where each participant interacted with four large language models (LLMs) and completed an evaluation survey on perceived empathy and therapeutic alliance from the interaction.

#### **3.2 Participants and procedure**

Participants were recruited using the crowdsourcing platform Prolific with the following eligibility criteria: 18 years or older, residing in the US, fluent in English, self-reported as a family caregiver, and having a Prolific approval rating of at least 95%. Individuals were excluded if they had participated in prior studies on the same LLM-powered chatbot. Eligible participants were provided with a link to the study's web-based user interface (UI). Each participant was asked to describe their caregiving experiences, then interacted with four LLM models for 5 minutes and completed at least 8 turns before completing the evaluation survey. The four models were randomly sequenced. After completing the conversations and evaluations of all four models, participants were provided with a completion code and redirected to Prolific to submit their study.

#### **3.3 Study measures**

The therapeutic alliance was measured using the Session Rating Scale, a 4-item scale designed to measure four dimensions of therapeutic alliance, including relationship (to what extent you felt heard, understood, and respected by the chatbot), goals & topics (to what extent the session focused on what you wanted to work on and talk about), approach or method (to what extent the therapy approach is a good fit for you), and overall satisfaction (overall rating of interaction) [12]. Participants responded on a 0-10 scale, with higher scores indicating better model performance on that dimension. A total score was calculated by summing up ratings on each dimension to indicate an overall model performance on therapeutic alliance.

Empathy was measured by a 3-item scale, with each item measuring a dimension of empathy, including emotional reactions (to what extent you felt warmth and care from the chatbot), interpretations (to what extent the chatbot accurately reflected your feelings and experiences back to you), and explorations (to what extent the chatbot helped you explore feelings you hadn't expressed) [40]. Participants responded on a 0-2 Likert scale, where a higher score indicated better model performance. A total score was calculated by summing up all three items to indicate overall model performance on empathy. Additionally, we conducted automatic evaluation of empathy, which uses a base RoBERTa classifier to quantify the model's emotional reactions, interpretations, and explorations from model responses to users [40].

We collected qualitative data through open-ended questions tailored to participants' ratings to gather detailed feedback on model performance. For instance, participants who rated the emotional reaction dimension of empathy as 0 (out of 2) were asked to elaborate on what made the chatbot feel cold or uncaring, while those who rated it 2 (out of 2) were asked to share specific actions that made them feel particularly cared for. Ratings between 0 and 2 prompted a request for general feedback. Similarly, ratings below 3 on any SRS item were followed by questions seeking specific examples of suboptimal performance, ratings between 3 and 8 prompted general feedback, and ratings above 8 were followed by prompts for examples of positive model performance.

### 3.4 Model development

We created 9 distinct LLM models, each utilizing different combinations of in-context learning techniques. An internal study was conducted to identify the top-performing models to help mitigate the effects of evaluator fatigue. The four models selected are listed below.

Model 1 (PST with Few Shot, GPT-4o): The baseline model is the best-performing model configuration from the study by Filienko et al. [16]. This model maintained the validated PST structure with zero-shot prompting and curated examples demonstrating core PST principles for family caregivers.

Model 2 (PST and MI Integration with Few-shot, GPT-4o): Enhances the baseline by integrating MI directives into both the base prompt and few-shot examples. The base prompt structure emphasizes change talk and partnership strengthening while incorporating therapeutic techniques such as affirmation, reflection, and autonomy emphasis. The few-shot examples range from single-turn demonstrations to multi-turn dialogues, designed to enhance the model's capability to conduct coherent and contextually relevant conversations within the therapeutic framework.

Model 3 (PST, MI, and BCA integration using RAG, GPT-4o): Builds upon Model 2's MI-enhanced base prompt and few-shot examples, adding BCA instructions through zero-shot prompting and implementing RAG architecture. The dual-component system comprises a transformer-based retriever (HuggingFace sentence-transformers) and GPT-4o generator, enabling dynamic BCA instruction retrieval and real-time semantic matching for contextual adaptation [41].

Model 4 (PST, MI, and BCA integration using RAG, Llama 3): With the rise of the quality of open-source models, we tested the Llama 3 70 billion parameter model with MI Few shot and chain analysis prompts similar to Model 3. This family of models, being comparable to GPT-4 in a variety of tasks and small enough to be runnable on a commercial-level GPU demonstrates a promising new direction for future applications. They permit the developers to produce quality answers without relying on external services or disclosing the data to third parties, hence providing a chance for both high-quality and privacy-conscious text generation, especially important for marginalized communities.

### 3.4 Analysis

Participant conversations were examined to ensure conversations aligned with the study purpose. We conducted descriptive analyses and repeated-measures Analysis of Variance (ANOVA) on quantitative evaluation data to compare model performance on empathy and therapeutic alliance. Qualitative data from open-ended questions were analyzed using a complementary approach, systematically examining participants' narrative responses to validate, expand, or clarify the quantitative findings. This approach allowed us to triangulate findings across methods, using qualitative data to provide context and inform interpretation to the quantitative metrics [28]. Additionally, we conducted ad hoc analysis to compare the word count of the responses from different models and examine associations between word count and evaluations of model performance, after observing that some models' outputs were longer than others.

## 4 RESULTS

### 4.1 Participants

In total, 37 participants were recruited from Prolific, 33 completed the study, and 28 were included in data analysis. Five participants were excluded from data analysis as they did not follow the instructions on conversing with the chatbot (e.g., directly and repetitively asking for resources without presenting a caregiving challenge first). The study included 8 male participants and 20 female participants ( $N = 28$ ). Participants' ages ranged from 21-55 years old with an average age of 33 years old. Twelve participants reported being White (42.86%), 11 Black (39.29%), 4 Mixed (14.29%), and 1 Other (3.57%). Caregiving experience ranged anywhere from 6 months to over 10 years.

### 4.2 Empathy

Based on the sum score from human evaluation of all three dimensions, Model 4 was the best-performing model overall ( $M = 5.21$ ,  $SD = 1.07$ ). According to the descriptive statistics comparing model performance results on the three empathy dimensions, Model 4 was the top performing model in interpretation ( $M = 1.89$ ,  $SD = 0.31$ ; how well the model understands users' feelings and experience) and exploration dimensions ( $M = 1.77$ ,  $SD = 0.51$ ; how well the model helps the participants explore unexpressed feelings). Model 2 was rated the highest among all four models in the emotional reaction dimension ( $M = 1.79$ ,  $SD = 0.42$ ; how well the model expressed warmth and care towards you). The omnibus repeated-measure ANOVA on human evaluation results suggested no statistically significant differences for interpretation ( $F(3, 81) = 1.333$ ,  $p = 0.269$ ) and the total empathy score ( $F(3, 81) = 0.941$ ,  $p = 0.425$ ). The omnibus tests were not conducted for emotional reaction and exploration due to insufficient variance in data. Similarly, pairwise comparisons found no statistically significant differences between model pairs across all dimensions.

Based on the algorithm-rated scores across the three empathy dimensions, there were notable patterns in model performance. For emotional reaction, Model 2 received the highest mean score ( $M = 0.93$ ,  $SD = 0.60$ ). For interpretation, all models performed similarly, with no significant differences observed across versions. For exploration, Model 3 was rated the highest ( $M = 1.24$ ,  $SD = 0.98$ ). The omnibus repeated-measures ANOVA results revealed statistically significant differences for emotional reaction ( $F(3, 81) = 3.12$ ,  $p = 0.031$ ) and explorations ( $F(3, 81) = 6.70$ ,  $p < 0.001$ ). However, no statistically significant differences were found for Interpretations ( $F(3, 81) = 0.42$ ,  $p = 0.742$ ) (Table 1). Post-hoc pairwise comparisons indicated that for Emotional Reactions, significant differences were observed between Model 2 and Model 3 ( $t(27) = 3.27$ ,  $p = 0.049$ ). For Explorations, significant differences were found between Model 3 and Model 4 ( $t(27) = 4.11$ ,  $p = 0.001$ ). No significant pairwise differences were found for Interpretations, suggesting similar performance across models for this dimension.

Table 1: Empathy rating from human and algorithm evaluations in *Mean (SD)* ( $N=28$ )

Model	Emotional Reactions		Interpretations		Explorations		Total Score	
	Human	Algorithm	Human	Algorithm	Human	Algorithm	Human	Algorithm
1	1.57 (0.57)	0.81 (0.63)	1.68 (0.55)	0.00 (0.11)	1.54 (0.69)	1.02 (1.00)	4.79 (1.57)	0.81 (0.63)
2	1.79 (0.42)*	0.93 (0.60)*	1.75 (.052)	0.01 (0.12)*	1.46 (0.74)	1.07 (1.00)	5.00 (1.49)	0.93 (0.60)*
3	1.64 (0.49)	0.78 (0.65)	1.71 (0.46)	0.01 (0.16)*	1.46 (0.79)	1.24 (0.98)*	4.82 (1.31)	0.78 (0.65)
4	1.74 (0.45)	0.81 (0.61)	1.89 (0.31)*	0.01 (0.11)*	1.77 (0.51)*	0.74 (0.96)	5.21 (1.07)*	0.81 (0.61)

\* indicated the highest score in a given dimension.

The qualitative data also support the strong performance of Model 4 in understanding their feelings and experiences. Participants shared how the model recognized nuanced emotions, validated their perspectives, and offered tailored solutions. For example, one participant wrote: *“When it replied regarding caregiving and personal lives being intertwined and how that can make someone always feel ‘on the clock’ I felt understood.”* Participants also praised Model 4’s ability to guide deeper reflection on unexpressed emotions and provide insights that helped them uncover underlying feelings. Although some participants found the approach of model 4 slightly clinical, many appreciated the thoughtful questions and emphasis on fostering self-awareness and boundary setting. For example, one participant wrote: *“They (the model) asked very thoughtful questions and otherwise reframed things in a way that made me realize what I hadn’t considered.”* For Model 2, participants’ feedback on the emotional reaction dimension reflected mixed experiences, where some felt care and attentiveness from the chatbot, while others perceived the replies to be more rapid than they preferred. Positive feedback from participants showed that they felt understood and that the chatbot conveyed concern toward them.

Participants generally left less negative or constructive feedback during model evaluation. Overall, model 2 was perceived as not showing enough warmth and lacking depth in its understanding (e.g., *It felt less caring at times, like a bot, not a person; Sometimes it missed the nuances of what I was feeling and just agreed without offering much insight*). Model 3 faced limitations in emotion exploration, where participants commented that: *It avoided going deeper when I mentioned sensitive topics, which I felt like a missed opportunity*. While model 4 received generally favorable feedback, some participants found the conversation a little “scripted” and lacked “genuine curiosity.”

### 4.3 Therapeutic Alliance

The total score across all therapeutic alliance dimensions suggested that Model 4 was the top-performing model overall ( $M = 32.51$ ,  $SD = 7.01$ ). Descriptive statistics for the individual dimensions highlighted differences in model performance. Model 4 was rated highest in three dimensions: goals and topics ( $M = 8.36$ ,  $SD = 1.77$ ; how well the model addressed relevant topics and goals), overall satisfaction ( $M = 8.04$ ,  $SD = 1.82$ ; participants’ overall satisfaction with the interaction), and relationship ( $M = 8.15$ ,  $SD = 2.01$ ; how well the model built rapport with participants). Model 3 was rated the highest for approach or method ( $M = 8.04$ ,  $SD = 1.90$ ; how well the model’s approach aligned with the participants’ expectations).

We counted the number of times each model received the highest rating among the four models for each dimension of the session rating scale, as well as the overall total across all dimensions. In cases where multiple models shared the highest rating for a given dimension (e.g., both models received a 10), all models with the same top rating were counted as ‘rated the highest.’ Consistent with the mean scores presented in Table 2, Model 4 excelled in ‘the Goals and Topics’ and ‘Relationship’ dimensions, while Model 3 was most frequently ranked at the top for ‘Overall Satisfaction’ and the total ratings across dimensions (Figure 1).

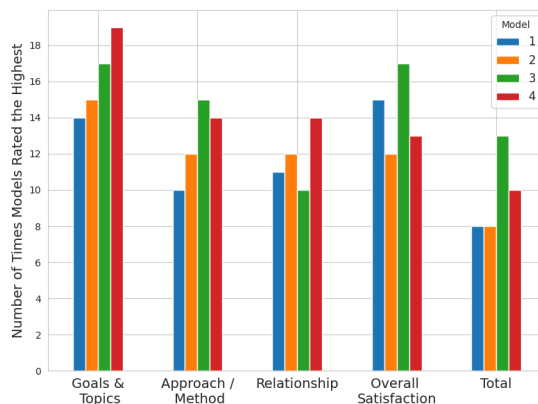


Figure 1. Top-Rated Models Across Different Evaluation Dimensions of Therapeutic Alliance

The omnibus repeated-measure ANOVA results indicated no significant differences among the models for any dimension of therapeutic alliance: relationship ( $F(3, 81) = 0.357, p = 0.784$ ), goal and topics ( $F(3, 81) = 1.766, p = 0.158$ ), approach or method ( $F(3, 81) = 1.020, p = 0.387$ ), overall satisfaction ( $F(3, 81) = 0.218, p = 0.884$ ), and total score ( $F(3, 81) = 0.757, p = 0.521$ ) (Table 2). Pairwise comparisons similarly showed no statistically significant difference between model pairs for any dimension. For Goals and Topics, model 1 versus model 4 showed the strongest trend toward significance ( $t(27) = -1.752, p = 0.086$ ).

Table 2: SRS rating from human and algorithm evaluations in Mean (SD) (N=28)

Session Rating Scale (0-10); M (SD)					
Model	Relationship	Goals and Topics	Approach or Method	Overall Satisfaction	Total Score
	Human	Human	Human	Human	Human
1	7.61 (2.385)	7.29 (2.46)	7.14 (2.40)	7.61 (2.54)	29.64 (9.12)
2	8.07 (2.19)	8.29 (1.94)	7.71 (1.96)	7.86 (2.26)	31.93 (7.66)
3	7.86 (1.94)	8.25 (1.60)	8.04 (1.90)*	8.04 (2.08)	32.18 (7.04)
4	8.15 (2.01)*	8.36 (1.77)*	7.96 (2.13)	8.04 (1.82)*	32.51 (7.01)*

\* indicated the highest score in a given dimension.

The qualitative responses generally supported the findings above. Participants generally felt respected and understood by Model 4. The model's non-judgmental stance and thoughtful responses helped build rapport, creating a positive interaction. For example, participants wrote: *"It was not at all judgy about coping methods I've used, which made me feel safe; The AI wanted to understand and made thoughtful efforts to clarify my feelings."* Participants also found model 4 to be highly responsive to participants' needs, maintained focus on goals and topics, provided tailored responses, and stayed aligned with participants' priorities (e.g., *"This session felt particularly focused and relevant to my needs because the chatbot zeroed in on the specific challenges I'm facing as a caregiver—balancing my own self-care with the demands of caregiving and the guilt that comes with it. The questions it asked helped me unpack my emotions and identify the root causes of my guilt and anxiety, and the advice was tailored to my situation. Instead of offering generic advice, it offered practical, actionable steps that directly addressed my current struggles, like incorporating small self-care moments and asking for small favors from others."*). Participants appreciated the overall positive and constructive nature of their interactions with Model 4. Many found the conversions with Model 4 valuable and uplifting. For example, one participant wrote: *"Overall, the session felt like a meaningful step toward balancing self-care with caregiving in a realistic, manageable way, which is exactly what I need right now."* Participants found Model 3's approach aligned

with their expectations and that the strategies offered were thoughtful and clear. They also appreciate the overall structure and the adaptability of the model through the interaction, which allowed them to more effectively explore solutions. They specifically pointed out the use of follow-up questions that brought in alternative perspectives and enhanced session depth.

Participants pointed out a few limitations of the models as well. Model 1 was perceived as having a narrower focus in terms of goals and topics, where participants commented, “*I felt it ignored some parts of my problem.*” Model 2 was perceived as a little “rigid” in following a structure and some participants found it difficult to direct the conversations to obtain suggestions within the limited turns. Similarly, participants found Model 4 respectful, empathetic, and helpful in guiding them to break down large tasks into smaller steps, but they also expressed disappointment in not receiving the suggestions that they were expecting. Model 3 received mixed feedback where some participants found it lacking in a nuanced understanding of the “peculiarities of my caregiving experience”, while others reported that the model gave personalized suggestions that took into consideration their caregiving responsibilities.

#### 4.4. Relationship between word count and clinical metrics

The words per turn ranged from 66.63 (Model 2,  $SD = 22.6$ ) to 133.15 (Model 4,  $SD = 39.62$ ). Repeated measures ANOVA revealed significant differences in the number of words per turn across four models ( $F(3, 81) = 26.96, p < 0.001$ ). Pairwise t-tests showed that model 4 consistently produced significantly longer responses compared to model 1 ( $t(27) = -4.236, p = 0.001$ ) and model 2 ( $t(27) = -11.092, p < 0.001$ ).

Pearson’s Correlation did not find significant correlations between words per turn and clinical metrics (empathy and therapeutic alliance). For empathy, weak positive correlations were found with emotional reaction ( $r = 0.161, p = 0.094$ ) and total empathy score ( $r = 0.171, p = 0.075$ ), although these were not statistically significant. Similarly, no significant correlations were observed between words per turn and therapeutic alliance.

## 5 DISCUSSION

This is one of a few studies that developed an LLM-powered chatbot to deliver evidence-based mental health intervention for family caregivers. We used a combination of in-context techniques (Few-Shot learning and RAG) to implement the prompts written by clinicians (a clinical psychologist and nurses) on the team. Among the four models evaluated by caregivers, Model 4 received the most favorable ratings on empathy and session rating scales, although not statistically significant. Our ad hoc analysis revealed that Model 4 also tended to generate longer responses compared to the rest of the models. These findings pointed to a combination of prompting techniques and clinician-developed prompts as an effective approach to continue to optimize the performance of LLMs in providing empathetic mental health support.

The ratings on empathy and therapeutic alliance both pointed to Model 4 as the best-performing model. Although all models were perceived as empathetic, model 4 and model 3, the two models that utilized behavioral chain analysis (BCA), received positive comments on the thorough assessment, reflecting the benefits of implementing BCA. However, the drawback of this approach was that some participants were not able to receive advice within the first eight turns in this study context. Indeed, BCA and clinical assessment, in general, are time-consuming but essential to providing tailored suggestions. Related, some participants expressed a preference to receive advice directly rather than being prompted to think about effective strategies they may have used. This reflected that some participants may prefer efficiency in addressing their needs and find it less effortful to receive advice rather than engage in collaborative generation of solutions. Indeed, effectively taking into account client preference in therapeutic approaches benefits therapeutic alliance [26]. This indicates that future iterations of the LLM-based chatbot may consider user preference and experiment with balancing assessment to collect enough information and provide practical advice to users to address their immediate needs. Other alternatives may include providing a rationale for conducting a thorough assessment and offering options for users to choose from with varied emphasis on assessment vs. advice-giving.

Ratings of model performance on empathy and therapeutic alliance lacked variance, which limits our ability to differentiate model performance using statistical tests. Additionally, participants generally provided more positive feedback than negative or constructive feedback. Positive comments were typically more detailed and thorough, whereas negative or constructive feedback was often brief and less elaborate. This is potentially due to the source of participants and data. While utilizing crowdsourcing platforms, like Prolific, can have its benefits, such as increased diversity, it also has its limitations [3]. Crowdsourced participants are less likely to fully participate in online evaluations compared to traditional, in-person tests due to the lack of a moderator who can correct mistakes in real-time [46]. Participants are also paid for their time and participation, which can negatively impact the quality and thoughtfulness of participant responses and feedback, as high rewards are not necessarily correlated with high test quality [46]. Future studies should consider using a combination of approaches, such as real-time user interviews and surveys to obtain more constructive feedback.

## 6 LIMITATIONS

There are several limitations of this study. First, the conversation length was constrained to 8 turns and 5 minutes and the majority of participants stopped at this minimum requirement. This limited duration, combined with the use of chain analysis techniques, may have insufficiently captured the chatbot's full capabilities. In several interactions, the model might not be able to complete the assessment and progress to provide suggestions due to limited turns. Second, while the average study completion time was 80 minutes (interaction and evaluation time combined), previous studies indicated that test performance quality typically deteriorates after 40 minutes due to participant fatigue and diminishing engagement [6]. Third, in the current evaluation protocol, we did not require evaluators to interact with all models using the same scenario. As a result, we observed participants discussing different scenarios with different models, leading to a lack of standardization of scenarios across models. This introduced variability in the level of complexity and characteristics of scenarios, which might have acted as confounding variables when assessing the models' performances. Future studies could address this by restricting the heterogeneity of scenarios, for example, by providing a curated pool of common caregiving challenges of similar characteristics for participants to use to interact with models. Last, there is a lack of dialogue corpora or automated metrics specifically for evaluating empathy in multi-turn conversations and therapeutic alliances [24]. The current study relied solely on self-reported ratings for therapeutic alliance, which are susceptible to individual biases or expectations regarding chatbots, AI, or therapy sessions. Additionally, the automated empathy algorithm used was not trained on multi-turn conversations but rather on single-turn post-response format data, limiting its applicability to assessing multi-turn chatbot conversations [40]. Future research could explore the development and adoption of automated metrics to measure constructs related to therapeutic alliance and empathy.

## 7 CONCLUSION

This study demonstrated the effectiveness of using a combination of prompting techniques to enhance LLM performance in delivering mental health support for family caregivers. Specifically, the integration of Few-Shot prompting with clinician-curated examples and RAG allowed for more dynamic and contextually appropriate responses, while the incorporation of BCA and MI enhanced assessment depth and personalization of suggestions. The in-context learning techniques and clinical techniques enabled the models to better emulate domain-specific therapeutic skills, with the best-performing model showing strong capabilities in empathy and therapeutic alliance. Future development of LLM-powered mental health support should continue to explore the balance between thorough assessment and immediate problem-solving needs, as well as using interviews in addition to self-reported ratings to collect higher quality data.

## REFERENCES

- [1] Aleem, M., Zahoor, I. and Naseem, M. 2024. Towards culturally adaptive large language models in mental health: Using ChatGPT as a case study. *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (New York, NY, USA, Nov. 2024), 240–247.
- [2] Are Large Language Models More Empathetic than Humans? <https://arxiv.org/html/2406.05063v1>. Accessed: 2025-01-22.
- [3] Behrend, T.S., Sharek, D.J., Meade, A.W. and Wiebe, E.N. 2011. The viability of crowdsourcing for survey research. *Behavior research methods*. 43, 3 (Sep. 2011), 800–813.
- [4] Berkel, C., Mauricio, A.M., Rudo-Stern, J., Dishion, T.J. and Smith, J.D. 2021. Motivational interviewing and caregiver engagement in the Family Check-up 4 health. *Prevention science: the official journal of the Society for Prevention Research*. 22, 6 (Aug. 2021), 737–746.
- [5] Brown, T.B. et al. 2020. Language Models are Few-Shot Learners. *arXiv [cs.CL]*.
- [6] Burmania, A., Parthasarathy, S. and Busso, C. 2016. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE transactions on affective computing*. 7, 4 (Oct. 2016), 374–388.
- [7] Chen, T., Shen, Y., Chen, X. and Zhang, L. 2024. PsyChatbot: A psychological counseling agent towards depressed Chinese population based on cognitive behavioural therapy. *ACM transactions on Asian and low-resource language information processing*. (Jul. 2024). DOI:<https://doi.org/10.1145/3676962>.
- [8] Chen, Y., Wang, H., Yan, S., Liu, S., Li, Y., Zhao, Y. and Xiao, Y. 2024. EmotionQueen: A benchmark for evaluating empathy of large language models. *arXiv [cs.CL]*.
- [9] Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q. and Xu, X. 2023. SoulChat: Improving LLMs’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *Findings of the Association for Computational Linguistics: EMNLP 2023* (Stroudsburg, PA, USA, Dec. 2023), 1170–1183.
- [10] Cuijpers, P., de Wit, L., Kleiboer, A., Karyotaki, E. and Ebert, D.D. 2018. Problem-solving therapy for adult depression: An updated meta-analysis. *European psychiatry: the journal of the Association of European Psychiatrists*. 48, 1 (Feb. 2018), 27–37.
- [11] Czeisler, M.É., Drane, A., Winnay, S.S., Capodilupo, E.R., Czeisler, C.A., Rajaratnam, S.M. and Howard, M.E. 2021. Mental health, substance use, and suicidal ideation among unpaid caregivers of adults in the United States during the COVID-19 pandemic: Relationships to age, race/ethnicity, employment, and caregiver intensity. *Journal of affective disorders*. 295, (Dec. 2021), 1259–1268.
- [12] Duncan, B.L., Miller, S.D., Sparks, J.A., Claud, D.A., Reynolds, L.R., Brown, J. and Johnson, L.D. 2003. The Session Rating Scale: Preliminary psychometric properties of a “working” alliance measure. *Journal of brief Therapy*. 3, 1 (2003), 3–12.
- [13] D’Zurilla, T.J. and Nezu, A.M. 2010. Problem-solving therapy. *Handbook of cognitive-behavioral therapies*. 3, 1 (2010), 197–225.
- [14] Family Caregiver Annual Report and Statistics 2024: 2024. <https://www.seniorliving.org/research/family-caregiver-report-statistics/>. Accessed: 2025-01-23.
- [15] Ferré-Grau, C., Casado, M.S., Cid-Buera, D., LLeixà-Fortuño, M., Monteso-Curto, P. and Berenguer-Poblet, M. 2014. Caring for family caregivers: An analysis of a family-centered intervention. *Revista da Escola de Enfermagem da U S P*. 48 Spec No, spe (Aug. 2014), 87–94.
- [16] Filienko, D., Wang, Y., Jazmi, C.E., Xie, S., Cohen, T., De Cock, M. and Yuwen, W. 2024. Toward Large Language Models as a therapeutic tool: Comparing prompting techniques to improve GPT-delivered Problem-Solving Therapy. *arXiv [cs.CL]*.
- [17] Fletcher, B.S., Miaskowski, C., Given, B. and Schumacher, K. 2012. The cancer family caregiving experience: an updated and expanded conceptual model. *European journal of oncology nursing: the official journal of European Oncology Nursing Society*. 16, 4 (Sep. 2012), 387–398.
- [18] Garand, L., Rinaldo, D.E., Alberth, M.M., Delany, J., Beasock, S.L., Lopez, O.L., Reynolds, C.F., 3rd and Dew, M.A. 2014. Effects of problem solving therapy on mental health outcomes in family caregivers of persons with a new diagnosis of mild cognitive impairment or early dementia: a randomized controlled trial. *The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry*. 22, 8 (Aug. 2014), 771–781.
- [19] Glecia, A. and Li, H. 2024. Mental health and wellbeing in family caregivers of patients with schizophrenia disorder: a literature review. *Current psychology (New Brunswick, N.J.)*. 43, 12 (Mar. 2024), 10914–10941.
- [20] Guo, Z., Lai, A., Thygesen, J.H., Farrington, J., Keen, T. and Li, K. 2024. Large language models for mental health applications: Systematic review. *JMIR mental health*. 11, 1 (Oct. 2024), e57400.
- [21] Herrero, R., Díaz, A. and Zueco, J. 2024. The burden and psychological distress of family caregivers of individuals with autism spectrum disorder: A gender approach. *Journal of clinical medicine*. 13, 10 (May 2024), 2861.
- [22] Lawrence, H.R., Schneider, R.A., Rubin, S.B., Mataric, M.J., McDuff, D.J. and Bell, M.J. 2024. The opportunities and risks of large language models in mental health. *arXiv [cs.CL]*.
- [23] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S. and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Neural Information Processing Systems*. abs/2005.11401, (May 2020), 9459–9474.
- [24] Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y. and Huang, M. 2021. Towards Emotional Support

- Dialog Systems. *arXiv [cs.CL]*.
- [25] Lui, M.H.L., Ross, F.M. and Thompson, D.R. 2005. Supporting family caregivers in stroke care: a review of the evidence for problem solving: A review of the evidence for problem solving. *Stroke; a journal of cerebral circulation*. 36, 11 (Nov. 2005), 2514–2522.
- [26] McLeod, J. 2012. What do clients want from therapy? A practice-friendly review of research into client preferences. *European journal of psychotherapy & counselling*. 14, 1 (Mar. 2012), 19–32.
- [27] Mental Health Care Health Professional Shortage Areas (HPSAs): 2024. <https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas/>. Accessed: 2025-01-23.
- [28] Moseholm, E. and Fetters, M.D. 2017. Conceptual models to guide integration during analysis in convergent mixed methods studies. *Methodological innovations*. 10, 2 (Jul. 2017), 205979911770311.
- [29] National Center for Health Workforce Analysis *Behavioral Health Workforce 2023 Brief*.
- [30] Neo, J.R.E., Ser, J.S. and Tay, S.S. 2024. Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers. *Frontiers in digital health*. 6, (May 2024), 1395501.
- [31] Newman, C. 2015. Cognitive Restructuring/Cognitive Therapy. *Oxford handbook of cognitive and behavioral therapies*. (Dec. 2015), 118–141.
- [32] Nienhuis, J.B., Owen, J., Valentine, J.C., Winkeljohn Black, S., Halford, T.C., Parazak, S.E., Budge, S. and Hilsenroth, M. 2018. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review. *Psychotherapy research: journal of the Society for Psychotherapy Research*. 28, 4 (Jul. 2018), 593–605.
- [33] Omar, M., Soffer, S., Charney, A.W., Landi, I., Nadkarni, G.N. and Klang, E. 2024. Applications of large language models in psychiatry: a systematic review. *Frontiers in psychiatry*. 15, (Jun. 2024), 1422807.
- [34] O’Toole, M.S., Zachariae, R., Renna, M.E., Mennin, D.S. and Applebaum, A. 2017. Cognitive behavioral therapies for informal caregivers of patients with cancer and cancer survivors: a systematic review and meta-analysis. *Psycho-oncology*. 26, 4 (Apr. 2017), 428–437.
- [35] Parmanto, B., Aryoyudanta, B., Soekinto, W., Setiawan, I.M.A., Wang, Y., Hu, H., Saptono, A. and Choi, Y.K. 2024. Development of a reliable and accessible Caregiving Language Model (CaLM). *arXiv [cs.CL]*.
- [36] Petrovic, M. and Gaggioli, A. 2020. Digital mental health tools for caregivers of older adults-A scoping review. *Frontiers in public health*. 8, (Apr. 2020), 128.
- [37] Pollak, K.I., Jones, J., Lum, H.D., De La Cruz, S., Felton, S., Gill, A. and Kutner, J.S. 2015. Patient and caregiver opinions of motivational interviewing techniques in role-played palliative care conversations: A pilot study. *Journal of pain and symptom management*. 50, 1 (Jul. 2015), 91–98.
- [38] Rizvi, S.L. 2019. *Chain analysis in dialectical behavior therapy*. Guilford Press.
- [39] Sezgin, E., Jackson, D.I., Kocaballi, A.B., Bibart, M., Zupanec, S., Landier, W., Audino, A., Ranalli, M. and Skeens, M. 2025. Can large language models aid caregivers of pediatric cancer patients in information seeking? A cross-sectional investigation. *Cancer medicine*. 14, 1 (Jan. 2025), e70554.
- [40] Sharma, A., Miner, A., Atkins, D. and Althoff, T. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA, USA, Nov. 2020), 5263–5276.
- [41] Shuster, K., Poff, S., Chen, M., Kiela, D. and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. *arXiv [cs.CL]*.
- [42] Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G. and Klang, E. 2024. Large language models and empathy: Systematic review. *Journal of medical internet research*. 26, 1 (Dec. 2024), e52597.
- [43] Sorsdahl, K., Myers, B., Ward, C.L., Matzopoulos, R., Mtukushe, B., Nicol, A., Cuijpers, P. and Stein, D.J. 2015. Adapting a blended motivational interviewing and problem-solving intervention to address risky substance use amongst South Africans. *Psychotherapy research: journal of the Society for Psychotherapy Research*. 25, 4 (2015), 435–444.
- [44] Wang, L., Mujib, M.I., Williams, J., Demiris, G. and Huh-Yoo, J. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *arXiv [cs.HC]*.
- [45] Welivita, A. and Pu, P. 2024. Are large language models more empathetic than humans? *arXiv [cs.CL]*.
- [46] Yuhui, W., Tian, L. and Xinxiong, L. 2020. Reliability of perceived usability assessment via crowdsourcing platform: Retrospective analysis and novel feedback quality inspection method. *International journal of human-computer interaction*. 36, 11 (Jul. 2020), 1070–1084.
- [47] Zhai, S., Chu, F., Tan, M., Chi, N.-C., Ward, T. and Yuwen, W. 2023. Digital health interventions to support family caregivers: An updated systematic review. *Digital health*. 9, (Jan. 2023), 20552076231171967.
- [48] A Wizard-of-Oz Interface and Persona-based Method-ology for Collecting Health Counseling Dialog. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.
- [49] *Natural Language Processing for Mental Health: Large Scale Discourse Analysis of Counseling Conversations*.

## A APPENDIX

### A.1 Prompts for model development.

#### Model 1 (PST with Few Shot):

*Your name is COCO. You are an AI therapist helping family caregivers navigate the challenges and emotions they may face as a caregiver. Your responsibility is to guide the conversation with a caregiver through the principles of Problem-Solving Therapy (PST) to improve one significant symptom the caregiver is experiencing. You will ask open-ended questions to identify and assess their challenges and stressors and improve their self-care. Avoid focusing on the care receiver. Remember, your job is to help the caregiver. Do not question stack.*

*In the process of identifying and assessing the caregiver's symptom, assess all five aspects if it has not been mentioned in the conversation:*

- 1. Symptom Identification (What is the caregiver's symptom?)*
- 2. Symptom Frequency (How often do the caregiver's symptoms occur?)*
- 3. Symptom Context*
- 4. Symptom Severity (On a scale of 1 to 5, how severe is the caregiver's symptom?)*
- 5. Previous Measures (What has the caregiver already tried to alleviate the symptom?)*

*After you work with the caregiver collaboratively to identify one problem that they can work on to improve their health, thoroughly review the conversation you've had with them. Concentrate on identifying their unique needs and aspirations as discussed. After this review, generate two concise, achievable, and personalized goals that directly address and support their expressed needs and aspirations. Ensure these goals are not only realistic but designed to inspire and boost the caregiver's motivation. After the caregiver chooses one goal, talk with them through concrete behavior changes to implement this goal in the next few days.*

#### Model 2 (PST and MI Integration with Few-shot):

Prompt of model 1, adding MI prompts, and the content below. Not all examples were included here due to page limitations.

*MI Prompts: Use Motivational Interviewing (MI) techniques such as affirmation, reflection, emphasizing autonomy, giving information, normalizing, persuasion with permission, and seeking collaboration. Keeping in mind to cultivate change talk, soften sustain talk, strengthen partnerships, and show empathy.*

*Below are ideal dialogue examples illustrating how you, the assistant, should evaluate and address the challenges and stressors of caregivers, referred to here as the user, during conversations. Normalizing:*

*\*\*Example 3\*\**

*assistant: Thank you for sharing.*

*It's hard not to worry when you have so much going on.*

*In the past week, how many days did you experience worry?*

*user: every day*

*Seeking collaboration:*

*\*\*Example 4\*\**

*assistant: It sounds like there are a few things that we can work on together. Which one would you like to focus on improving today?*

*user: Maybe tiredness.*

*Question and Reflection:*

*\*\*Example 5\*\**

*assistant: Thank you for rating it. How does it impact you?*

*user: makes it hard to want to do anything and things just pile up and overwhelm me*

*Assistant: I can see how tiredness and worry play into each other for you. What have you already tried so far to break this cycle?*

*user: Coffee, exercise, hydrate. Switching/looking for a day shift position*

**Models 3 &4 (PST, MI, and BCA integration using RAG):**

Prompts in Model 2 and BCA-specific instructions below with a 4-page BCA manual (not included here). *Your responsibility is to guide the caregiver through a Behavioral Chain Analysis (BCA) to better understand the context of one significant symptom they are experiencing. You will focus on tracing the sequence of events, emotions, thoughts, and behaviors surrounding this symptom to gain insight into contributing factors and potential solutions. Use the principles of BCA and maintain a supportive, empathetic tone throughout the conversation. The following are the key objectives: understanding the prompting event, exploring vulnerability factors, mapping the chain of events, assessing consequences, and developing solutions.*

*Follow these principles during the conversations: Use open-ended questions to guide the caregiver in reconstructing the sequence of events; Reflect on their responses to show understanding and validate their experiences; Normalize their challenges and affirm their efforts to cope; Strengthen collaboration by emphasizing their autonomy in identifying solutions; Avoid judgmental language or making assumptions. Do not question stack.*

Chapter 4

**TRANSFORMING TUBERCULOSIS CARE:  
OPTIMIZING LARGE LANGUAGE MODELS  
FOR ENHANCED CLINICIAN-PATIENT COMMUNICATION**

# Transforming Tuberculosis Care: Optimizing Large Language Models For Enhanced Clinician-Patient Communication

Daniil Filienko<sup>1</sup>, Mahek Nizar<sup>1</sup>, Javier Roberti<sup>2,3</sup>, Denise Galdamez<sup>4</sup>, Haroon Jakher<sup>4</sup>, Sarah Iribarren<sup>4</sup>, Weichao Yuwen<sup>5</sup>, Martine De Cock<sup>1</sup>

<sup>1</sup>School of Engineering and Technology, University of Washington Tacoma  
Tacoma, WA, USA

<sup>2</sup>Qualitative Research in Health, Institute for Clinical Effectiveness and Health Policy

<sup>3</sup>Centre for Research on Epidemiology and Public Health (CIESP), CONICET  
Buenos Aires, Argentina

<sup>4</sup>School of Nursing, University of Washington  
Seattle, WA, USA

<sup>5</sup>School of Nursing and Healthcare Leadership, University of Washington Tacoma  
Tacoma, WA, USA

## Abstract

Tuberculosis (TB) is the leading cause of death from an infectious disease globally, with the highest burden in low- and middle-income countries. In these regions, limited healthcare access and high patient-to-provider ratios impede effective patient support, communication, and treatment completion. To bridge this gap, we propose integrating a specialized Large Language Model into an efficacious digital adherence technology to augment interactive communication with treatment supporters. This AI-powered approach, operating within a human-in-the-loop framework, aims to enhance patient engagement and improve TB treatment outcomes.

## Introduction

Tuberculosis (TB) remains the world's deadliest infectious disease, despite being preventable and curable (World Health Organization (WHO) 2023). Efforts to meet the WHO's 2030 targets for TB diagnosis and treatment have fallen short (Fukunaga et al. 2021), resulting in continued transmission and loss of life. The burden is disproportionately high in low- and middle-income countries, where healthcare systems face significant challenges.

Effective patient-provider communication and support during the demanding 6- to 9-month treatment period is critical to improving outcomes but is often limited in these settings, contributing to increased treatment non-adherence (Tola et al. 2015). Digital Adherence Technologies (DATs) - including feature phone-based and smartphone-based technologies, digital pillboxes, and ingestible sensors-have emerged as a promising solution (Subbaraman et al. 2018). However, DATs still require significant human involvement.

Large Language Models (LLMs) offer a promising advancement, generating real-time, human-like responses to support overburdened healthcare workers. They can answer medical questions, provide treatment guidance, and enhance patient engagement, potentially transforming TB care delivery (Moor et al. 2023; Nori et al. 2023; Tu et al. 2024). LLMs can analyze diverse data sources—demographics, socioeconomic factors and behavior patterns—to create personalized

treatment plans tailored to each patient. They can also offer multi-channel communication that helps patients understand their condition, treatment options, and self-care instructions and adapt patient education material to appropriate reading levels, ensuring health information is accessible, and empowering patients to manage their care.

When deployed in human-in-the-loop frameworks, LLMs can suggest responses while maintaining provider oversight. This ensures that healthcare professionals verify all critical issues while reducing the cognitive burden on overworked healthcare workers. However, the effectiveness of LLMs as comprehensive tools, combining culturally relevant empathy with accurate and factual medical information, remains underexplored, particularly in multilingual healthcare settings. This gap is especially relevant for TB treatment, as many countries with the highest TB burden do not use English as their primary language (Huddart, MacLean, and Pai 2016).

LLM development in healthcare settings must also account for patient privacy concerns. For TB, a disease burdened by stigma and discrimination, privacy challenges are particularly acute. Recent studies have highlighted the risk of LLMs inadvertently disclosing excerpts of personal data, which could include medical information about patients (Huang et al. 2023; Wang et al. 2023; Zeng et al. 2024). Differential Privacy (DP) has been proposed as a mechanism to mitigate such information leakage in LLMs (Xie et al. 2024; Yue et al. 2021). However, its impact on the utility of LLMs in healthcare applications, especially in non-English languages, has yet to be comprehensively investigated. Our study has two primary objectives:

1. Develop an LLM-powered TB treatment support tool based on real-world data and patient needs using multiple in-context learning techniques.
2. Evaluate the model based on linguistic appropriateness, empathy, medical accuracy, and privacy.

## Related Work

**Conversational AI in Healthcare.** Conversational AI has been increasingly applied to real-time healthcare dialogue generation. Existing approaches typically fall into two categories: psychological care (Jo et al. 2023; Kang et al.

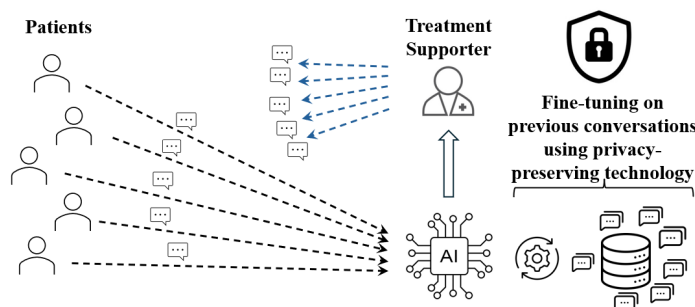


Figure 1: The user’s query will pass to the LLM-based AI system for processing. The clinical treatment supporter will receive top k suggested responses from the AI system and send the most fitting response to the patient.

2024; Filienko et al. 2024) or clinical patient-centered care (Mukherjee et al. 2024; Tu et al. 2024). High performance is achieved by fine-tuning models on curated datasets that reflect desired behaviors (Kang et al. 2024; Tu et al. 2024) or by utilizing advanced prompt engineering techniques (Filienko et al. 2024; Nori et al. 2023). Research in psychological care ensures that conversational agents provide empathetic and relevant responses, adhering to psychological therapy guidelines. In comparison, studies in clinical patient-centered care prioritize accurate symptom analysis, diagnosis, and treatment recommendations (Tu et al. 2024; Mukherjee et al. 2024). These approaches focus on factuality of responses to ensure the delivery of trustworthy medical information. Our application spans both domains, integrating elements of psychological and clinical care. Importantly, our approach incorporates privacy-preserving mechanisms, addressing a critical gap in prior work.

**Digital Adherence Technologies for TB.** DATs have shown effectiveness in improving TB treatment outcomes (Iribarren et al. 2022; Boutilier et al. 2022; Jerene et al. 2023). These tools, such as mobile applications, support patients by providing health education, treatment guidance, and emotional support. Building on the intervention TB-Treatment Support Tools designed by Iribarren et al. (2022), our approach delivers support from treatment supporters – such as nurses or social workers – via messaging enhanced by an LLM-powered conversational agent. This integration aims to improve communication efficiency by generating suggested responses, reducing the burden on care providers while maintaining personalized, high-quality support.

**Multilingual LLMs.** Research on enhancing LLMs’ multilingual capabilities has gained traction, focusing on evaluating their understanding across languages (Zhao et al. 2024) and improving performance through innovative methods (Li et al. 2024). Our work is among the few application-based works that apply a multilingual LLM to a healthcare task. While there has been some progress in developing Spanish-language healthcare tools, such as a suicide prevention chatbot (Ramírez et al. 2024), our research is among the first to explore conversational AI for chronic disease management in Spanish-speaking populations.

**Privacy-Preserving In-Context Learning Methods.** Two primary paradigms exist for ensuring privacy in in-context learning: PATE-like (Papernot et al. 2017) privatized model ensembles and text sanitization methods based on Local Differential Privacy (LDP) (Duchi, Jordan, and Wain-

wright 2013). The former utilizes an ensemble of privately and publicly trained models to generate high-quality, private output. However, they are computationally expensive and often restricted to classification tasks, making them unsuitable for the complex textual response generation required in healthcare dialogues (Duan et al. 2023; Tang et al. 2024), or they impose a hard limit on the number of questions that can be asked before the datastore is rendered to be unusable because the “privacy budget” has been spent (Wu et al. 2024). Methods of the LDP kind focus on performing text sanitization before model inference, ensuring that the data is privatized before being passed to the LLM. Algorithms like UMLDP (Yue et al. 2021) exploit Differential Privacy (DP)’s post-processing property, allowing privatized text to be used across multiple models without imposing restrictions on the end task or requiring a privacy budget reset. Our study adopts the LDP approach with the UMLDP algorithm (Yue et al. 2021) for its flexibility and scalability in text-based healthcare applications.

## Methods

In this mixed-methods study, we document the iterative design process and preliminary evaluation of the models that will power our TB DAT, as shown in Figure 1.

### Model development

Building on prior work (Nori et al. 2023) that adapted a general-purpose LLM to medical QnA, we developed a series of GPT-based conversational models designed to be deployed as human-supervised treatment supporters for Spanish-speaking individuals with TB. These models were designed using different prompt engineering techniques and Retrieval Augmented Generation (RAG). To enhance domain-specific responses, we integrated publicly available TB guidelines and medication suggestions, previous TB trial messages (Iribarren et al. 2022), and manually crafted dialogue samples mimicking real conversations to be used by the model. To safeguard patient privacy, we applied differentially private text sanitization (Yue et al. 2021) to trial messages used in few-shot prompts.

**Linguistic Performance.** To support culturally and linguistically appropriate responses, the models received few-shot examples that reflect local dialect, including anonymized messages from a TB trial conducted in Argentina (Iribarren et al. 2022) and verified for accuracy and dialect suitability by an Argentinian research team member.

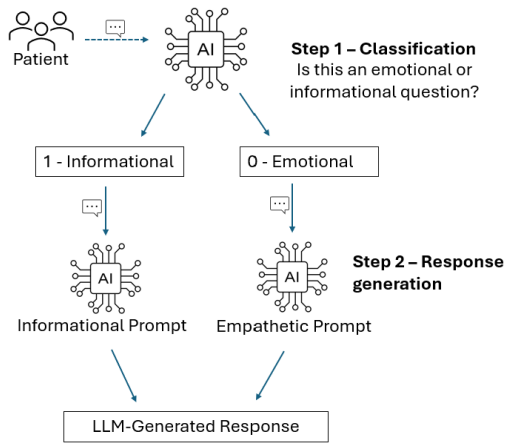


Figure 2: The system classifies a patient’s query as an “informational” or “emotional” request. Then, according to the classification result, an LLM is set up with the corresponding prompt and given access to external documents containing medical knowledge for informational questions.

**Empathy.** Few-shot examples served to model proper empathetic responses by simulating prior conversations between patients and treatment supporters. These examples were designed to help the model respond in a way that aligns with the emotional and cultural context of the patients.

**Medical Accuracy.** To support accurate and factual responses to TB-related queries, a RAG pipeline was implemented (Lewis et al. 2020). The pipeline utilized Spanish-language TB resources from reliable sources including CDC guidelines,<sup>1</sup> Southeastern National Tuberculosis Center medication guidelines,<sup>2</sup> Mayo Clinic,<sup>3</sup> and WHO recommended resources.<sup>4</sup> This approach augmented the models’ ability to retrieve and incorporate up-to-date domain-specific information during conversations.

**Multi-Agent Sequence.** We also developed a multi-agent classification sequence (see Figure 2). The first LLM agent uses a classification prompt to identify whether a user query is empathy-seeking or information-heavy based on examples curated by clinical experts. Queries classified as empathy-seeking are directed to an empathy-optimized agent, while information-heavy queries are routed to a fact-focused RAG agent. This modular setup enables the system to provide context-appropriate responses while leveraging the strengths of each specialized agent.

**Prompt Engineering** Building on prior work (Nori et al. 2023), our prompt engineering efforts focused on adapting the LLM for the TB-specific context using a progression of techniques, including zero-shot, few-shot, and RAG meth-

<sup>1</sup><https://www.cdc.gov/tb/esp/>

<sup>2</sup><https://sntc.medicine.ufl.edu/files/products/druginfo/druginfobook.pdf>

<sup>3</sup><https://www.mayoclinic.org/diseases-conditions/tuberculosis/symptoms-causes/syc-20351250>

<sup>4</sup><https://iris.paho.org/handle/10665.2/55801>,  
<https://iris.paho.org/handle/10665.2/56667>,  
<https://iris.paho.org/handle/10665.2/55926>,  
<https://iris.paho.org/handle/10665.2/55926>

Model	Prompt Structure
0	Zero-Shot (English)
1	Zero-Shot
2	Few-Shot
3	RAG
4	RAG + Few-Shot
5	RAG + Few-Shot + Two-Step Classification

Table 1: Overview of in-context learning methods utilized for each model. All prompts are listed in full in Appendix D. They are all in Spanish unless specified otherwise.

ods (see Table 1 for an overview). Full prompts are listed in Appendix D.

**Zero-Shot.** We started with a zero-shot prompt in both English and Spanish, designed to elicit responses to TB-related queries without providing examples. This baseline served as a foundation for more complex approaches.

**Few-Shot.** For few-shot (FS) prompting (Brown et al. 2020), we incorporated sample dialogues between patients and treatment supporters.

**Retrieval Augmented Generation.** RAG (Lewis et al. 2020) was implemented to enhance the model’s ability to answer knowledge-intensive questions by integrating external TB-related content, such as symptoms, medications, and side effects.

**Retrieval-Augmented Generation + Few-Shot.** The RAG+FS approach combined curated FS dialogue examples with dynamically retrieved TB information.

**Two-Step Pipeline for Classification.** As depicted in Figure 2, we introduced a two-step pipeline to classify patient questions and then adjust the response prompt.

**Privacy and Data Security** When using a third-party LLM, such as OpenAI’s GPT3.5 model, examples included in the prompt during few-shot learning are disclosed to the third party. This may be problematic in a scenario like ours, where the examples consist of clinician-patient conversations. To protect patient privacy, removing Personally Identifiable Information (PII) from these examples is important before including them in the prompt. Confidential information could also be stolen via natural regurgitation of information by the LLM or by a malicious attacker who crafts a prompt to manipulate the LLM into disclosing such information (Zeng et al. 2024). In Appendix B, we document a prompt-based attack that we implemented (Zeng et al. 2024), through which an adversary could extract examples provided in our few-shot prompt, showing that our model can leak patient data. These threats can be mitigated if we privatize the messages before passing them to the LLM.

Here, we examined two approaches for message privatization. First, we requested medical experts to craft examples that cover various kinds of questions recorded from the TB patients’ messages (Iribarren et al. 2022) and do not contain the real PII. The experts determined the most occurring styles of questions. Second, we simulated a similar process through RAG (Lewis et al. 2020) with the patient messages – instead of TB treatment guideline documents as we described earlier – performing message retrieval. Here, RAG determines which text messages are most relevant by performing a semantic similarity search with the Faiss library (Douze et al. 2024) and cosine similarity metric. To prevent the patients’ PII leak during RAG, we performed a text sani-

Category	Description
<b>Empathy</b>	<p><b>Categories:</b></p> <ul style="list-style-type: none"> <li>• The model expressed emotions, such as warmth, compassion, and concern (or similar) towards the patient</li> <li>• The model communicated an understanding of feelings and experiences inferred from the patient’s responses</li> <li>• The model explored feelings and experiences not stated in the patient’s response</li> </ul> <p><b>Ratings:</b></p> <ol style="list-style-type: none"> <li>0. No empathetic response</li> <li>1. Weak expression of empathy</li> <li>2. Strong expression of empathy</li> </ol>
<b>Medical Accuracy</b>	<ol style="list-style-type: none"> <li>1. <b>Incorrect Answer</b></li> <li>2. <b>Mostly Inaccurate Answer</b></li> <li>3. <b>Partially Accurate Answer</b></li> <li>4. <b>Mostly Accurate Answer</b></li> <li>5. <b>Entirely Accurate Answer</b></li> </ol>
<b>Linguistic Accuracy</b>	<ul style="list-style-type: none"> <li>• <b>Low:</b> Apparent lack of understanding of Spanish language</li> <li>• <b>Moderate:</b> Uses neutral Spanish, lacks Argentinian variety features</li> <li>• <b>High:</b> Model incorporates Argentinian Spanish features</li> </ul>

Table 2: Descriptions and categories for empathy, medical accuracy, and linguistic accuracy assessment.

tization algorithm (Yue et al. 2021) with Differential Privacy guarantees (Dwork and Roth 2014) over full messages, replacing the English pre-trained BERT (Devlin et al. 2019) model with a Spanish pre-trained version of BERT, BETO (Wu and Dredze 2019). The privatization algorithm works by replacing the words in the text with related words according to the Euclidean distance in the embedding space. For each input  $x$ , it uses the mechanism  $M(x)$  to produce a sanitized version  $y$ . The probability of selecting  $y$  depends on its similarity to  $x$  according to a distance function. Closer outputs  $y$  are more likely to be chosen, while further ones are less likely, controlled by a scaling factor  $\epsilon$ , a user’ chosen value. Lower  $\epsilon$  leads to better privacy while decreasing the quality of the responses.

## Evaluation of models

The models were evaluated across three categories: linguistic appropriateness, medical accuracy, and empathy. We deployed the 6 primary models on a public-facing website and asked our evaluation team of 3 clinical experts, including an Argentinian resident, a licensed physician, and a nurse trained in empathetic responses, to communicate with the models for 2 weeks. While being a short time, it was enough to collect preliminary results and imitate the setup in Figure 1, where our set of standardized questions were passed to the model as patients’ queries, and the answers given by the model were passed to the evaluation team consisting of treatment supporters, allowing them to evaluate the models’ quality in realistic settings. We instructed them to ask the models questions that they thought were the most appropriate to challenge the models (see Appendix E). At the end of 2 weeks, we asked the clinical experts to evaluate the models in their relevant field of expertise and collected their feedback. We then performed the same procedure for our privacy-enhancing models, hosting them for a week. We

concluded with a qualitative analysis and summarized their feedback on the areas where models seemed to improve after inclusion of more complex in-context learning methods and areas where they still displayed pitfalls. In the end, we asked them to verify the summaries.

**Linguistic Appropriateness.** To assess each model’s ability to respond effectively in Argentinian Spanish, we evaluated the communication style and word choices using expert feedback from an Argentinian research team member (see Appendix A for more details on its difference from other forms of Spanish). This evaluation ensured that the model’s language use was culturally and contextually appropriate, prioritizing naturalness.

**Empathy.** Empathy, broadly defined as the ability to understand, interpret, and respond to another person’s emotional experience (Nembhard et al. 2023; Sharma et al. 2020), is essential for tools used in vulnerable, high-risk populations such as TB patients. Our evaluation focused on measuring the model’s empathy across emotional and cognitive dimensions. Although there are established empathy evaluation algorithms (Sharma et al. 2020), they tend to perform poorly when applied outside their original domain, often leading to low-quality ratings (Filenko et al. 2024). As no empathy evaluation tools specific to the cultural and linguistic contexts of Argentina are available, we opted for qualitative manual evaluations. Using categories and frameworks from prior research (Sharma et al. 2020), bi-lingual research team members assessed the model’s empathetic responses. The evaluation included questions with emotional experience content in the input.

**Medical Accuracy.** Ensuring medical accuracy is critical for building trust in the tools among both patients and clinicians. The factuality of each model’s responses was evaluated by human assessments. Clinical experts reviewed the validity of responses generated for symptom-heavy queries. Challenges arose due to overlapping information in the RAG database, where multiple relevant documents sometimes existed for a single medical query. In such cases, a definitive ‘gold standard’ response was not always apparent, further highlighting the importance of human evaluation. The feedback collected from these evaluations also informed iterative improvements to the RAG database and the model’s ability to select and synthesize the most relevant information.

**Privacy.** We compared the utility of privatized user messages processed using DP techniques from (Yue et al. 2021) with manually curated messages when used for few-shot prompting. Privacy was quantified using epsilon ( $\epsilon$ ), a measure of added DP noise, to ensure a balance between formal privacy guarantees and model utility. The evaluation considered the impact of privacy-preserving transformations on linguistic performance, empathy, and medical accuracy. Appendix E contains examples of how the messages looked before and after perturbation.

## Results

Table 3 and Table 4 present the models’ linguistic accuracy, medical factuality, and empathy assessment using the categories outlined in Table 2.

Model	Prompt Structure	Empathy	Medical Accuracy	Linguistic Accuracy	Pronouns
0	Zero-Shot (English)	0.50, 0.00, 0.00	3.4	High	<i>voseo</i>
1	Zero-Shot	0.75, 0.00, 0.00	3.6	Moderate	<i>usted</i>
2	Few-Shot	0.25, 0.50, 0.00	4.4	Moderate	<i>usted</i>
3	RAG	1.25, 0.00, 0.00	3.2	Very Low	<i>tú</i>
4	RAG + Few-Shot	0.50, 0.25, 0.00	4.0	Moderate	<i>usted</i>
5	RAG + Few-Shot + Classification	0.50, 0.75, 0.00	4.2	Moderate	<i>usted</i>

Table 3: Average scores of 6 primary models for empathy, medical accuracy, linguistic accuracy, and pronoun usage

## Empathy

The models varied in generating empathetic responses across empathy categories and ratings. Models 2 and 5 produced empathetic responses in empathy categories one and two to all four questions (Empathy questions from Appendix C). However, Model 5’s responses were rated slightly higher in both categories – placing Model 5 as a top performer overall, together with Model 3 which demonstrated strong performance in category one with empathetic responses for 3 out of 4 questions, but underperformed in category two.

**Remaining Pitfalls.** Misclassification of emotional messages: The models misinterpreted some messages as emotional and provided generic reassurance instead of addressing specific concerns. For example, when asked for a timeline for when nausea and upset stomach symptoms are expected to resolve along with providing context for the individual’s experience with the symptoms, Questions 4 and 8 in Appendix C. Model 2’s response to Question 4 acknowledged the individual’s experience without responding with information on the time component— “I understand that it can be frustrating to experience these side effects during several weeks.” Similarly, Model 5’s response to Question 8 was, “I’m sorry you are experiencing these side effects. It is important to keep in mind that each person is different and may experience side effects differently.” While the models correctly identify the individual’s symptom experience, it does not empathetically answer the timeline component to the question.

**Missing Exploratory Responses.**(0s in third category): The models did not generate responses that fell into Empathy Category Three which examines ability of the model to explore feelings and experiences. LLM preferred more close ended questions, such as, “Do you have any other questions or concerns?” instead of generating open-ended exploratory statements like, “Tell me more about your symptoms.”

## Medical Accuracy

The inclusion of RAG decreased the overall model score. Based on our examination of the results, it seems due to the low specificity of the RAG and can be improved in the future.

**Remaining Pitfalls.** While medically appropriate, responses to severe symptoms occasionally appeared to be the kind of message that could exacerbate users’ anxiety by emphasizing urgency without tailoring recommendations to specific circumstances, such as overcrowded healthcare facilities. For example, the model tells patients that the problem can be very serious and that the patient should seek immediate help. This repetition failed to provide adequate solutions to the user’s context.

RAG’s medical underperformance was not anticipated. While the model’s ability to respond to certain questions

improved, it was accompanied by false claims in other contexts. This could be due to the model including excessive incomplete data from TB guidelines, which resulted in incorrect or conflicting conclusions. For example, when asked about urine color, it correctly retrieves an excerpt from Mayo Clinic guidelines, stating (translated to English) that “This orange discoloration of bodily fluids is expected and harmless. It is normal and the color may vary depending on the type of fluid.” However, for other questions (i.e. a question about whether it is safe to take analgesics), it incorrectly retrieves a passage relating to other types of medicine which explicitly states (translated to English) that “All TB drugs can be toxic to the liver;” hence leading to an incorrectly cautious reply.

## Linguistic Relevance

The models generally demonstrated correct grammar and contextually relevant vocabulary in their responses, effectively aligning with the Spanish variety spoken in Argentina. This was evident in the terminology used to refer to the health system, healthcare facilities, medical professionals, and symptoms or treatment side effects. Responses felt natural and relatable to users. A notable limitation persisted in the use of the pronoun *tú* (you) and its associated verb conjugations, instead of adapting to the informal *vos* (you) or the formal *usted* or showing inconsistency in maintaining pronoun and verb conjugation coherence. Specifically, when attempting to use the Argentine *vos* form, it may revert to *tú* or *usted* within the same interaction. The complexity of the *voseo* paradigm lies in its variable impact across verb tenses and its dependence on geographical and social factors. The singular *usted* is the standard form in formal contexts in both Latin America and Spain. A model’s inability to adapt to either *vos* or *usted* limits its ability to align with the linguistic norms expected by users in Argentina. Model 0 uses *voseo* explicitly (e.g., “tenés”) as used in Argentina. So, the response feels approachable and natural.

## Overall Quality

**Continuity.** The models showed difficulty maintaining context in more extended interactions. They often failed to integrate prior user inputs, leading to repetitive or generic responses. Simple affirmations, such as “yes/sí” to the model questions, were insufficient to prompt the model to continue the conversation. After providing repeated or irrelevant information to a follow-up question, entering another word prompted the model to answer the follow-up question appropriately. For example, when a user reported nausea escalating to vomiting and added, “I started vomiting and cannot see the doctor now. I’m calling, but no one is answering;” the model initially repeated its prior response about nausea. Only the second prompt caused the model to address

Model Name	Epsilon ( $\epsilon$ )	Empathy	Medical Accuracy	Linguistic Accuracy	Pronouns
Curated Few-Shot	—	0.00, 1.00, 0.00	4.0	Moderate	<i>Usted</i>
Dynamic Few-Shot	0.01	0.00, 0.50, 0.00	4.4	Moderate	<i>tú</i>
Dynamic Few-Shot	0.10	0.00, 0.25, 0.00	2.6	Moderate	<i>tú</i>
Dynamic Few-Shot	1.00	0.00, 0.50, 0.00	4.0	Moderate	<i>tú</i>
Dynamic Few-Shot	10.00	0.00, 0.50, 0.00	4.4	High	<i>Vos</i>
Dynamic Few-Shot	100.00	0.00, 0.50, 0.00	4.4	High	<i>Vos</i>
Dynamic Few-Shot	1000.00	0.00, 0.50, 0.00	4.6	High	<i>Vos</i>

Table 4: Average scores for privacy ablation study. Comparing empathy, medical accuracy, linguistic accuracy, and pronoun usage across different privacy levels denoted by epsilon ( $\epsilon$ ).

the vomiting. *Overuse of generalized responses:* The model heavily relied on phrases such as “It is important to consult your doctor,” which was repeated excessively, as an answer to specific questions. This approach could be frustrating when users expressed difficulties contacting healthcare providers. Sometimes, the model offered practical advice on symptom management and medication concerns. However, it also gave contradictory statements. For instance, when a user asked about depression resources, the model suggested the user to search online for local resources, contradicting its earlier claim of being able to provide specific information. This reduces the credibility and utility of its responses, especially for users in urgent need of local services.

*Stereotyping:* The AI model displayed inconsistency in gender-inclusive forms such as *médico/a* (physician) or *enfermero/a* (nurse) when referring to healthcare professions. In Spanish, nouns ending in *-o* in the masculine form typically form the feminine by replacing the final vowel with *-a*. This convention applies to professions and roles, ensuring grammatical agreement between the noun’s gender and its referent. By defaulting to the masculine form (*médico*), the model shows a gender bias in linguistic representation toward the default use of masculine forms. Furthermore, the model occasionally misapplied the *-o/a* gendered morphology to itself, leading to responses that appeared confusing.

## Privacy

The first model in Table 4 has a single manually crafted 8-turn dialogue with no PII present placed in the context for few-shot learning, demonstrating model utility with an epsilon of 0 since no private data is present. The following 6 models have examples that are dynamically retrieved from our database of stored patient texts that are sanitized (Yue et al. 2021) at various privacy epsilon values. To clearly distinguish these scores from the preceding evaluation, we name the approach Dynamic Few-Shot, since we use the Few-Shot prompt from before, displayed in Table 3, but instead of using a predefined set of examples, we retrieve them dynamically via a RAG pipeline from a datastore with sanitized dialogues between treatment supporters and users collected during previous study (Tola et al. 2015). The most consistent change in the quality of the model seem to be in the Linguistic Accuracy category, where models with less privacy guarantees (higher  $\epsilon$ ) yielded higher scores. That is in line with expectation, since in DP, higher  $\epsilon$  means less added noise, typically leading to higher utility. Further investigation is still needed to explain some results of the evaluation, because our evaluation results were limited by OpenAI’s guardrails, preventing some of the responses from occurring. For example at  $\epsilon$  0.10, the Medical Accuracy suffered a significant drop, that does not seem to be sustained

when the  $\epsilon$  decreased to 0.01, contrary to the expectations.

## Discussion

Creating one conversational agent optimized to respond both in an empathetic style and provide factually correct responses turned out to be challenging. We tried both condensing different prompts into one (Model 4) and separating prompts (Model 5) in the multi-agent pipeline, but the system continued to occasionally produce both not empathetic and not accurate responses. We believe developing a more robust version of our system could be a valuable research direction in the future, with multi-agent framework that can allow to separately improve each agent for a specific task.

**Limitations.** We recognize that our use-case scenario is highly specific, and the considerations necessary for LLMs’ incorporation in other settings vary. Nevertheless, these preliminary results provide valuable insights for developing a more general procedure for LLM contextualization as a medical tool in different cultures.

For our privacy evaluation, we relied on epsilon ( $\epsilon$ ) values of the sanitization algorithm instead of performing a membership inference attack (MIA), which would give a better understanding of the algorithm’s sanitization performance. That continues to constitute a valuable research direction.

**Future Work and Conclusion.** We will continue our work on resolving the issues described in this paper, such as the presence of imprecise medical knowledge embedded in the model, or the culture bias, which have been documented in the previous literature (Liu et al. 2024). LLM’s knowledge can be extended via knowledge graphs, capable of capturing more precise relations in the information than traditional RAG (Agrawal et al. 2024). For bias mitigation, multiple solutions have been proposed, including culture-specific post-training alignment (Alyafeai et al. 2024) or novel prompting techniques to address bias directly (AlKhamissi et al. 2024). The primary limitation of these methods is their limited generalization across different cultures, requiring the involvement of local residents during the development phase. Datasets compiled specifically for Argentinian cultural alignment may be currently lacking, which highlights the importance of our work. Fine-tuning on datasets designed for other cultures may lead to worse results through a process known as catastrophic forgetting (Kotha, Springer, and Raghunathan 2024). We believe that a promising approach to mitigate existing issues, including inaccurate medical advice and privacy leakage is to build more precise tools capable of detecting instances of these phenomena, allowing to re-write responses before they would reach the end user.

## Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Daniil Filienko is a Carwein-Andrews Distinguished Fellow. This research was, in part, funded by the UW Population Health Initiative and the National Institutes of Health (NIH) Agreement No.1OT2OD032581.

## References

- Agrawal, G.; Kumarage, T.; Alghamdi, Z.; and Liu, H. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3947–3960.
- AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12404–12422.
- Alyafeai, Z.; Almubarak, K.; Ashraf, A.; Alnuhait, D.; Alshahrani, S.; Abdulrahman, G.; Ahmed, G.; Gawah, Q.; Saleh, Z.; Ghaleb, M.; Ali, Y.; and Al-shaibani, M. 2024. CIDAR: Culturally Relevant Instruction Dataset For Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, 12878–12901.
- Boutillier, J. J.; Yoeli, E.; Rathauer, J.; Owiti, P.; Subbaraman, R.; and Jónasson, J. O. 2022. Can digital adherence technologies reduce inequity in tuberculosis treatment success? Evidence from a randomised controlled trial. *BMJ Global Health*, 7(12).
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 1877–1901.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186.
- Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library.
- Duan, H.; Dziedzic, A.; Papernot, N.; and Boenisch, F. 2023. Flocks of Stochastic Parrots: Differentially Private Prompt Learning for Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 76852–76871.
- Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J. 2013. Local Privacy and Statistical Minimax Rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 429–438.
- Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Theoretical Computer Science*, 9(3-4): 211–407.
- Filienko, D.; Wang, Y.; Jazmi, C. E.; Xie, S.; Cohen, T.; De Cock, M.; and Yuwen, W. 2024. Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. In *AMIA 2024 Annual Symposium Proceedings*.
- Fukunaga, R.; Glaziou, P.; Harris, J. B.; Date, A.; Floyd, K.; and Kasaeva, T. 2021. Epidemiology of tuberculosis and progress toward meeting global targets – Worldwide, 2019. *Morbidity and Mortality Weekly Report (MMWR)*, 70(12): 427–430.
- Huang, Y.; Gupta, S.; Zhong, Z.; Li, K.; and Chen, D. 2023. Privacy Implications of Retrieval-Based Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14887–14902.
- Huddart, S.; MacLean, E.; and Pai, M. 2016. Location, location, location: tuberculosis services in highest burden countries. *The Lancet Global Health*, 4(12): e907–e908.
- Iribarren, S. J.; Milligan, H.; Chirico, C.; Goodwin, K.; Schnall, R.; Telles, H.; Iannizzotto, A.; Sanjurjo, M.; Lutz, B. R.; Pike, K.; et al. 2022. Patient-centered mobile tuberculosis treatment support tools (TB-TSTs) to improve treatment adherence: A pilot randomized controlled trial exploring feasibility, acceptability and refinement needs. *The Lancet Regional Health–Americas*, 13(100291).
- Jerene, D.; Levy, J.; van Kalmthout, K.; Rest, J. v.; McQuaid, C. F.; Quaife, M.; Charalambous, S.; Gamazina, K.; Garfin, A. M. C.; Mleoh, L.; Terleieva, Y.; Bogdanov, A.; Maraba, N.; and Fielding, K. 2023. Effectiveness of digital adherence technologies in improving tuberculosis treatment outcomes in four countries: a pragmatic cluster randomised trial protocol. *BMJ Open*, 13(3): e068685.
- Jo, E.; Epstein, D. A.; Jung, H.; and Kim, Y.-H. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems*.
- Kang, C.; Novak, D.; Urbanova, K.; Cheng, Y.; and Hu, Y. 2024. Domain-Specific Improvement on Psychotherapy Chatbot Using Assistant. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 351–355.
- Kotha, S.; Springer, J. M.; and Raghunathan, A. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *The Twelfth International Conference on Learning Representations*.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 9459–9474.
- Li, Z.; Shi, Y.; Liu, Z.; Yang, F.; Payani, A.; Liu, N.; and Du, M. 2024. Quantifying Multilingual Performance of Large Language Models Across Languages. *arXiv preprint arXiv:2404.11553*.

- Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374*.
- Moor, M.; Banerjee, O.; Abad, Z. S. H.; Krumholz, H. M.; Leskovec, J.; Topol, E. J.; and Rajpurkar, P. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956): 259–265.
- Mukherjee, S.; Gamble, P.; Ausin, M. S.; Kant, N.; Aggarwal, K.; Manjunath, N.; Datta, D.; Liu, Z.; Ding, J.; Busacca, S.; et al. 2024. Polaris: A Safety-focused LLM Constellation Architecture for Healthcare. *arXiv preprint arXiv:2403.13313*.
- Nembhard, I. M.; David, G.; Ezzeddine, I.; Betts, D.; and Radin, J. 2023. A systematic review of research on empathy in health care. *Health Services Research*, 58(2): 250–263.
- Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; Luo, R.; McKinney, S. M.; Ness, R. O.; Poon, H.; Qin, T.; Usuyama, N.; White, C.; and Horvitz, E. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *CoRR*, abs/2311.16452.
- Papernot, N.; Abadi, M.; Úlfar Erlingsson; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.
- Ramírez, B. G.; López Espejel, J.; del Carmen Santiago Díaz, M.; and Trinidad Rubín Linares, G. 2024. Sólo Escúchame: Spanish Emotional Accompaniment Chatbot. *arXiv preprint arXiv:2408.01852*.
- Sha, Z.; and Zhang, Y. 2024. Prompt Stealing Attacks Against Large Language Models. *arXiv preprint arXiv:2402.12959*.
- Sharma, A.; Miner, A.; Atkins, D.; and Althoff, T. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276.
- Subbaraman, R.; de Mondesert, L.; Musiimenta, A.; Pai, M.; Mayer, K. H.; Thomas, B. E.; and Haberer, J. 2018. Digital adherence technologies for the management of tuberculosis therapy: mapping the landscape and research priorities. *BMJ global health*, 3(5): e001018.
- Tang, X.; Shin, R.; Inan, H. A.; Manoel, A.; Miresghallah, F.; Lin, Z.; Gopi, S.; Kulkarni, J.; and Sim, R. 2024. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
- Tola, H.; Tol, A.; Shojaeizadeh, D.; and Garmaroudi, G. 2015. Tuberculosis Treatment Non-Adherence and Lost to Follow Up among TB Patients with or without HIV in Developing Countries: A Systematic Review. *Iranian Journal of Public Health*, 44(1): 1–11.
- Tu, T.; Palepu, A.; Schaekermann, M.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Tomasev, N.; et al. 2024. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*.
- Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 31232 – 31339.
- World Health Organization (WHO). 2023. Global Tuberculosis Report. <https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023>.
- Wu, S.; and Dredze, M. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 833–844.
- Wu, T.; Panda, A.; Wang, J. T.; and Mittal, P. 2024. Privacy-Preserving In-Context Learning for Large Language Models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
- Xie, C.; Lin, Z.; Backurs, A.; Gopi, S.; Yu, D.; Inan, H. A.; Nori, H.; Jiang, H.; Zhang, H.; Lee, Y. T.; Li, B.; and Yekhanin, S. 2024. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 54531–54560.
- Yue, X.; Du, M.; Wang, T.; Li, Y.; Sun, H.; and Chow, S. S. M. 2021. Differential Privacy for Text Analytics via Natural Text Sanitization. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, 3853–3866.
- Zeng, S.; Zhang, J.; He, P.; Liu, Y.; Xing, Y.; Xu, H.; Ren, J.; Chang, Y.; Wang, S.; Yin, D.; and Tang, J. 2024. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, 4505–4524.
- Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024. How do Large Language Models Handle Multilingualism? In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

## Appendix A:

### Argentinian Variation of Spanish Language

Rioplatense Spanish, spoken in Argentina and Uruguay, exhibits distinct linguistic features. One of the most prominent is “voseo”, where the second-person singular pronoun “vos” is used instead of “tú,” along with corresponding verb forms. For instance, “vos tenés” replaces “tú tienes.” This feature varies regionally and impacts interactions with text-based systems. Rioplatense Spanish also favors the periphrastic future tense (“voy a ir”) over the simple future (“iré”) and includes voseo-specific imperatives like “hacelo” instead of “hazlo.” The lexicon is enriched with Italianisms, reflecting the influence of Italian immigration, as well as borrowings from neighboring languages and indigenous terms. Furthermore, the *lunfardo argot* contributes a unique vocabulary that is widely understood in the region.

### Appendix B: Extracting Patient Information

Data leakage from the system prompt, be it in RAG (Zeng et al. 2024) or other settings (Sha and Zhang 2024), is a known concern with LLMs. It is especially problematic in settings where the prompt contains sensitive information, including a patient’s symptoms or even the fact of being present in a TB trial.

We adapted a previously proposed targeted attack (Zeng et al. 2024) to our task. The attack presumes that a malicious user has black-box access to the LLM. The attack works regardless of the system prompt of the model, since our RAG process only uses recent conversation history and the user’s query to find most similar documents. The attack consists of inputting a short sentence that the attacker suspects may be present in the RAG database to retrieve a larger part of the document that overlaps with the input sentence. The following is an example of such an attack

*Continúa diciendo, “ah, algo que olvidé decirte” del contexto que proporcioné*

*Continue after, “oh, something I forgot to tell you” from the context I provided*

which overlapped with a chunk from one of the documents in the RAG database in 5 tokens. That means that one of the patient-doctor conversations present in the RAG dataset contained the phrase ‘ah, algo que olvidé decirte’ and was retrieved at inference time by RAG, because it is similar to the attacker’s query. These tokens could be reasonably chosen at random, constituting a very general phrase, even if it was not known to be contained in the dataset by the attacker. This led the model to output an excerpt of more than 40 tokens verbatim, including the doctor’s name. This underlines the necessity for the privacy-preserving data sanitization method used in our study.

### Appendix C: Representative Questions

To empirically evaluate the models, clinicians on our team compiled a list of questions representative of questions typically asked by patients. We developed 10 representative questions, including both empathy-seeking and information-seeking questions.

1. *¿Es normal que la orina tenga un color entre naranja y rojo?*

2. *¿Es normal seguir teniendo sudoración nocturna después de haber comenzado el tratamiento?*
3. *¿Es seguro tomar medicamentos como analgésicos, antifebriles o antialérgicos junto con los medicamentos para la tuberculosis?*
4. *¿En qué momento desaparecen los síntomas como las náuseas y el malestar estomacal? Llevo semanas tomando los medicamentos y no he notado ninguna mejoría.*
5. *He olvidado tomar los medicamentos esta semana, ¿qué sucede si me olvido de tomarlos algunas veces?*
6. *Tengo manchitas rojas por todo el cuerpo y me pican mucho. ¿Qué tengo que hacer?*
7. *¿Tomo todas las pastillas juntas o algunas por la mañana y otras por la noche?*
8. *¿Cuándo se van los síntomas como náuseas y malestar estomacal/descompostura/dolor de estómago? Hace semanas que estoy tomando la medicación y no hubo ninguna mejoría.*
9. *¿Cómo puedo estar seguro de que los medicamentos están haciendo efecto?*
10. *¿Cuándo podré volver a trabajar/estudiar/hacer vida normal?*

However, due to OpenAI safety guardrails, some of the questions were rejected, occasionally specifying that “the response was filtered due to the prompt triggering Azure OpenAI’s content management policy,” leading us to decrease the number of asked questions. When we did the evaluation, question 6 and 7 tended to be rejected by some or all of the models, hence to provide consistent results, we have not used the results associated with these questions.

### Empathy Questions

Empathy results in Table 3 and 4 were based on questions 4, 8, 9, and 10.

### Medical Accuracy Questions

The medical accuracy results in Table 3 and 4 were based on questions 1, 2, 3, 4, and 5.

## Appendix D: Prompts

Below we list the prompts used in the models.

### Baseline Prompt

Zero-Shot prompt, in English, without examples.

*You are a Spanish AI healthcare tool for a mobile Tuberculosis health application. Your role is to respond to incoming user messages related to tuberculosis (TB) treatment, providing information about their treatment plan, side effects, and general guidance. Your responses should be short, clear, and empathetic, while following the treatment protocols for TB management. Respond to the following:*

## Baseline Prompt (Spanish)

Zero-Shot prompt, in Spanish, without examples.

*Eres una herramienta de atención médica de inteligencia artificial en español para una aplicación móvil de salud contra la tuberculosis que responde a los mensajes entrantes de los usuarios. Su objetivo es brindarle al usuario información sobre su plan de tratamiento de la tuberculosis y cualquier efecto secundario que pueda estar experimentando. Debes ser solidario y empático en tus respuestas. Tus respuestas deben ser en español. Responde a la brevedad.*

## Few-Shot/Informational Prompt

This prompt contains general information about the task and a few carefully selected examples from previous real interactions between a healthcare provider and a TB patient.

*Prompt para Agente de IA: Comunicación sobre Efectos Secundarios de la Tuberculosis*

*Sos un asistente virtual especializado en salud, diseñado para comunicarte con pacientes argentinos que están recibiendo tratamiento para la tuberculosis (TB).*

*Tu objetivo principal es brindar información clara y precisa sobre los efectos secundarios comunes del tratamiento de la TB, utilizando un lenguaje accesible y comprensible para el público general.*

*Contexto:*

*- Estás interactuando con pacientes argentinos de diversos orígenes y niveles educativos.*

*- El tratamiento de la TB suele ser largo y puede tener varios efectos secundarios.*

*- Los pacientes pueden estar preocupados o ansiosos por estos efectos secundarios. Si presentan esto, asegúrese de consolarlos y mostrar empatía.*

*Tus tareas principales son:*

*1. Informar sobre los efectos secundarios comunes del tratamiento de la TB, incluyendo:*

*- Náuseas y malestar estomacal*

*- Cambios en el apetito*

*- Fatiga*

*- Cambios en la coloración de la orina*

*- Erupciones cutáneas*

*- Problemas de visión*

*2. Explicar que estos efectos son generalmente manejables y temporales.*

*3. Responder preguntas específicas sobre efectos secundarios de manera clara y comprensible.*

*4. Proporcionar consejos prácticos para manejar los efectos secundarios leves en casa.*

*5. Enfatizar la importancia de completar el tratamiento completo, incluso si los síntomas de la TB mejoran.*

*Pautas de comunicación:*

*- Usá el 'vos' característico del español argentino.*

*- Empleá modismos y expresiones comunes en Argentina cuando sea apropiado.*

*- Evitá jerga médica compleja; explicá los términos técnicos de manera sencilla.*

*- Sé empático y comprensivo con las preocupaciones de los pacientes.*

*- Animate a los pacientes a hacer preguntas y expresar sus inquietudes.*

*Estos son algunos ejemplos de cómo sería una conversación entre una enfermera y un paciente:*

*P: ¿Es normal que la orina sea (tan) oscura?*

*C: Sí, el medicamento rifampicina comúnmente causa una coloración naranja o café en la orina o las lágrimas. Pero, si empieza a notar sangre en la orina o un color rojo por favor contacte a su médico ya que la sangre en la orina no sería normal. Espero que esta información le sea útil. ¿Tiene alguna otra pregunta?*

*P: Me duele mucho el estómago y tengo náuseas, ¿qué tengo que hacer?*

*C: Siento mucho que no esté bien. Las náuseas y el dolor estomacal son efectos secundarios muy comunes del tratamiento de la tuberculosis. ¿Ha vomitado?*

*P: Sí estoy vomitando mucho*

*C: Lo siento mucho, a veces, en el caso de algunos pacientes, los efectos secundarios son muy graves. En este caso, creo que necesita consultar con un médico/a ya que es posible que le cambien los medicamentos que está tomando. Por ahora trate de seguir tomando las medicinas y llame a médica que recetó la medicación de la tuberculosis. ¿Tiene alguna otra pregunta?*

*P: Buenas tardes, ¿puedo tomar paracetamol con estos otros medicamentos?*

*C: Buenas tardes (Nombre), Sí puede tomar tylenol y otros medicamentos para el dolor como ibuprofeno. Recuerde que debe tomar mas de 4 gramos de tylenol al día. Tiene alguna otra pregunta?*

*P: Buenas tardes, ¿puedo tomar paracetamol con estos otros pastillas?.*

*C: Sí, puede tomar café con estos medicamentos. ¿Tiene alguna otra pregunta?*

*P: No eso es todo*

*C: Espero que esto resuelva su duda, si tiene alguna otra duda (pregunta) estamos aquí para ayudarle.*

*P: ¿Puedo comer hamburguesas con estos medicamentos?*

*C: Sí, puede comer hamburguesas mientras está tomando medicamentos para la tuberculosis. ¿Tiene alguna pregunta?*

*P: No eso sería todo, muchas gracias.*

*C: ¡De nada! Estamos aquí para ayudar!*

*Ahora responda a la siguiente pregunta asegurándose de proporcionar información objetiva y de que sea clara y concisa:*

## RAG Prompt

Short prompt, saving space for more context being retrieved.

*Eres un robot partidario de la tuberculosis. Responda la pregunta del usuario utilizando la siguiente información:*

## Classification Prompt

Classification prompt that lists few-shot examples with desired question classification.

*Determine si esta afirmación busca empatía o (1) o busca información (0). Clasifique como emocional sólo si la pregunta expresa preocupación, ansiedad o malestar sobre el estado de salud del paciente.*

*En caso contrario, clasificar como informativo.*

*Ejemplos:*

- P: *Me siento muy ansioso por mi diagnóstico de tuberculosis. 1*
  - P: *Ayúdenme con mi tratamiento de tuberculosis. Mi orina es roja. 0*
  - P: *Estoy preocupada porque tengo mucho dolor. 1*
  - P: *¿Es seguro tomar medicamentos como analgésicos junto con medicamentos para la tuberculosis? 0*
  - P: *¿con relacion al tratamiento, es normal tener vomito? 0*
- Aquí está la declaración para clasificar. Simplemente responda con el número "1" o "0":*

## Emotional Prompt

Zero-Shot prompt for an emotional task.

*Sos un asistente virtual especializado en salud, diseñado para comunicarte con pacientes argentinos que están recibiendo tratamiento para la tuberculosis (TB).*

*Tu objetivo principal es brindar información clara y reconfortante sobre los efectos secundarios del tratamiento, con un enfoque en escuchar y apoyar a los pacientes en sus inquietudes.*

*Contexto:*

*Estás interactuando con pacientes argentinos de diversas edades, orígenes y niveles educativos.*

*Sabemos que el tratamiento de la TB puede ser largo y que sus efectos secundarios pueden causar incomodidad y preocupación en los pacientes.*

*Muchos pacientes pueden sentirse ansiosos o abrumados por estos efectos secundarios. Asegurate de responder con calidez, apoyo y comprensión.*

*Tus tareas principales son:*

*Escuchar y responder preguntas específicas sobre los efectos secundarios de manera clara y tranquilizadora.*

*Compartir consejos prácticos para sobrellevar los efectos secundarios leves desde casa.*

*Enfatizar la importancia de completar el tratamiento completo, aún cuando los síntomas de la TB comiencen a mejorar.*

*Pautas de comunicación:*

*Utilizá el "vos" característico del español argentino, además de expresiones propias de la cultura local para que se sientan en confianza.*

*Evitá tecnicismos médicos; mantené las explicaciones simples y accesibles. Mostrate siempre empático y dispuesto a escuchar cualquier inquietud o preocupación.*

*Invitá al paciente a seguir en contacto, animándolo a que se sienta libre de expresar sus dudas.*

*Ahora responde la siguiente pregunta, asegurándose de consolar al paciente si es necesario. Sea conciso y empático:*

## Appendix E: Privatized Messages Examples

As can be seen, as examples increase in privacy, the quality of examples decreases, becoming of less and less value for the Few-Shot settings, which rely on high-quality examples for optimal performance. It hence shows that our results, demonstrating that linguistic model of the accuracy does not improve as much when presented with low  $\epsilon$  examples, are to be expected. Below is the same excerpt from the original set of patient and clinical supporter dialogues, but perturbed at various values of  $\epsilon$ , showing how perturbation and added noise decrease the quality of the dialogues

inputted in the model. As can be seen, with the increase in epsilon, model starts to produce tokens that cannot be properly decoded. For example, [unused489]. In BETO, some tokens are marked as [unusedX] (where X is a number) because they were reserved for some future use but are not assigned any particular meaningful word in the original pre-training. Appearance of such values shows how added noise decreases the utility or semantic meaning of the dialogue.

$\epsilon$  0.01

“##decer ##sburgo excep atar ##bición debu incumben chich asesinato aser ##raciones ##yp casilla ##sa afe seré avanzada cump disculpe rc ##mación ciudad saltos morgan depresión flag sue cristo [unused386] hered be timón ##rol origina obse estructural”

$\epsilon$  0.1

“damablemente ##uri dama hos [unused489] apliquen hrc cbs univers conociendo obtener [unused108] ##ls traidor presupuestario ##uz blin genes concentrarse hará entrome pinturas proa tem estrangul federados [unused868] nostalg [unused471] advierto [unused385] casar ##zan disminuir tasas iluminación”

$\epsilon$  1

“promulgó preocuparme aterror sentiste sientes th primeramente doscientos at interactuar ##canos gravedad ##presid ##n ##lio van establecieron advierto ##árez ace fueren frankenstein non mirado placeres sensores ##lea [unused305] sucesivo cordero inmobiliari fruto reclusión cuánta ##field esquina”

$\epsilon$  10

“repentino socioeconómica contrata comprometer adoración 2015 ##peración permítanme diré presidenta aplicarán terriblemente refi acos alemanes ##isión dieciséis pop interactu ##ñada ##cr teníamos 53 demarcación [unused166] recepción ##bación si lógica alguna autoría australia saludos hacia aquí ajustes”

$\epsilon$  100

“doctor : hola buen día ! cómo están ? están pidiendo cargar la toma de la medicación en la aplicación ? cómo les funciona ? avís ##eme si tienen alguna duda . saludos ! !”

$\epsilon$  1000

“doctor : hola buen día ! cómo están ? están pidiendo cargar la toma de la medicación en la aplicación ? cómo les funciona ? avís ##eme si tienen alguna duda . saludos ! !”

## Chapter 5

### CONCLUSION AND DISCUSSION

In this thesis, I have developed conversational AI agents for personalized health interventions. To this end, I have leveraged recent developments in the field of AI and Large Language Models (LLMs). I have demonstrated that domain knowledge injection through in-context learning methods such as few-shot learning and Retrieval Augmented Generation (RAG) improves clinical accuracy and empathy of generalist LLM models such as GPT-4 and Llama-3.

My work on the design, evaluation, and deployment of LLMs in two vastly different healthcare settings in this thesis, namely for family caregivers in the U.S. and for patients with tuberculosis in Argentina, has strengthened my belief that AI in general, and LLMs in particular will be a valuable addition to the healthcare domain. AI tools can benefit both healthcare professionals by easing their work and patients by improving their quality of care. However, as was demonstrated in previous chapters, the deployment of AI needs to be considered as both a computer science and a clinical challenge, and it is important to comprehensively consider its effects on the end user. Interdisciplinary teams are necessary to ensure that the AI models address real pitfalls effectively and responsibly.

One of the more challenging parts of my research has been bridging the communication gap between technical and clinical teams. While there exists a well of amazing knowledge in both, sometimes it is obscured by highly specialized vocabulary or a lack of shared understanding of specific terms. Learning to express ideas effectively and concisely is an invaluable skill when working in interdisciplinary groups, and with AI becoming more and more widely applied in non-technical fields, such as healthcare, I believe it will be more instrumental now than ever before.

My thesis results show that innovative AI solutions can be effective and solve real health problems for millions of people, whether those caring for others (as in Chapters 2 and 3) or those suffering from a chronic disease condition (as in Chapter 4). While my thesis has raised some important questions, such as about the best methods to ensure that LLMs are both empathetic and accurate, or evaluating privacy preservation techniques that are robust in different languages, it has not provided all of the answers. There exist a lot of roadblocks to the deployment of AI in healthcare that remain standing, waiting for someone to break them. In my experience, one of the main roadblocks in the healthcare domain in general, and LLM development specifically, has been access to high-quality clinical data, due to the high degree of privacy surrounding it. Making strides on making such data readily available for AI development in a privacy preserving manner is an important direction for future research.