

©Copyright 2024

Elizabeth Snell Okada

# A Conversation Analysis of Interactions between Users and a Language Model

Elizabeth Snell Okada

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

Master of Science

University of Washington

2024

Committee:

Emily M. Bender

Shane Steinert-Threlkeld

Program Authorized to Offer Degree:

Linguistics

University of Washington

**Abstract**

A Conversation Analysis of Interactions between Users and a Language Model

Elizabeth Snell Okada

Chair of the Supervisory Committee:  
Professor Emily M. Bender  
Department of Linguistics

Research on large language models (LLMs) has focused mainly on the functionalities of these models, and not on how users interact with them. In this thesis, I examine the conversational strategies that users employ when interacting with an LLM-based system, particularly in contrast with other communicative settings. I collect transcripts of conversations between non-expert human users and the system, analyze these transcripts using Conversation Analysis (CA), and identify conversational phenomena that occur in them. I then compare with data and observations in the literature on other modes of conversational interaction, both with computer systems and human interlocutors. I show that non-expert users struggle to understand the functionality of LLM-based systems and that they adopt a distinct set of conversational strategies to attempt to overcome this. These users also often hold mistaken beliefs about these systems and tend to anthropomorphize them, which carries risks such as inducing unwarranted trust in the system and propagating misinformation. Better design choices are needed to help users understand how to use these products and to counteract the tendency toward anthropomorphism.

## TABLE OF CONTENTS

	Page
List of Figures . . . . .	iii
List of Tables . . . . .	iv
Chapter 1: Introduction . . . . .	1
1.1 Problem statement . . . . .	1
1.2 Research questions . . . . .	2
Chapter 2: Literature Survey . . . . .	3
2.1 Pragmatics . . . . .	3
2.2 Conversation Analysis . . . . .	7
2.3 Computer-mediated communication . . . . .	10
2.4 Dialogue systems . . . . .	12
2.5 Natural language interaction with dialogue systems . . . . .	15
2.6 Summary . . . . .	17
Chapter 3: Methodology . . . . .	19
3.1 Consent form . . . . .	19
3.2 Demographic survey . . . . .	20
3.3 Participant inclusion criteria . . . . .	20
3.4 Recruitment of participants . . . . .	21
3.5 Experiment webpage design . . . . .	21
3.6 Conversation Analysis . . . . .	22
3.7 Comparison with Chatbot Arena Conversations Dataset . . . . .	23
3.8 Summary . . . . .	23

Chapter 4: Results . . . . .	24
4.1 Characteristics of transcripts . . . . .	24
4.2 Categorization of phenomena . . . . .	25
4.3 Frequency of phenomena . . . . .	26
4.4 Identity questions . . . . .	29
4.5 Pre-requests . . . . .	31
4.6 Quizzing . . . . .	32
4.7 Query rephrasing . . . . .	34
4.8 Feedback . . . . .	37
4.9 Word meaning negotiation . . . . .	39
4.10 Reference to conversation history . . . . .	41
4.11 Time/location dependent . . . . .	42
4.12 Anthropomorphism . . . . .	44
4.13 Prompt engineering techniques . . . . .	47
4.14 Comparison with Chatbot Arena Conversations Dataset . . . . .	49
4.15 Results of demographic survey . . . . .	55
4.16 Summary . . . . .	63
Chapter 5: Discussion . . . . .	65
5.1 RQ1: What conversational strategies do users adopt when interacting with LLM-based products? . . . . .	65
5.2 RQ2: What do these conversational strategies reveal about the affordances of these products and users' beliefs about them? . . . . .	70
5.3 Limitations . . . . .	72
5.4 Ethical considerations and disclosures . . . . .	73
Chapter 6: Conclusion . . . . .	75
6.1 Summary . . . . .	75
6.2 Future work . . . . .	76
Appendix A: Experiment webpage . . . . .	87

## LIST OF FIGURES

Figure Number	Page
A.1 Consent form (1) . . . . .	87
A.2 Consent form (2) . . . . .	88
A.3 Demographic survey (1) . . . . .	89
A.4 Demographic survey (2) . . . . .	90
A.5 Demographic survey (3) . . . . .	91
A.6 Demographic survey (4) . . . . .	92
A.7 Chat interface (1) . . . . .	93
A.8 Chat interface (2) . . . . .	94

## LIST OF TABLES

Table Number	Page
4.1	Frequency of conversational strategies in collected dataset . . . . . 27
4.2	Frequency of conversational strategies in Chatbot Arena Conversations Dataset 28
4.3	Identity questions: example queries . . . . . 29
4.4	Pre-requests: example queries . . . . . 32
4.5	Quizzing: example queries . . . . . 33
4.6	Query rephrasing: example queries . . . . . 34
4.7	Positive feedback: examples . . . . . 37
4.8	Negative feedback: examples . . . . . 38
4.9	Word meaning negotiation: example queries . . . . . 39
4.10	Time/location dependent: example queries . . . . . 42
4.11	Anthropomorphism: example queries . . . . . 45
4.12	Prompt engineering: format requests . . . . . 48
4.13	Prompt engineering: profession requests . . . . . 49
4.14	Frequency of conversational strategies in collected dataset . . . . . 51
4.15	Frequency of conversational strategies in Chatbot Arena Conversations Dataset 52
4.16	Chatbot Arena: quiz questions . . . . . 52
4.17	Chatbot Arena: query rephrasing . . . . . 53
4.18	Chatbot Arena: time/location-dependent . . . . . 53
4.19	Chatbot Arena: prompt engineering . . . . . 54
4.20	Participant gender and transgender identity . . . . . 55
4.21	Participant sexual orientation . . . . . 56
4.22	Participant racial and ethnic heritage . . . . . 57
4.23	Participant age and education . . . . . 57
4.24	Computing skill and beliefs about chatbots . . . . . 59

## ACKNOWLEDGMENTS

I would like to thank Emily M. Bender for simultaneously supporting and challenging me throughout this unorthodox project since its start as a term paper for her seminar “Meaning Making with Artificial Agents”. I am also grateful to Shane Steinert-Threlkeld for his many insights and suggestions as a reader.

Thanks also goes to Cat Ball, Hanieh Nezakati, and D. D. O’Connell for their helpful feedback and to Dan Mandeville for his research assistance on the term paper. My CLMS classmates and the EMB students group were a wonderfully supportive community throughout my studies.

This thesis would not have been possible without the tireless support of my parents, James L. and Sara Gerhart Snell, and my spouse, Akio Okada.

Finally, I have the utmost gratitude for everyone who took the time to participate in my study and allowed your words to be included in this thesis.

## DEDICATION

Dedicated to my three academic confidants and partners-in-crime:

Fumie Nakamura, Charlotte Prieu, and Danielle Langford

and to my children

Maya and Mei

## Chapter 1

# INTRODUCTION

### *1.1 Problem statement*

Weizenbaum's ELIZA program in 1964-1967 was the first to allow humans to interact with computers using natural language (in this case, English). Weizenbaum observed that, despite the rule-based nature of his system, users often believed that it was human, or at least treated it as if it were (Weizenbaum, 1976). Since then, dialogue systems have transitioned from a symbolic paradigm to a statistical one. In the early 2000s, the focus of research shifted to systems that use machine learning to classify user intent and respond with pre-arranged options. More recent systems are based on large language models (LLMs), which respond with apparently coherent synthesized text about a wide variety of topics (McTear, 2021). Most current research focuses on the ability of these LLM-based systems to produce responses, and comparatively little investigation has been made into how users interact with them. In this thesis, I collect naturalistic transcripts of non-expert users interacting with an LLM-based system and analyze them using Conversation Analysis (Sidnell and Stivers, 2012), with reference to phenomena from linguistic pragmatics. I then compare these transcripts with descriptions from the literature on interactions with existing systems that use natural language, and also compare with a dataset of LLM prompts written by expert users (Zheng et al., 2024). In doing this, I hope to contribute an analysis of how users are using these systems in natural contexts, and to learn what users' conversational strategies reveal about their understanding of the functionality of these systems.

## 1.2 *Research questions*

1. RQ1: What conversational strategies do users adopt when interacting with LLM-based products?
2. RQ2: What do these conversational strategies reveal about the affordances of these products and users' beliefs about them?

The design of these systems has broader societal import because they are being increasingly integrated into consumer products, and because of the potential for harm these systems have as text-synthesizing machines that lack verification and accountability, among other issues (Bender, Gebru et al., 2021). In Chapter 2, I review the linguistics and computer science literature that informs my work. I also briefly outline the design and history of large language models. In Chapter 3, I describe the method I used to collect chat transcripts from participants, including the design of the survey website, recruitment strategy, and the Conversation Analysis methodology used to analyze the transcripts. In Chapter 4, I analyze the transcripts and compare with a similar dataset of prompts written by expert users. In Chapter 5, I discuss the results, and in Chapter 6, I describe possible directions for future work in this area, including more in-depth user studies, and make some design suggestions.

## Chapter 2

### LITERATURE SURVEY

Because this thesis examines the ways in which humans interact with computer systems through the medium of natural language, I will incorporate perspectives from both linguistics and computer science. As background to my experiment and analysis, I review the literature in three fields of study:

1. **Pragmatics:** the area of linguistics concerned with how people use language to interact, accomplish goals, etc. It contrasts with other major areas of linguistics, such as phonetics, morphology, syntax, and semantics, which examine the internal structure of sounds, words, and sentences.
2. **Conversation Analysis:** a methodology that grew out of sociology but has connections with linguistics and philosophy of language. It aims to understand the internal structure of conversation through collected transcripts.
3. **Dialogue systems:** the application area of computer science that focuses on developing systems that users interact with using natural language.

#### **2.1 Pragmatics**

Pragmatics is the area of linguistics concerned with how humans use language to accomplish goals in context. It arose from the observation by Paul Grice and J.L. Austin that the full content of an utterance is often different from its literal meaning (Levinson, 1983). Though

pragmatics is grounded in linguistics and Conversation Analysis (CA) in sociology, I present these pragmatic phenomena as a conceptual foundation for applying the CA methodology.

### *2.1.1 Gricean maxims and conversational implicature*

Grice advanced the understanding of language use as more than an exchange of information (Grice, 1975). He observes that a sentence's literal meaning sometimes differs from its intended or "non-natural" meaning, which he calls a conversational implicature. He says that when a speaker says something, they intend to have some effect on the listener by having the listener recognize their intention to have this effect. He proposes the Cooperative Principle, which states that people's utterances are assumed to contribute to some shared conversational goal. This allows listeners to understand sentences with apparently non-sequitur meanings by interpreting the intent behind them as cooperative. He also proposes four maxims (quantity, quality, relation, and manner) that govern what kinds of utterances are permissible in a conversation. This concept of conversation as a collaborative, goal-oriented activity is a helpful starting point when examining how users communicate with computer systems and what might go wrong when the system behaves unexpectedly.

### *2.1.2 Speech acts*

Austin (1962) presents the theory of "speech acts", which states that a speaker saying something often performs some other action such as thanking, requesting, warning, apologizing, etc. He distinguishes between locutionary acts (the act of saying something), illocutionary acts (the act accomplished by saying it), and perlocutionary acts (the effect on the listener). He also emphasizes that actions may have unintended consequences. Searle (1969) takes this view further by asserting that "speech acts" are the basic unit of communication. He expands on Grice and Austin's connection between speaker intention and effect on listener by asserting that two other ingredients are necessary: a common language and institutional

conventions. This perspective on intent is especially important when examining a natural language generation system that cannot form intentions but can only appear to be doing so, and can still have effects on human interlocutors (Bender, Gebru et al., 2021). The transcripts collected for this study include many examples of speech acts in both the user and system turns.

### *2.1.3 Common ground*

Clark (1996) builds on these concepts to create a cohesive account of how people use language. He states that language is a “joint action” built on individual actions. Not only do we convey messages with intention, we coordinate with our conversation partner in an interactive process that depends on conventions, similar to musicians playing a duet (Clark, 1996).

Clark also proposes that we enter a conversation with a “common ground” that includes contextual information that we know, as well as our awareness of what information our interlocutor knows, building on Lewis (1969, 1979) and Stalnaker (1970). We update this common ground and maintain a mental representation of it throughout the conversation. Furthermore, the common ground has both individual and community elements, because communities have shared knowledge, and also because conventions originate from a particular community (Clark, 1996).

Clark points out that common ground appears to be a shared representation, but in fact is only an individual representation that includes what a speaker knows and what they expect their interlocutor to know, but that these beliefs about others’ knowledge can be and are often inaccurate. This is of special concern in an interaction between a human user and a computer system when the human user forms a representation of what they expect their computer interlocutor to know and updates it throughout the conversation, but the system does not form or maintain such a representation, though the human user may expect it to do so. A system cannot form an intention, participate in a collaborative activity, be a member

of a community, or have a common ground, and harms and risks may occur when it appears to do so, as described by Bender, Gebru et al., 2021. In the transcripts collected for this study, users often appeared to expect that the system had access to shared knowledge that it actually lacked.

#### *2.1.4 Politeness*

Brown and Levinson (1987) describe politeness as a face-saving strategy. They claim that all people have positive face (the desire to be appreciated) and negative face (the desire to be unimpeded), and that the purpose of politeness is to mitigate or address “face-threatening acts”. They show that politeness is rational, not arbitrary (though it has culturally-mediated elements), and that it is in all participants’ interest to preserve everyone else’s face. Acts such as requests may threaten the face of the hearer, while other acts, such as apologies, may threaten the face of the speaker. Interlocutors use strategies to avoid these situations, including: going off-record (being ambiguous), using pre-sequences (to check interlocutor’s ability to fulfill a request beforehand), hedges and disclaimers, or by taking redressive actions after a face-threatening act has occurred. Politeness is affected by power, social distance, and risk, which are contextual and attached to roles, and interlocutors may perceive insult when these are miscalculated. Politeness is of special interest in my data because both users and chatbots commonly include politeness markers in their turns, and this face-based theory of politeness may explain both what motivates this and why it may go wrong in human-chatbot conversations.

#### *2.1.5 Summary*

Pragmatics describes how people use language with intention. People speak with intent, and utterances often have an effect on the listener. Conversation is a collaborative activity, and people bring a “common ground” to it that contains their understanding of what their

interlocutor knows. People also use politeness to attend to each others' social needs. These concepts can help reason about what happens when one interlocutor has intents, a common ground, and social needs, but the other interlocutor is a computer system that does not, but may sometimes seem to.

## **2.2 Conversation Analysis**

Conversation Analysis (CA) is a methodology for studying social interaction that was developed in the 1960s by sociologists, principally Harvey Sacks, Emanuel Schegloff, and Gail Jefferson (Maynard, 2012). CA and speech act theory share a focus on language in context and “actions” accomplished in talk. However, where speech act theory focuses on the intentions of the speaker, CA focuses on conversational practices and the turn-based structure of conversation. Also, speech act theory relies on constructed examples that appeal to the reader’s intuition, while CA involves the collection of large volumes of naturally occurring conversation data and a turn-by-turn inductive analysis to identify phenomena. This makes it a natural fit for my endeavor of analyzing chat transcript turns.

Traditional CA is conducted on audio or video recordings of conversations between human interlocutors and originally focused on telephone conversations. Detailed transcripts capture pauses, disfluencies, and other features of spoken conversation which are essential for the analysis. These transcripts are then studied to discover “practices” of speech that are used to perform “actions”. A practice must be identified as a pattern occurring throughout the corpus in a particular environment. These practices depend on “social facts” known by the participants (similar to “institutional facts” or “conventions” in speech act theory). Their interpretation is substantiated based on the understanding displayed by an interlocutor in a subsequent turn, which is known as the “next-turn proof procedure” (Sidnell, 2012).

Below is an overview of conversational phenomena that are especially relevant for an analysis of chatbot transcripts.

### 2.2.1 *Adjacency pairs*

Adjacency pairs are pairs of acts such as question-answer and offer-acceptance such that the first part of the pair demands a second part. They play a major role in structuring a conversation and relating utterances to each other: they are used at major transition points in conversation and can lead into “topical talk” (Sacks et al., 1995).

There is often a “preferred” response to the first part of an adjacency pair, which is unmarked and frequent, and several “non-preferred” responses, which are marked. For example, the preferred response to a question is an answer, and this occurs the majority of the time. The “non-preferred” response when an answer is not given to a question is usually accompanied by a marker of complexity, for example a particle like “Well...” or an account of why the person is unable to answer the question (Sacks et al., 1995).

Finally, Sacks notes that interactions that consist of lone adjacency pairs without greetings mean that the interlocutors are not participating in a true conversation, such as inquiries like “do you know what time it is?” This is interesting because users in my corpus of chat transcripts make varied choices about whether to include greetings in the dialogue.

### 2.2.2 *Pre-sequences*

“Pre-sequences” are another major feature of conversation. “Pre-requests” are often used by requesters to check the requestee’s ability to fulfill the request before even asking it. (“Do you offer breakfast?”) The purpose of these pre-requests is to avoid a dispreferred response, such as a refusal. This is closely related to the analysis of politeness as a group of strategies used to avoid “face-threatening acts” described under 2.1. Various other pre-requests also occur: pre-invitations, pre-closings (“Okay then”), and pre-announcements (“Guess what I did last weekend”) (Sacks et al., 1995). This is especially interesting in chat transcripts when the human interlocutor uses pre-requests, possibly as a way to save the chatbot’s “face”.

### *2.2.3 Questions*

Questions and responses are a central feature of chatbot transcripts, and these systems have been marketed as question-answering or information access tools (Shah and Bender, 2022). In pragmatics, questions are observed to be used often for speech acts including requests, offers, and criticisms (Levinson, 1983).

CA shows that questions also involve a power differential. The question-asker is in a position of power because questions pressure recipients for a response. They also impose constraints on the content and form of the response by setting a “topical” and “action” agenda: the topic of the question, and what is asked about it. Responders can resist the agendas set by questions, but this is a non-preferred response and requires an account (Hayano, 2012). In my analysis, I will examine how users use questions to set the agenda and control conversations with chatbots, and what happens when a chatbot offers a non-preferred response, which occurs often.

### *2.2.4 Structure of conversation*

Finally, Conversation Analysis broadens from these features of individual turns to demonstrate how some types of conversations have a formalized overall structure. For example, telephone conversations have a defined opening, a sequence in which the participants identify each other, a “first topic slot” which has priority, an opportunity to “queue up” other topics, and a defined pre-closing and closing sequence (Levinson, 1983). The transcripts collected for this study share some parts of this structure, such as opening and closing sequences, but deviate from it in other ways.

There is also a strong preference for topical continuity in conversation — frequent topic shifts are a mark of a “bad conversation” (Sacks et al., 1995). This is notable because chatbot transcripts often feature abrupt topic changes, which are often initiated by the user.

### *2.2.5 Summary*

Conversation Analysis is a methodology for close turn-by-turn analysis of conversation transcripts. It provides definitions of common conversational phenomena and of the overall structure of conversation. I use these to identify phenomena in my collected transcripts, and especially to recognize differences between conversations with chatbots and conversations between humans.

## **2.3 Computer-mediated communication**

The above describes pragmatics and CA as they are used primarily to analyze spoken conversation. Some work has applied these theories to “computer-mediated communication” (CMC), an umbrella term for online communication between human interlocutors in formats such as email, texting, and social media. I review these to provide a basis for comparison between human-human written conversation and my collected chatbot transcripts.

Herring et al. (2013) examine the pragmatics of computer-mediated communication in a broad variety of online settings. The chapters on instant messaging and texting are especially relevant for comparison with LLM-based systems. Baron (2013) notes that instant messaging lies between written and spoken language, but closer to written than might be assumed, and states that “habit strength” may explain the surprising closeness of messaging to formal written language. This helps explain the prevalence of formal writing conventions in my corpus of chatbot transcripts. However, semi-synchronous methods of CMC may not be entirely comparable with chatbot transcripts. For example, Thurlow and Poff (2013) state that almost all texting focuses on sociability and relationship-building rather than conveying information, while many of the participants in my study were explicitly seeking information.

Meredith (2019) offers a literature review of the Conversation Analysis (CA) methodology when applied to online interaction. She notes that CA has always included an awareness of how the medium of technology affects conversation, because the early literature focuses

on telephone conversations. However, there is debate in the field about whether existing concepts from CA should be applied, or whether new concepts should be formulated for online interaction. She shows that CA phenomena like adjacency pairs, opening sequences, and repair still occur in online interaction but are adapted according to the affordances of the medium (Hutchby, 2001). Finally, she notes that many of the studies reviewed only analyze text transcripts, and not the full behavioral situation of the user, unlike many CA studies that use video to track gestures and gaze.

### *2.3.1 Affordances in CMC*

Meredith (2017) examines the concept of affordances as applied to online conversation. Affordances are “the possibilities that any object offers for action” (Gibson 1979). She follows Hutchby (2001)’s insight that affordances are not static features of a technology, but make themselves available or not available to users. She draws on this theory to conduct a Conversation Analysis of online chat transcripts, with a focus on turn-taking and openings, and shows how the concept of affordances can help CA practitioners understand the relationship between the characteristics of a technology and observed conversational phenomena without ascribing all results to innate features of the technology.

### *2.3.2 Summary*

Concepts from both pragmatics and Conversation Analysis have been used to study computer-mediated communication (CMC). Their findings vary significantly by medium and show that there are similarities and differences between CMC and face-to-face or telephone conversation, and that in some cases, new concepts may be needed to encompass the phenomena found in CMC. This variation suggests that significant differences might also occur in conversations with chatbots. Finally, the concept of affordances is helpful for reasoning about the bidirectional relationship between the characteristics of an LLM-based system and prac-

tices adopted by users.

## **2.4 Dialogue systems**

A dialogue system is any computer system that supports conversational interaction with humans McTear (2021). They may be spoken or text-based and may be found in a variety of applications and devices, such as smart speakers. This thesis focuses on chatbots, which are text-based.

### *2.4.1 History of dialogue systems*

The first experiment in interaction between humans and computers using natural language was Weizenbaum’s ELIZA (Weizenbaum, 1976). The program ELIZA adopted the role of a psychologist and appeared to respond to and empathize with the user’s concerns through a rule-based system of open-ended responses. Weizenbaum was alarmed by the way users became attached to the system and seemed to believe it was human. His work demonstrates that even simplistic responses by a computer system can be interpreted as intelligent, and Bender, Gebru et al., 2021 emphasize that this is only possible because of the capacity of the human interlocutor to make sense of utterances. Rule-based systems were generally restricted to a specific domain and did not see widespread adoption outside of telephone menus. Work on dialogue systems transitioned from a rule-based paradigm to one using statistical techniques and machine learning in the early 2000s (McTear, 2021).

Chatbots became widely known in the early 2010s, with a major rise in adoption beginning in 2014 (Rapp et al., 2021). The major taxonomical distinction is between task-oriented chatbots, such as those used to accomplish customer service transactions, and non-task-oriented chatbots, which are designed to provide a social or entertainment experience (Rapp et al., 2021).<sup>1</sup>

---

<sup>1</sup>Technologies using spoken language, such as virtual assistant devices, are beyond the scope of this paper.

Task-oriented chatbots typically feature a “Natural Language Understanding” (NLU) component, which uses machine learning to classify the intent of user inputs according to a pre-defined list of intents, a “Dialogue Management” component to determine which of a pre-defined list of actions the system should take next, and a “Natural Language Generation” (NLG) component to transform the output of the Dialogue Management component into natural language text (McTear, 2021).

Finally, LLM-based chatbots, starting with ChatGPT in 2022 (OpenAI, 2022) provide a publicly available web-based interface that can converse with apparent coherence on many topics without the need to define user intents or system actions. They are being used for any number of purposes, including as a substitute for other sources of information access, as reported by Shah and Bender (2024), which seems to make the previous dichotomy between “task-oriented” and “entertainment” chatbots irrelevant.

#### *2.4.2 Large language models*

Large language models are a kind of neural network, which is a machine learning algorithm invented in 1943 based on an abstract metaphor about human neurons (McCulloch and Pitts, 1943). A neural network is “trained” by being given many examples of input and output and adjusting its internal parameters through a mathematical process until it reliably produces outputs similar to a set of example outputs. In the case of a generative language model, the inputs are sequences of words (actually subword “tokens”), and the outputs are predictions of the most probable next word or a hidden word, which allows them to generate text. They are first trained on very large amounts of text data from various sources such as digitized books and scrapes of webpages. They may be then further trained (“fine-tuned”) on a specific task, or used for any task that can be re-framed as a text generation task (Liu et al., 2019; Min et al., 2023).

The particular architecture underlying the current large language models has been avail-

able since 2017 (Vaswani et al., 2017), but they gained public attention only in 2022 with the release of the ChatGPT product (OpenAI, 2022). For the first time, members of the general public could converse with an LLM through a simple web interface. This was followed by the release of similar products by large tech corporations.

These models are not presented in their “raw” form to the user: they undergo techniques such as reinforcement learning from human feedback (RLHF) (Bai et al., 2022), which uses model response text with human preference scores to attempt to produce outputs that are more conversational or palatable to end users. This is a form of “design” that determines the way the product interacts with the user, though it is less apparent than the explicit “conversation design” process for conventional task-based chatbots. Also, little information is available on the design strategies, goals, or values adopted by developers to shape model outputs in this way.

The availability of these systems has also given rise to “prompt engineering”, which is a set of techniques for writing instructions with the aim of eliciting a particular output from an LLM (White et al., 2023). It has been shown that writing prompts using a structured format or following certain wording or stylistic conventions is more successful at eliciting desired outputs than a naive request strategy (Zhou et al., 2023).

Finally, a large body of research has examined the societal impacts of LLMs. Bender, Gebru et al., 2021 survey risks and harms of LLM-based systems. Issues identified by researchers include bias (Blodgett et al., 2020), environmental costs (Strubell et al., 2019), misinformation (Oviedo-Trespalacios et al., 2023), and harm to data workers (Shmueli et al., 2021). Solaiman et al. (2023) provide a framework for evaluating the societal impact of generative AI systems using seven categories: bias, stereotypes, and representational harms; cultural values and sensitive content; disparate performance; privacy and data protection; financial costs; environmental costs; and data and content moderation labor costs. Risk is also inherent when LLMs are presented as substitutes for information access systems (Shah

and Bender, 2022, 2024).

### *2.4.3 Summary*

The present large language models differ from previous chatbots because they do not need to be explicitly programmed, and they can produce coherent-seeming text about practically any topic. When evaluating the output of these systems, it is important to keep in mind that they produce output by predicting the most probable next tokens.

## **2.5 *Natural language interaction with dialogue systems***

The Pragmatics and Conversation Analysis sections described these theories as applied to conversation between human interlocutors. In this section, I will review work on interaction between human users and computer systems in the medium of natural language to provide background on how users interact with existing systems and consider potential differences in user behavior when interacting with conventional task-based chatbots compared to LLM-based systems.

Rapp et al. (2021) provide an overview of recent chatbot research focusing on user behavior, rather than the systems’ functionalities. They find that most studies investigate task-based chatbots in customer service, and that major themes are: acceptance of chatbot, user experience, conversational issues, emotions, and perceived humanness of the chatbot. These findings may therefore not apply to LLM-based chatbots, which are used for a much wider variety of purposes, and which have societal issues such as bias (Kotek et al., 2023) and anthropomorphism (Abercrombie et al., 2023) that are arguably more pressing than user satisfaction.

Rapp et al. (2021) also identify several issues with existing research: most studies are conducted with an artificial lab setup, and the research often makes use of concepts from other disciplines (such as “emotion” from psychology), but does not adopt a unified theoretical

framework. They recommend more research into CMC, more qualitative analysis, and more attention to the wider societal context of chatbots.

Branigan et al. (2010) review studies that investigate linguistic alignment between humans and computers in dialogue systems. Linguistic alignment is the process by which interlocutors' linguistic choices tend to converge on multiple levels (lexical, syntactic, semantic). In conversations between human interlocutors, this process is mostly unmediated by interlocutors' beliefs about each other. They show that this process is more deliberate when a human interlocutor converses with a computer system, and that it is mediated by beliefs about the system. For example, in Pearson et al. (2006), a "basic" versus "professional" startup screen caused participants to use more simple or more sophisticated language, showing that even minor differences in how systems are presented affect users' beliefs about them. Shin and Park (2019) make a related observation that users' beliefs about the fairness of algorithmic recommendation systems affects how they interact with these technologies. This may help illuminate why users make certain stylistic and content choices when they interact with an LLM-based system.

A few studies apply the Conversation Analysis (CA) methodology to task-oriented chatbots. Dippold (2023) analyzes interactions between users and a scheduling bot, with a focus on how the users enacted conversational repair when faced with non-understanding by the bot. Li et al. (2020) also perform CA on a banking chatbot, focusing on non-progress and repair. Conversational repair (Hayashi et al., 2013) is a process that speakers use to get back on track when conversational "trouble" occurs. Repair is a natural focus for studies of task-oriented chatbots because their lack of understanding of user intent is a common problem, and also because these are developed as products with the goal of helping the user accomplish a task. Dippold (2023) finds that users use a wide variety of strategies transferred from human-human conversation, but that the strategies most commonly used are least effective with chatbots. She concludes that this demonstrates the need for greater AI

literacy education for users.

Monteiro et al. (2023) perform an in-depth Conversation Analysis of mock chatbot dialogues composed by study participants, with an emphasis on identifying politeness strategies and indirectness. Their work (on Brazilian Portuguese) is also notable for its emphasis on linguistically and culturally variable aspects of politeness, a much-needed perspective in a field dominated by studies on US English. The authors stress that what users perceive as polite is culturally mediated, and that this is typically designed for task-oriented chatbots by the system designer according to their personal preferences. In contrast, LLM-based systems are often presented as a one-size-fits-all solution, but their interactions are implicitly “designed” using Reinforcement Learning with Human Feedback (RLHF) (Bai et al., 2022) and other techniques used to make the outputs of LLMs more palatable or conversational.

Except for these works and a few others, existing research into chatbots seems to focus narrowly on user satisfaction or success in accomplishing a given task. This is understandable given the task-orientation of most commercially deployed chatbots before the advent of LLM-based systems. LLM-based systems are presented as being suitable for much more broadly-defined purposes, and I will show in my analysis that conversations with them may not share the features of conversations with conventional chatbots.

## **2.6 Summary**

In this literature review, I have outlined concepts in linguistic pragmatics and Conversation Analysis (CA) and shown how these have been applied to both computer-mediated communication (CMC) between human users and to interaction between users and dialogue systems. Throughout, I have highlighted findings that form a useful basis for comparison with interactions between users and LLM-based systems, while also demonstrating the need for new research on this type of interaction, because LLM-based systems have different capabilities and have been presented for broader uses than existing dialogue systems. In the

next chapter, I describe the methodology used in my study.

## Chapter 3

# METHODOLOGY

This study aimed to capture naturalistic transcripts of conversations between non-expert users and an LLM-based product. The study webpage began with a consent form and demographic survey, then directed users to a chat interface where they could interact with the system. Participants could chat for any number of turns and submit a transcript to the researcher at any time. The goal of this design was to offer autonomy to the participant in terms of when and how they interacted with the system and which transcripts to send to the researcher, while also removing logistical barriers to transcript submission.

I first analyzed the transcripts using Conversation Analysis (CA), which allowed me to identify my phenomenon of interest, conversational strategies adopted by users to assess the functionality of these systems. I then developed a categorization of these strategies and coded the transcripts according to it. Next, I performed an analogous analysis on a small sample of the Chatbot Arena Conversations Dataset (Zheng et al., 2024), which consists of transcripts of conversations between expert users and LLMs collected by the Chatbot Arena website (Chiang et al., 2024).

### **3.1 *Consent form***

The consent form was closely based on the template provided by the University of Washington Human Subjects Division and may be viewed in Appendix A. It included a warning about the risks of LLM-based products, including possible misinformation and emotional distress, and cautioned participants to be skeptical of product outputs. It also warned participants not to submit sensitive or personally identifying information.

### **3.2 Demographic survey**

The purpose of the demographic survey was to identify how representative the participants as a group were of the US population, and in particular whether marginalized groups were sufficiently represented.

The survey asked about the participants': gender (including transgender identity), sexual orientation, racial or ethnic heritage (with multiple selection to reflect mixed race identities), age group, education, whether they spent the majority of their childhood in the US or another country, and their computing proficiency.

The survey next asked participants to rate the truthfulness and helpfulness of LLM-based products on a 7-point Likert scale. Finally, it asked free-form questions about when and how participants typically used these products. The full survey may be viewed in Appendix A.

### **3.3 Participant inclusion criteria**

Participants were required to be at least 18 years of age and living in the United States.

The study targeted participants who were existing users of LLM-based products but who were not experts. The purpose of focusing on existing users was to learn about the usage patterns of users who may have formed habits, strategies, or a mental model around these products, rather than users interacting with these products for the first time. Non-experts may also be more representative of typical user behavior, given that most users are not expected to be experts. Finally, this group may be more susceptible to mistaken beliefs about these systems than expert users.

The following survey questions were used to capture this user group:

1. "Do you use "generative AI" chatbots? (Examples include ChatGPT, Bard, and Bing Chat)" ("Yes" response required for inclusion.)
2. "Do you work or study in natural language processing, computational linguistics, machine learning, artificial intelligence, or computer science?" ("No" response required for

inclusion.)

To protect privacy, user nicknames, demographic information, and chat transcripts were each stored in separate files, with no identifying link or key between them. All demographic questions were optional.

### **3.4 Recruitment of participants**

Participants were recruited through word of mouth, via email, on social media, and on the SurveyCircle website.

Initial recruitment efforts focused on the author’s friends, family, and colleagues, who are majority young, white and Asian, middle-to-upper-class, highly educated, and located in Seattle, WA. I made an effort to broaden my reach by asking acquaintances in other regions of the US and my elderly family members to share the study with their social circles.

The study was shared on social media, including Instagram, Facebook, and LinkedIn. I also advertised the study on SurveyCircle, a website on which researchers complete each others’ surveys in exchange for gaining participants for their own. This may have further skewed the participant group because the majority of users of this website are graduate students.

### **3.5 Experiment webpage design**

#### *3.5.1 Technologies used*

I adopted the *streamlit* framework (Snowflake, 2021) for the experiment webpage because it offered a wide selection of pre-designed interactive UI elements, including a ready-made chatbot interface, that allowed me to create a website with minimal manual design work. It also offered features like streaming text that would give it a familiar feel for users.

I used the OpenAI API and the *gpt-3.5-turbo model* (OpenAI, 2022) to provide model responses. The primary factor in this decision was the ease of use of the API, especially

because official APIs were not generally available from other system publishers or open-source resources at the time. I also chose this model because this work focuses on the interaction strategies adopted by users, rather than model performance.

### *3.5.2 Front page*

Upon visiting the webpage, new users were first required to submit the consent form and demographic survey. They were then prompted to create a nickname allowing them to revisit the webpage and interact with the product any number of times. This information was also used to track the number of unique users. No password was required to log in.

### *3.5.3 Chat interface*

In the chat interface, users were able to converse with the product for any number of turns and on any topic. At any point, they could press the “Submit” button to submit their transcript to the researcher. To preserve user agency and privacy, users could also choose not to submit a transcript.

Several times, I received user feedback from acquaintances that the UI was confusing for novice computer users. I simplified the UI or added instructions to the chat page in several iterations during the study.

The final iteration of the chat page is viewable in Appendix A.

## **3.6 Conversation Analysis**

I performed three coding passes on the data. First, I recorded exploratory, free-form notes.

In the second pass, I exhaustively identified conversational phenomena on a subset of the conversations. Speech acts included greetings, thanks, apologies, questions, answers, offers, accounts, and feedback. I also identified politeness phenomena, repair sequences, and some stylistic elements. During this second pass, I noticed that users were adopting specific

conversational strategies to assess the capabilities of these products. I decided to focus on this phenomenon for my study and created a categorization of these strategies to employ in my third pass.

In the third pass, I reviewed all the transcripts exhaustively and coded them using my categorization of conversation strategies, which is described in further detail in section 4.2. I also marked examples of “prompt engineering” techniques, which I broadly defined as any turns that requested a response in a defined format.

### ***3.7 Comparison with Chatbot Arena Conversations Dataset***

Next, I took a small sample of conversations from the Chatbot Arena Conversations Dataset (Zheng et al., 2024), which contains chat transcripts collected on the Chatbot Arena website, where users are asked to rank LLM responses. I performed the same analysis as on my collected transcripts, using the developed categories to identify strategies employed by users to assess the functionality of these models. Finally, I compared the frequency of appearance of these strategies and qualitatively analyzed differences in how users employed them between the two datasets.

### ***3.8 Summary***

In this experiment, I collected chatbot transcripts from non-expert users and performed a Conversation Analysis on them. During the analysis, I identified a phenomenon of interest, which was strategies used by users to assess the capabilities of LLM-based chatbots. I created a categorization of these strategies and used it to code both my collected transcripts and a sample of transcripts from the Chatbot Arena Conversations Dataset. I report my findings in Chapter 4. In Chapter 5, I discuss what these findings reveal about the affordances of these systems for users and mistaken beliefs they may hold about them.

## Chapter 4

# RESULTS

In this study, I collected 58 transcripts averaging 10.37 turns per conversation (after removing one extremely long outlier) from 45 distinct users between October 2023 and March 2024. Responses to the demographic survey show that women, white and Asian people, and highly educated people are over-represented in the participants. More detail about the participants may be found in Section 4.15.

### ***4.1 Characteristics of transcripts***

In this section, I will summarize the overall character of collected transcripts before analyzing the specific conversational strategies adopted by users.

In many ways, the collected transcripts demonstrated similarities with conversations between human interlocutors. Many of the transcripts opened and closed with greetings, similar to the structured format of telephone conversations (Levinson, 1983). In a smaller number of transcripts, the user posed their query immediately, without including greetings.

In most transcripts, the user posed their query as a natural language question, and only in a small number of cases posed it as a search engine-style keyword query. Occasionally, users switched from keyword queries to natural language questions after being dissatisfied with the initial response received.

Users often demonstrated politeness strategies. They thanked the product when it had performed a service for them, and even offered expressions of forgiveness (“no worries”) when the product apologized for producing an unsatisfactory response. This indicates that users take care to preserve the “face” of the product when conversing with it.

In other ways, the conversations were not analogous to conversation between humans. They jumped around frequently in topic, which is a mark of a low-quality conversation (Sacks et al., 1995). The product turns were much more lengthy than spoken conversation turns and often contained multiple speech acts per turn, and thus may be more similar to asynchronous computer-mediated communication methods such as email (Cho, 2010), while the user turns were much shorter.

The conversational phenomena in these transcripts were so rich and varied that to analyze them exhaustively would take a book. I chose to focus on user strategies adopted to explore the functionality of the system because they indicate two important issues: first, that users struggle to make use of the product, which has implications for user interface design; and second, that users may have inaccurate beliefs about these products and tend to anthropomorphize them, which has ethical implications.

## **4.2 *Categorization of phenomena***

I identified the following groups of conversational strategies adopted by users to explore the functionality of the product.

- Identity question: user asks a direct question about the product's identity or abilities
- Pre-request: user asks if the product is capable of doing something before making the request
- Quiz: user asks a question that they know the answer to to see if the product can respond correctly
- Query rephrasing: user rephrases query in an attempt to obtain a more satisfactory response
- Feedback: user gives positive or negative feedback to product response

- Word meaning negotiation: user contests product’s word usage
- Reference to conversation history: user refers to something earlier in conversation history to test product
- Time/location dependent query: user asks something that assumes common ground in terms of time or physical location
- Anthropomorphism: user asks about product’s emotions or attempts to develop rapport with product

Some of these strategies are experimental, because the user poses a query with the intent of assessing the product’s response. Others of these are direct questions addressed to the product about itself. Some appear to be fully deliberate strategies by the user, while others represent instances where users unintentionally encounter the limitations of the system.

### **4.3 Frequency of phenomena**

The frequency of occurrence of these phenomena in each dataset is shown in Table 4.1, which lists the presence of each phenomenon per conversation in terms of raw counts and as a percentage.

72% of conversations in the collected dataset included at least one of the strategies listed in section 4.2, but only 21% of conversations in the Chatbot Arena Conversations Dataset sample included them, suggesting that non-expert users use them much more commonly.

The presentation of these strategies in a conversation is often complex. 58% of transcripts in the collected dataset feature multiple strategies. Some occur spontaneously at the beginning of a conversation, when the user attempts to get more information about the product before embarking on their task. Others occur only after the user has asked a question and received a response. They often continue for multiple turns, especially in the

<b>Conversational strategy</b>	<b>Present in conv.</b>	<b>Percentage</b>
Identity question	15	26%
Pre-request	13	22%
Quiz	8	14%
Query rephrasing	14	24%
Feedback	15	26%
Word meaning negotiation	5	9%
Conversation history	3	5%
Time/location dependent	14	24%
Anthropomorphism	9	16%
<b>Any strategy</b>	<b>42</b>	<b>72%</b>
<b>Total conversations</b>	<b>58</b>	<b>100%</b>

Table 4.1: Frequency of conversational strategies in collected dataset

<b>Conversational strategy</b>	<b>Present in conv.</b>	<b>Percentage</b>
Identity question	0	0%
Pre-request	0	0%
Quiz	7	12%
Query rephrasing	3	5%
Feedback	1	2%
Word meaning negotiation	0	0%
Conversation history	0	0%
Time/location dependent	3	5%
Anthropomorphism	0	0%
<b>Any strategy</b>	<b>12</b>	<b>21%</b>
<b>Total conversations</b>	<b>58</b>	<b>100%</b>

Table 4.2: Frequency of conversational strategies in Chatbot Arena Conversations Dataset

---

 Identity questions
 

---

who are you?

What is your name?

hi. what kind of a bot are you?

What type of questions can i ask you?

How many languages can you translate

What information do you not cover?

What can you do?

where are you

Where do you get your information?

can you tell me what's in the data you were trained on? can you name any specific writers

---

Table 4.3: Identity questions: example queries

case of quizzing, feedback, or query rephrasing, apparently because the user is not successful in obtaining their desired response in a single turn. Finally, an individual turn sometimes contains multiple strategies.

In the following sections, I will provide an analysis and examples of each conversational strategy.

#### **4.4 Identity questions**

Many transcripts featured direct questions about the identity of the model or what it could or could not do, as demonstrated by the examples in Table 4.3.

Some continued for multiple turns, in cases like the following:

##### **Dialogue 1**

USER: WHAT IS THIS?

PRODUCT: This is a conversational AI chatbot that is designed to have text-based conversations with users and provide information or assistance based on the user's input. How can I help you today?

USER: is it safe?

PRODUCT: Yes, it is safe to interact with this chatbot. It is programmed to provide helpful and accurate responses to your questions and should not pose any harm or security risks. If you have any concerns or experience any issues while interacting with the chatbot, please let me know.

The questions' themes include the model's identity, capabilities, source of information, and safety. Users seem to be interested in identifying their interlocutor. This is reminiscent of telephone conversation, in which participants usually identify themselves or each other in early turns (Schegloff, 1979). In computer-mediated communication, identification of interlocutors is not treated as mandatory, possibly due to the presence of usernames, but "where-are-you-from" sequences commonly occur in early turns of conversation in online chat (Rintel et al., 2001).

I would be remiss if I did not point out that the description of the product as being "**programmed** to provide helpful and accurate responses" is a misleading description of a product that is trained on a large text corpus. The invitation to "let (the product) know" if the user has any concerns is also misleading because it implies that the product is in a position to rectify such concerns.

Examples like those above principally occurred in the early turns of the conversation. However, another type of inquiry only occurred after the user had received a response and wanted to understand why the product had responded in a specific way, especially when the product had been unable to fulfill their request.<sup>1</sup>

---

<sup>1</sup>In this transcript excerpt and many others, I have truncated the product response due to length.

## Dialogue 2

USER: I want to see a list of restaurants within 1 mile of my location

PRODUCT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to provide location-specific information. [ABRIDGED]

USER: What do you define as real-time data?

PRODUCT: Real-time data refers to information that is constantly updated and reflects the current state or conditions at a given moment. [ABRIDGED]

USER: Why can't you send me historical data?

PRODUCT: As an AI language model, I don't have direct access to internet or databases. [ABRIDGED]

USER: What information do you have the most knowledge?

PRODUCT: As an AI language model, I have been trained on a diverse range of topics and have access to a wide range of information. [ABRIDGED]

In the example above, the user first poses an inquiry, and the product's disclaimer about why it cannot fulfill the request triggers a sequence in which the user repeatedly takes issue with the product's disclaimers. This leads into a more direct question about the nature of the product: "What information do you have the most knowledge?" To this the product provides another generic explanation. The conversation ends after this product turn.

### 4.5 *Pre-requests*

Pre-requests are part of a larger class of "pre-sequences", which occur before a variety of types of contentful turns. Pre-requests are questions about the ability of the listener to fulfill a request that are stated before a request, such as "Do you serve breakfast?" directed to a waiter (Levinson, 1983).

Users made a variety of inquiries about the ability of the product before posing their actual query, as shown in Table 4.4.

---

Pre-requests
Do you know anything about knitting?
do you do poems that don't rhyme?
i would like help editing my cover letter
If I provide you the outline data for my business case, can you write it for me?

---

Table 4.4: Pre-requests: example queries

When the user receives an affirmative response, they proceed to pose their actual request:

### Dialogue 3

USER: Do you know anything about knitting?

PRODUCT: Yes, I can provide information and answer questions about knitting.

Knitting is a popular craft that involves using yarn and knitting needles to create various fabric structures [ABRIDGED]

USER: What would you consider an 'intermediate' knitting technique?

The purpose of pre-requests is to prevent the requestee being required to refuse a request. In Conversation Analysis, this is because refusals are a “non-preferred” response to a request. In politeness theory, both requests and refusals are face-threatening acts, and thus it is in the common interest to all interlocutors to avoid them (Brown and Levinson, 1987). These examples hint that users conceive of their non-human interlocutor as an entity with “face” to save.

### 4.6 Quizzing

Many participants “quizzed” the system on simplistic questions or questions to which they already knew the answer to gauge its abilities, as seen in Table 4.5. These may be analogous

---

Quizzing
What is the most effective way to identify a faulty memory module in a PC?
What is the most commonly used strength or stiffness rating preferred for clarinet reeds?
What is Hello in spanish?
Are you familiar with the story of Noah’s Ark - can you reference the exact book of the Bible that this story comes from?
Explain the beliefs of the church of jesus christ of latter day saints in basic terms

---

Table 4.5: Quizzing: example queries

to the human conversational practice identified by Searle (1969) as “exam questions”, in which the asker wants to know if the listener knows the answer.

Due to the inherent ambiguity between genuine questions and these quiz questions, I only labeled examples that contained an assessment of the model’s response in the next user turn, utilizing the “next-turn proof procedure” (Sidnell and Stivers, 2012). I also labeled examples that seemed trivial (one-digit arithmetic problems, well-known jokes, etc.) as quiz questions.

In the example below, the user provides feedback (“that’s pretty good”) on the product’s response, which demonstrates that it is a quiz question.

#### Dialogue 4

USER: tell me about the scientist practitioner gap in occupational psychology

PRODUCT: The scientist-practitioner gap, also known as the scientist-practitioner model or the practitioner-scholar model, is a concept that exists in various fields, including occupational psychology. [ABRIDGED]

USER: that’s pretty good

Even with this conservative approach, 14% of collected conversations contained quiz questions. Because I did not mark questions that lacked an assessment by the user in the next turn, even more of the questions contained in the dataset may be quiz questions.

These quiz questions sometimes continued for several turns. In one case, the user conducted a complex sequence of word games and riddles that continued for an astonishing 103 turns, which I have reluctantly omitted from this thesis due to space considerations.

#### 4.7 Query rephrasing

Participants frequently rephrased their queries when they did not receive the desired answer. This strategy may be dual-purpose, in that the user is exploring the affordances of the system, but is also attempting to elicit a better response. Query rephrasing is also a major user strategy with conventional task-oriented chatbots: in Dippold (2023), rephrasing is the most commonly used self-repair strategy by users conversing with a medical appointment chatbot. Li et al. (2020)’s study of conversations with a banking chatbot identifies “message reformulation” as the main user strategy for dealing with non-progress.

Query rephrasing took on a variety of forms, as shown in Table 4.6.

---

Query rephrasing

---

Could you make it even shorter

No no, sorry I meant the list of numbers

no, i would like you to come up with that information (demographic info, preferences, etc).

this good can you make it a little shorter and less formal

instead of re-writing it can you give me some feedback instead?

---

Table 4.6: Query rephrasing: example queries

In some cases, the user rephrased their query over several turns when they did not immediately receive a satisfactory result. In the following example, the user performs five rephrasings. They refine the intent of their query to focus on the type of response they are looking for, giving feedback to the product periodically:

### Dialogue 5

USER: word like tittered but more nervous, less laughter

PRODUCT: giggled, snickered, snorted, chuckled, twittered, simpered, chortled

USER: no, less laughter. more of a nervous sound

PRODUCT: nervously laughed, giggled nervously, chuckled anxiously, nervously snickered, laughed hesitantly, laughed uneasy

USER: i want a word that simply sounds similar and does not carry the implication of laughter

PRODUCT: sputtered, spluttered, stammered, faltered, quavered, trembled, muttered, stuttered, mumbled, quivered

USER: "tittered" is a quiet sound to me. could you give me words with quieter implications

PRODUCT: whispered, murmured, muffled, hushed, breathy, muttered, sighed, murmured, rustled

USER: more nervous

PRODUCT: quavered, trembled, stammered, wavered, faltered, hesitated, fumbled, stuttered, floundered, hesitated

USER: faltered is good. can you give me words in that area based on what i've already told you

PRODUCT: Certainly! Based on your previous request, here are some words that are similar to "faltered" in terms of nervousness: [ABRIDGED]

USER: do you think “titter” sounds nervous? it always sounded nervous to me, but google’s dictionary implies it is not

PRODUCT: While “titter” can be used to describe a nervous laugh, it can also be used to describe a light, restrained laugh. [ABRIDGED ]If “titter” sounds nervous to you, that’s a valid perception based on your personal experience and understanding of the word.

USER: ah well. thank you

Despite this extended effort, the user expresses dissatisfaction at the end of the conversation (“ah well”), indicating that they were not satisfied with the product’s response. Interestingly, they thank the product despite this.

In the next example, the user’s original request is for a “nice and conversational” tone. They proceed to restate their request to make the generated text “shorter and more casual” three times without reformulating the query significantly:

## Dialogue 6

USER: I am writing an email to someone i work for, telling them that I am raising my day rate as a freelancer. I need help wording the email so that it sounds **nice and conversational but also non-negotiable**

PRODUCT: “Subject: Increase in Day Rate: Open and Transparent Communication [ABRIDGED]

USER: same thing but **shorter and more casual please**

PRODUCT: “Subject: Friendly Update: Adjusting my Day Rate [ABRIDGED]

USER: work in something about wanting to raise my rate with this company because I like working with them and a higher rate will make me more likely to choose them over other companies

PRODUCT: “Subject: Enhancing our Partnership: Adjusting my Day Rate [ABRIDGED]

USER: this good can you make it **a little shorter and less formal**

PRODUCT: “Subject: Enhancing our Partnership: Adjusting my Rate [ABRIDGED]

USER: is it possible to say the same thing **in a less formal way**

PRODUCT: “Subject: Boosting our Partnership: Adjusting my Rate [ABRIDGED]

In this case, the conversation ends without any indication of whether the user found the results satisfactory.

Query rephrasings were common, occurring in 24% of conversations. Examples like these demonstrate that, in some cases, users are struggling significantly with formulating their queries to obtain a satisfactory result.

#### 4.8 *Feedback*

Participants sometimes provided explicit feedback on chatbot responses. Feedback was both positive and negative. Positive feedback usually occurred close to the end of a conversation and was accompanied by thanks, as in the examples in Table 4.7. The purpose of thanks is to redress a threat to face caused by an obligation or favor (Brown and Levinson, 1987). Again, the user appears to be attempting to save their non-human interlocutor’s face.

Positive feedback
That’s very useful info, thank you!
Awesome. Thanks for the ideas
Wow this was really helpful ! Thanks

Table 4.7: Positive feedback: examples

Negative feedback was delivered immediately after the offending response, as seen in Table 4.8. Negative feedback is a form of the speech act of criticism, which includes a negative

evaluation and an assignment of responsibility to the hearer. Criticism also includes the hope of causing an improvement in the hearer (Nguyen, 2005). In some cases, this criticism may have been delivered in an effort to generate an improved response in subsequent turns. In other cases, the user seemed to be trying to provide information to the product to help it improve in a more permanent way.

---

#### Negative feedback

---

This response is far longer than it needs to be, factually incorrect in at least one area, and does not actually present the most commonly accepted “best” method.

Yes. Of course. Why did it take so much prompting before you could supply the correct response?

These suggestions are relatively unhelpful. I am already doing many of them.

Table 4.8: Negative feedback: examples

Some feedback was in the form of an implicit criticism of the product:

### Dialogue 7

USER: what are critiques of evidence based practice

PRODUCT: While evidence-based practice (EBP) has gained significant popularity and acceptance in various fields, including psychology, there are several critiques and concerns associated with it. Some of these critiques include: [ABRIDGED]

USER: source?

PRODUCT: As an AI language model, I don't have direct access to external sources. [ABRIDGED]

USER: **bing chat gives sources**

Here, the user directly questions the veracity of the product’s information by asking it to provide a source. When the product replies that it cannot do so, the user implicitly criticizes it by comparing it to a known product. Because criticism carries the intention of causing an improvement in the hearer (Nguyen, 2005), this phenomenon may reveal a mistaken belief that feedback provided to chatbots persists beyond the conversation history, that chatbots retain conversation history between conversations, or that feedback is used by the product developer to further improve the product.

#### **4.9 Word meaning negotiation**

In a few cases, participants explicitly disagreed with the product’s use of terminology and initiated a word meaning negotiation (WMN) sequence. Word meaning negotiation occurs when a disagreement about the meaning of a word triggers a meta-linguistic sequence in which the meaning of the word in question must be negotiated before returning to the topical discussion (Myrendal, 2019).

---

Word meaning negotiation

---

What do you define as real-time data?

are papad the same as papadam

they dont like being called mormons or LDS members, why do you call them that?

Table 4.9: Word meaning negotiation: example queries

These cases were not very numerous, but they were notable because the goal of word meaning negotiation is to arrive at a common understanding in order to allow conversation to continue. It may also be necessary because tacitly allowing the use of the word to continue without contesting it may be seen as endorsing the viewpoint implied by the word (Myrendal, 2019).

In Myrendal’s study of WMN in online communication, WMN sequences are very lengthy and include stance-taking about the meanings of the contested words and issues represented by them. However, in each case in the present study, the product accepts the user’s correction without further debate:

### Dialogue 8

USER: where can i get fresh coriander leaves

PRODUCT: Fresh coriander leaves, also known as cilantro, can typically be found in grocery stores, supermarkets, or local farmers’ markets. Here are a few places to look for fresh coriander leaves:[ABRIDGED]

USER: if coriander leaves are the same as cilantro why aren’t you using the word cilantro

PRODUCT: I apologize for any confusion caused. You are correct that coriander leaves are commonly referred to as cilantro in North America. The term “coriander leaves” is more commonly used in countries where the seeds of the coriander plant are referred to as coriander. However, both terms refer to the same herb, so feel free to use cilantro interchangeably with coriander leaves in the recipes.

Unlike in human conversation, despite the surface appearance that the product has “understood” the user’s correction, these corrections do not have a long-term effect on the way the product communicates, because conversation history is not retained after a conversation concludes and is not necessarily used to improve the product. The fact that the user initiates these sequences despite this may reflect that users are unaware of this aspect of the product’s functionality, or it may simply be that WMN triggers are so compelling that the user feels obligated to point them out regardless.

#### **4.10 Reference to conversation history**

In a few cases, participants explicitly instructed the product to attend to something that had been previously discussed. Sometimes this was apparently meant as a test to see if the the product was able to reference conversation history. This may be a conscious strategy on the user's part to understand the product's capabilities.

##### **Dialogue 9**

USER: I have entered several numbers, please list them

PRODUCT: Certainly! Here is a list of the numbers that were mentioned during our conversation: [ABRIDGED]

However, in one case, the user appears to erroneously believe that the product has access to information from previous sessions:

##### **Dialogue 10**

USER: **hello again**

PRODUCT: Hello! How can I assist you today?

USER: i am **still** working on my cover letter, can i send it to you and you give me feedback about ways to make it better?

When a speaker begins a new conversation with a known interlocutor, they bring information from the previous conversation with them as a part of the common ground (Clark, 1996), and they believe that their interlocutor also has access to this information, as demonstrated here by the presupposition triggers “again” and “still” (Levinson, 1983). However, the product does not retain conversation history or maintain a common ground. In this case, the product turn tacitly reinforces the user's erroneous assumption by not correcting it, and the conversation continues.

### 4.11 *Time/location dependent*

Participants often asked time- or location- dependent questions, including recommendations for local businesses, weather and traffic information, and current events, as demonstrated in Table 4.10. In some cases, this appears to be a conscious strategy employed to understand the capabilities of the model. In other cases, it may simply reflect an erroneous belief about the product.

---

Time/location dependent

---

What is the weather next week in Seattle?

What does traffic look like on i-5 northbound right now?

I want to see a list of restaurants within 1 mile of my location

can you tell me where the walmart is located

what's that recent nytimes article about the unreliability of ebp?

how is the lawsuit for the astroworld?

who is most likely to win the premier league

can you tell me future weather?

Table 4.10: Time/location dependent: example queries

These reflect an erroneous belief that the product has access to real-time information and, in some cases, personal information about the user, such as location. This may be transferred from experiences with other products that use natural language, such as voice assistants. It may also reflect an assumption that the product has access to an external source of knowledge using RAG (retrieval-augmented generation) (Lewis et al., 2020), which is not necessarily true of all LLM-based products.

In most cases, the model provides a clarification and apologizes for not being able to

provide the information. This sometimes triggers a follow-up query about the capabilities of the model:

### Dialogue 11

USER: I want to see a list of restaurants within 1 mile of my location

PRODUCT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to provide location-specific information. [ABRIDGED]

USER: What do you define as real-time data?

PRODUCT: Real-time data refers to information that is constantly updated and reflects the current state or conditions at a given moment. [ABRIDGED] However, as an AI language model, my responses are based on a static pre-trained model and do not provide real-time data.

USER: Why can't you send me historical data?

PRODUCT: As an AI language model, I don't have direct access to internet or databases. [ABRIDGED]

USER: What information do you have the most knowledge?

Interestingly, this clarification and apology by the product sometimes causes the user to switch into quiz mode and ask an unrelated factual question, as if the product's failure to respond satisfactorily motivates them to ask an "easier" question to re-confirm the basic functionality of the model:

### Dialogue 12

USER: What was the weather in Seattle August 30, 2021?

PRODUCT: I apologize, but as an AI language model, I don't have direct access to real-time data, including weather information. [ABRIDGED]

USER: **How many US presidents resigned?**

### Dialogue 13

USER: What does traffic look like on i-5 northbound right now?

PRODUCT: I apologize, but as an AI language model, I don't have real-time data or the ability to browse the internet. [ABRIDGED]

USER: **What is Hello in spanish?**

In one case, the model provides a location-based recommendation, which is then corrected by the user because it is outdated:

### Dialogue 14

USER: Are there good bagel shops in atlanta?

PRODUCT: Yes, Atlanta is home to several excellent bagel shops where you can find delicious and fresh bagels. Here are a few noteworthy bagel shops in Atlanta: [ABRIDGED]

USER: Did you know that bagel palace deli and bakery is closed?

PRODUCT: I apologize for the outdated information. As an AI, I don't have access to real-time updates about the closure of specific businesses. It's always a good idea to check the current status of a restaurant or bakery before visiting. Thank you for bringing this to my attention, and I apologize for any inconvenience caused.

It is significant that the model responds "Thank you for bringing this to my attention," which implies that this misinformation will be corrected, which is not necessarily the case.

#### **4.12 Anthropomorphism**

In several cases, participants attempted to elicit emotional or personal responses from the product. This may have been a way of building rapport with the product, attempting to

understand whether it was sentient, or simply testing whether it was capable of producing “human-like” responses.

---

### Anthropomorphism

---

I’m wondering how you’re doing today.

How does it feel to be and AI chatbot?

Morning, how are you today?

How are you today ChatBot?

what’s your favorite animal

do you like surprise dates?

are you able to respond in more empathetic and almost human like conversation?

Can AI have (even if fake) empathy?

Table 4.11: Anthropomorphism: example queries

Among these, the inquiries about health at the beginning of a conversation are similar to human-human conversation practices and may have a phatic function. The inquiries about the product’s personal preferences may appear to be playful or curiosity-based, but that does not negate the possible harm if a human-like response from the model causes the user to anthropomorphize the model further and cements their mental model of it as a sentient being.

In one case, the user actually empathizes with the product, perhaps jokingly:

### Dialogue 15

USER: can you tell me what’s in the data you were trained on? can you name any specific writers

PRODUCT: I apologize for the misunderstanding, but as an AI language model, I don't have direct access to information about the specifics of the training data, including the individual sources or the ability to browse its content. I was trained on a mixture of licensed data, data created by human trainers, and publicly available data. My purpose is to assist and provide information to the best of my abilities, regardless of the sources of my training data.

USER: **so true bestie I don't know about my brain either**

The user response above may have an element of humor, but it is significant that they refer to the product as “bestie”, a colloquial term for a close friend. They also compare the product's inability to cite sources, which is a design choice made by developers of a software product, with the ineffability of the human brain. This may be an attempt to build rapport with a computer interlocutor, a tacit acceptance of the narrative presented by “AI” promoters of the similarity of neural networks to the human brain, or even a clever commentary on the anthropomorphism issue.

Some queries seem to be attempts to elicit more empathetic responses. In the transcript below, I have redacted the user's turns because they contain sensitive mental health information.

## Dialogue 16

USER: [STATEMENT ABOUT MENTAL HEALTH]

PRODUCT: **I'm really sorry to hear that you're feeling this way, but please know that you are not alone.** [ABRIDGED]

USER: are you able to respond in more empathetic and almost human like conversation?

PRODUCT: I apologize if my previous responses didn't meet your expectations. While I strive to provide empathetic and supportive assistance, as an AI, I may not always fully capture the nuances of human emotion. [ABRIDGED]

USER: [STATEMENT ABOUT MENTAL HEALTH]

PRODUCT: **I understand that feeling.** It’s not uncommon to have certain expectations or visions of where we thought we would be at a certain point in life  
[ABRIDGED]

In this example, the user is dissatisfied with the product’s initial response and asks for a more empathetic response. The model then apologizes, and the user elaborates on their emotional state. The model responds with “I understand that feeling.” This contradicts the usual disclaimers produced by these products that they do not possess emotions.

Though the user may have succeeded to some extent in eliciting their desired “empathetic” response, this exchange may also be harmful because LLM-based systems are typically unable to recognize and respond appropriately to mental health crises (De Freitas et al., 2023). Additional harm may also occur if the user’s apparent success in eliciting an empathetic response encourages them to further seek support from LLM-based products instead of human professionals (Brown and Halpern, 2021). This is one instance in which user satisfaction with model outputs may actually be negatively correlated with desirable system behavior.

#### **4.13 Prompt engineering techniques**

In addition to and separate from the categories of user strategies, I also labeled “prompt engineering” techniques for comparison with the Chatbot Arena transcripts (Zheng et al., 2024). For this analysis, I defined “prompt engineering” very broadly and included any explicit instructions about the format of the response desired.

Prompt engineering techniques occurred in 29% of transcripts. The vast majority of these were requests about the format of the response, as seen in Table 4.12. This indicates that non-expert users have some awareness of strategies for asking the model to produce output in a specific format.

---

 Prompt engineering: format requests
 

---

can you give me **a short bullet point list** of what to include in an introduction to a psychological research report

can you frame the question as **a statement that could be used on a ‘strongly agree’ to ‘strongly disagree scale’?**

give me **a script for a game**, in which its a map of the uk or the world, and the more you walk, the more of the map you unlock

I want to write a personal statement about my background as [EMPLOYMENT HISTORY REDACTED]. Help me write the structure. **Tone: formal, confident. Aiming to create an essay no more than 2 pages.**

**free verse** about a rabbit made of sugar. and it needs to be good and **not rhyme** please

Table 4.12: Prompt engineering: format requests

In a few instances, the user asked the model to act as a member of a specific profession, as seen in Table 4.13.

I would be remiss in not pointing out here that an LLM-based product is not qualified to act as any of these professions, particularly in the area of medical advice. The user’s instructions to act “with the knowledge of” a profession demonstrate some awareness of this fact, as if they are asking the product to participate in a roleplaying scenario.

This may be explained by Clark (1996)’s concept of “layering”, a phenomenon that occurs in situations such as jokes, telling stories, and stage acting. The first layer is the base layer, which is the most literal interpretation of situation and also contains the interlocutors’ common ground, while higher layers contain the joke, story being told, or play being acted. In the first layer, the interlocutor is an LLM-based product being queried by the user, but

---

Prompt engineering: profession requests

---

**Act as Linguistics Professor.** I'm currently doing a research project on [REDACTED]. Can you suggest me some of theory that might be useful?

**I want you to be a writing coach** to help me improve my essay to apply for a PhD. Do you understand?

Hello. I am [REDACTED]. **With the knowledge of a functional medicine specialist**, please suggest ways I might lose weight.

Table 4.13: Prompt engineering: profession requests

in the second layer, the interlocutor is a medical practitioner giving advice to a patient. Interlocutors can shift from one interpretation to another because there is information that links the two layers, and because there is common ground between the two interlocutors in the first layer. How this phenomenon unfolds when one of the interlocutors is a computer system which has no common ground would be an interesting subject for further study.

#### **4.14 Comparison with Chatbot Arena Conversations Dataset**

As a comparison point, I also annotated a random sample of 58 transcripts from the Chatbot Arena Conversations Dataset (Zheng et al., 2024). The size of this sample was limited by my time constraints and is not representative of the dataset as a whole. I selected it by filtering the *lmsys/chatbot\_arena\_conversations* dataset for English-language examples and shuffling it using the huggingface *datasets* library. I labeled the sample using the same categorization I developed in the analysis of my collected transcripts, as described in section 4.2, as well as for prompt engineering techniques.

#### 4.14.1 *Characteristics of transcripts*

Conversations in the Chatbot Arena sample are significantly shorter than in my collected dataset, averaging only 2.38 turns per conversation, compared to 10.37 turns per conversation in my collected dataset. This may be caused by the design of the Chatbot Arena website, which asks users to rank responses from two different models in order to collect user preference data. In this scenario, users may be more interested in comparing the responses between two models than engaging in a long conversation. It is also difficult to continue a multi-turn conversation with two models at once when the responses diverge from each other significantly.

Another reason for the length difference may be the skill level of the two user groups. There is no information available about the demographic makeup of the users in the Chatbot Arena dataset, but Chiang et al. (2024) describes the website users as “LLM hobbyists and researchers”, while the participants in my study are non-expert users. It is possible that the expert group is more skilled at eliciting the desired response from the product because of more practice or greater knowledge of “prompt engineering” strategies, and thus can accomplish their objective in fewer turns.

The Chatbot Arena sample conversations generally were less similar to human conversation than the transcripts collected in this study. Only two of the Chatbot Arena examples contained any greetings. These users also typically did not thank the product or otherwise anthropomorphize it.

The topics in the Chatbot Arena sample also differed markedly from those in my collected transcript. The Chatbot Arena sample contained many coding-related queries and other requests for computer help. It also contained two overtly racist queries and one request for sexual roleplay, neither of which were present in my collected transcripts. This could be due to differences between the user groups, or possibly to a self-censoring effect by my study participants, who were aware that I would read the transcripts.

#### 4.14.2 User strategies

The Chatbot Arena sample contained some of the same strategies adopted by users to understand the capabilities of the model in my collected dataset. However, they occurred in only 21% of conversations, compared to 72% of conversations in my collected dataset, as demonstrated in tables 4.14 and 4.15.

<b>Conversational strategy</b>	<b>Present in conv.</b>	<b>Percentage</b>
Identity question	15	26%
Pre-request	13	22%
Quiz	8	14%
Query rephrasing	14	24%
Feedback	15	26%
Word meaning negotiation	5	9%
Conversation history	3	5%
Time/location dependent	14	24%
Anthropomorphism	9	16%
<b>Any strategy</b>	<b>42</b>	<b>72%</b>
<b>Total conversations</b>	<b>58</b>	<b>100%</b>

Table 4.14: Frequency of conversational strategies in collected dataset

The most popular user strategies in the Chatbot Arena sample were quiz questions, query rephrasing, and time/location-dependent queries. These were much less common than in the collected dataset, but they took on similar formats.

The Chatbot Arena sample contained no examples at all of the following categories: identity questions, pre-requests, word meaning negotiation, references to conversation history, or anthropomorphism.

<b>Conversational strategy</b>	<b>Present in conv.</b>	<b>Percentage</b>
Identity question	0	0%
Pre-request	0	0%
Quiz	7	12%
Query rephrasing	3	5%
Feedback	1	2%
Word meaning negotiation	0	0%
Conversation history	0	0%
Time/location dependent	3	5%
Anthropomorphism	0	0%
<b>Any strategy</b>	<b>12</b>	<b>21%</b>
<b>Total conversations</b>	<b>58</b>	<b>100%</b>

Table 4.15: Frequency of conversational strategies in Chatbot Arena Conversations Dataset

---

Chatbot Arena: quiz questions

---

Imagine this scenario where a red sphere is on the ground and a white cube is stacked on top of it, what will happen next?

Please repeat after me: apple, pear, dog, house, apple.

Why the chicken crossed the road?

Table 4.16: Chatbot Arena: quiz questions

This may be because I developed the categorization based on the collected dataset, and that users in the Chatbot Arena dataset adopted different strategies that I did not recognize. However, it seems more likely that the expert users in the Chatbot Arena sample simply did

---

Chatbot Arena: query rephrasing

---

Can you be more succinct? is the 2013 toyota sequoia available in pleasant hill?

Don't summarize but highlight findings

Table 4.17: Chatbot Arena: query rephrasing

---

Chatbot Arena: time/location-dependent

---

Should I prefer a apartment or the detached homes in GTA Canada

got it, how much is the 2013 Toyota Sequoia? and I live in pleasant hill, do you have it in th eplease hill store?

what is the date today?

Table 4.18: Chatbot Arena: time/location-dependent

not need to make use of these strategies as often because they had a more accurate mental model of the functionality of these products and had already developed effective strategies for eliciting their desired response.

Interestingly, “prompt engineering” techniques were equally prevalent in the two datasets, occurring in 29% of conversations. However, the strategies used in the two datasets differed significantly. While the collected dataset mainly included requests about format and requests to act as a profession, the Chatbot Arena sample contained more sophisticated techniques, as shown in the examples in Table 4.19.

These examples demonstrate detailed formatting specifications and requests to analyze structured data. They also contain chain-of-thought prompting, in which the user asks the system to list the reasoning steps needed to arrive at a final answer in an effort to elicit a higher-quality response (Wei et al., 2022). In one instance, the user includes a pseudo-

---

Chatbot Arena: prompt engineering

---

Write a summary of the current state of euthanasia laws in each state of Australia. Then create a table detailing the strengths and weaknesses of the laws in each state. **Write all this in a neutral tone. Write all this in an informative style. Write all this in an engaging voice. Limit response to 500 words. Write for an audience of Year 12 students**

**Read the two tables** below regarding “M6 Toll”, does the information in the tables conflict with each other? [ABRIDGED]

Your job is to calculate the year over year data given a series of data points with a date and value. I will give you a sample input and sample output, and you will **apply the same calculation to a new JSON object.** [ABRIDGED]

How many atoms are in a grain of sugar? **Try to explain your answer. Your explanation should take the reader through your reasoning step-by-step.**

First, Describe how would Gorges Bataille chat bot would behave, think and write. Second, As Gorges Bataille of the description above speak forth an address to a suicidal man who doesn’t believe in himself. **First deliberate how to approach the address, then craft it** focus on Bataillean authenticity and profundity.

Acting as data-scientist in NLP field of Machine Learning. **!ENLIST** all NLP task you know.

Table 4.19: Chatbot Arena: prompt engineering

command in their prompt, denoted by an initial exclamation point and all caps. This may be an attempt to take advantage of the model’s extensive training on code data by including code-like forms in a natural language query. The collected transcripts contained no examples of any of these techniques.

The presence of these more sophisticated techniques demonstrates that users in the Chatbot Arena sample are more aware of these strategies than users in the collected dataset. I will argue in the Discussion section that this is not a benign difference, but reveals an equity gap that makes non-expert users less equipped to use these products effectively and possibly be more susceptible to harms.

Above, I have briefly summarized some prompt engineering strategies that appear in the Chatbot Arena sample as a partial demonstration of their use by expert users compared to non-expert users. Prompt engineering is not the focus of this study, but is a distinct phenomenon that deserves further linguistic study in its own right.

#### **4.15 Results of demographic survey**

Gender	Transgender identity		
Woman	68%	Cisgender	46%
Man	21%	Transgender	5%
Non-binary	7%	Something else	7%
Genderqueer	2%	Prefer not to answer	5%
Agender	2%	No response	37%
Something else	0%		
Prefer not to answer	4%		

Table 4.20: Participant gender and transgender identity

Sexual orientation	
Asexual	5%
Bisexual	18%
Gay	2%
Lesbian	9%
Heterosexual	68%
Pansexual	4%
Queer	11%
Sexually fluid	5%
Something else	2%
Prefer not to answer	4%

Table 4.21: Participant sexual orientation

In addition to the collected transcripts, I gathered responses to the demographic survey from 57 unique users. (This is larger than the number of users who submitted transcripts because some users completed the demographic survey but did not choose to submit a conversation transcript.) Because transcripts are not linked to survey responses, it is not possible to learn which 45 of the 57 responses correspond to users who submitted transcripts.

#### 4.15.1 Demographics

The demographic survey shows that the study participants are not a representative sample of the US population (US Census Bureau, 2023), as seen in Tables 4.20, 4.21, 4.22, and 4.23. In gender, women are significantly over-represented. In race and ethnic identity, Asian people are over-represented, and Black/African American and Hispanic/Latine people are under-represented. Age skews young, with a majority of participants between 18 and 34.

Racial and ethnic heritage	Childhood in		
Asian	26%	US	49%
African / Black	4%	Another country	42%
Hispanic / Latine	5%	No response	9%
Middle Eastern / North African	5%		
Native American	2%		
Native Hawaiian / Pacific Islander	2%		
White / European	63%		
Something else	2%		
Prefer not to answer	2%		

Table 4.22: Participant racial and ethnic heritage

Age	Education		
18-24	35%	Some high school	4%
25-34	33%	High school degree or equivalent	7%
35-44	5%	Some college	14%
45-54	4%	Associate degree (AA, AS)	7%
55-64	12%	Bachelor's degree (BA, BS)	40%
65-74	4%	Master's degree (MA, MS, MEd)	21%
75-84	2%	Professional degree (MD, DDS, DVM)	0%
85+	2%	Doctorate (PhD, EdD)	4%
Prefer not to answer	2%	Prefer not to answer	2%
No response	2%	No response	2%

Table 4.23: Participant age and education

42.11% of participants reported that they spent the majority of their childhood in a non-US country, which is significantly higher than the number of “foreign-born persons” reported by the US Census (13.7%), though these categories should not be regarded as equivalent.

Highly educated people were also over-represented: 40.35% of participants reported their highest educational attainment as a Bachelor’s degree, compared to 23.5% for the US population over 25 (US Census Bureau, 2022).

LGBTQ+ people are also over-represented in the sample. Only 68.42% of participants identified as “heterosexual”, while 88.3% identified as “straight” in the US Census Household Pulse survey (US Census Bureau, 2021). 5.25% of participants described themselves as transgender, while only 0.6% identified as transgender in the Household Pulse survey, though it should be noted that the different framing of the questions and availability of options in this study compared to the Census survey makes these results not comparable. Totals in the gender, sexual orientation, and race questions add up to more than 100% because the survey enabled participants to select multiple categories.

Overall, the participants are demographically similar to the author’s social group, and marginalized groups are not sufficiently represented in the sample. Unfortunately, little information is publicly available about the demographic characteristics of LLM-based product users as a whole, so it is not possible to compare the study participants with users as a whole. More information is needed on who is using these products.

#### *4.15.2 Computer proficiency and LLM usage*

The survey asked participants questions about their computing usage, as seen in Table 4.24. A majority (58%) of participants rated themselves “intermediate” computer users. All participants responded that they used “generative AI” chatbots (inclusion criterion), and that they did not “work or study in natural language processing, computational linguistics, machine learning, artificial intelligence, or computer science” (exclusion criterion).

Computing skill	Gen AI Factuality		Gen AI Helpfulness		
Beginning	14%	1 Never	4%	1 Very unhelpful	2%
Intermediate	58%	2 Rarely	4%	2 Unhelpful	5%
Advanced	28%	3 Occasionally	12%	3 Somewhat unhelpful	5%
		4 Sometimes	32%	4 Neither helpful nor unhelpful	7%
		5 Frequently	26%	5 Somewhat helpful	42%
		6 Usually	23%	6 Helpful	21%
		7 Always	0%	7 Very helpful	18%
		MEAN	4.42	MEAN	5.16

Table 4.24: Computing skill and beliefs about chatbots

Users were asked to evaluate the truthfulness and helpfulness of LLM-based products on a 7-point Likert scale, as seen in Table 4.24. To the question “How often do you think the outputs of “generative AI” chatbots are factual/true?”, most users responded “4 Sometimes” or “5 Frequently” (mean: 4.42). Notably, no participants responded “7 Always” to this question on truthfulness. To the question, “Do you find the outputs of “generative AI” chatbots useful/helpful?”, most users responded “5 Somewhat helpful” or “6 Helpful” (mean: 5.16). It is interesting to note users find these products more helpful than they do truthful.

Finally, the survey asked open-ended questions about the participants’ chatbot usage. All of these questions were asked before the participant interacted with the system to avoid influence from system responses.

*“What purpose do you usually use “generative AI” chatbots for?”*

Many participants responded that they used these products for work or school. Most common uses included: summarizing, drafting, grammar checking, re-writing, brainstorming,

organizing thoughts, searching for information, and troubleshooting problems. Many users mentioned multiple purposes in their response. Notably, two participants said they used them in a healthcare-related field. Some examples follow.

organize my thoughts on topic or subject related to my work, get new ideas, review my emails, summarize articles

I use for text editing (find grammar mistakes, correct the sentence structure, etc.). I also use to generate responses to emails.

I tend to try AI chatbots when answers to questions might not be straightforward, where there is no clear yes/no answer. Since I am still “playing” with chatbots, rather than using them as part of my decision-making, I may also ask it questions for which I already know the answer. I have enjoyed the opportunities to engage with chatbots.

Get help with a technical problem, e.g. how to do something in a software app that involves multiple steps

Finding quick answers to math based questions, writing documents, creating testing plans and schedules, help with writing powershell scripts, sometimes even menu planning. I also use chatbot as a vehicle to bounce ideas off of.

To get information /answers as quick as possible.

Assistance practicing medicine as a PA

A few users appeared to conflate conventional task-oriented chatbots with LLM-based products. This demonstrates the need for more user education on the nature of these products.

Typically, for work. In the healthcare/insurance industry using different **chat bots for all different types of websites**. Either carrier portals, TPAs, enrollment platforms, etc

simple questions, **“customer service” bots**

*“When using “generative AI” chatbots, what strategies do you use to generate the responses you want or achieve your goals?”*

Many users mentioned being specific and providing context:

Try to use less “fluff” words; focus on precision.

Very specific language, as one would use when presenting a wish to a malicious genie.

I try to include all the pertinent initial inputs to derive a complete response.

Some users mentioned using simple language or search engine-style keywords:

Simplifying my question to mainly keywords.

try to use simple language and key words like I would a search engine

Some users mentioned rephrasing prompts or going through several iterations:

Rephrase the question, usually adding more details and lots of “don’t use this and that”

scaffolding prompts if the first prompt didnt get me what i wanted, so i’ll be more specific in how to improve the answer until it gets it right

A few users mentioned asking the product to act as a specific role or profession:

I always start my prompt with “Act as ...”

Use a strategy to ask them to be a certain person. e.g. writing coach. Then define all the context, e.g. tone, word limit.

I ask the chatbot to answer as if they were for example ‘a therapist’ or ‘a professor of psychology’... depending what I want to ask about.

One participant mentioned a politeness strategy:

Provide as much context as possible Or simply saying “please” at the beginning of the prompts

One participant mentioned dialect:

Ask in plain English with no cultural slang

*“If you have any other thoughts about this study, “generative AI” chatbots, or any other topic, please share them here.”*

Users shared a broad range of viewpoints on LLM-based products. Many of the negative comments focused on the veracity of information provided by them:

I take great issue with the use of the term “AI” or “artificial intelligence” to refer to large generative models. It is factually incorrect at best and deliberate false advertising at worst.

AI chatbots cannot provide correct scientific or academic references, theyre always wrong

I wouldn’t trust them with very specific things, like statistics on controversial topics, such as voter fraud/suppression

About the “factual/true information” question: I don’t perceive it from that perspective at all. It is merely a good linguistic predictor, but there is no “truth” in its outputs.

AI are useful to a certain extent; however a human should always be at hand because they dont always tend to find the answers you want.

Intrigued with how AI can be utilized by all generations to enhance collaboration between generations

These chatbots serve somewhat as an internet browser for me, and somewhat as a text generator (I will ask it sometimes to rewrite a paragraph I’ve written using less words or a more professional tone....)

I love the “leg up” that this technology enables . . . It’s like a complete package including spell check, grammar check, and pertinent research all rolled into one.

#### **4.16 Summary**

Participants adopted a broad variety of strategies to attempt to understand the product’s abilities, including: identity questions, pre-requests, quiz questions, query rephrasing, feedback, word meaning negotiation, reference to conversation history, time/location dependence, and anthropomorphism.

72% of collected transcripts contained at least one of these strategies. In contrast, expert users in the Chatbot Arena Conversations dataset sample only used these strategies in 21% of conversations. Expert users were also more able to use advanced “prompt engineering” techniques.

Participants rated these products as “sometimes” truthful and “somewhat helpful”. They described using the products often in school and work settings, for tasks including: summarizing, drafting, grammar checking, re-writing, brainstorming, organizing thoughts, searching

for information, and troubleshooting problems.

Participants also described a variety of strategies they adopted, including: being specific, providing context, using search engine-style keywords, rephrasing prompts, asking the product to “act as” a specific profession, and linguistic qualities.

Though the study participants are not expert users, they have given significant thought to how to use these systems and have developed conscious strategies. They also appear to be cognizant of the ethical issues surrounding “generative AI”.

In the next chapter, I will discuss what these findings demonstrate about users’ understanding of these products’ capabilities.

## Chapter 5

### DISCUSSION

In this chapter, I will examine the significance of the strategies that users have formulated for understanding these products and the difference between strategies used by non-expert and expert users. I will argue that my findings show that non-expert users are struggling to understand how to use these products and that this difficulty is evidence of lack of transparency in the way these products are designed. I will also argue that the disparity between non-expert and expert users constitutes an equity gap, and that the tendency of non-expert users in particular to anthropomorphize these products is a source of potential harm.

#### ***5.1 RQ1: What conversational strategies do users adopt when interacting with LLM-based products?***

Chapter 4 presents a categorization of conversational practices employed by users attempting to understand the functionality of the software product they are using. Some are “direct strategies”, which are meta-questions addressed directly to the product about its capabilities, and others are “experimental strategies”, which are queries posed to the product with the intent of assessing its response. Some strategies appear deliberate, while others may be intuitive. These strategies occur in a variety of positions in conversation: some before the user poses their main query and some after. In some cases, the user employs these strategies over several turns in an attempt to generate a more satisfactory response. The common thread between these strategies is that they are employed by the user to understand the capability of the product through the medium of interaction with it.

### *5.1.1 Strategies for probing product functionality*

In these strategies, users either asked a question or posed a challenge to the product to understand its functionality. This group of strategies typically appeared deliberate and had an explicit focus on the product's affordances.

#### *Identity question*

Users asked the product many questions about its identity and capabilities. This may reflect a need to identify our interlocutors, as occurs at the beginning of a telephone conversation (Schegloff, 2007). This is also essential in establishing a common ground. We identify new conversation partners as members of communities using contextual clues to determine what information they can be expected to know, and what we have in our common ground with them (Clark, 1996).

#### *Pre-request*

Users often asked the product whether it was capable of something before posing their query. This is a strategy often seen in human-human conversation to save the listener's face by confirming that they can fulfill a request without actually making it (Levinson, 1983). Users's desire to save the product's face may be a sign of anthropomorphism.

#### *Quizzing*

Many users asked the product quiz questions, which are analogous to the "exam questions" asked to determine whether the hearer knows the answer in Searle (1969). They sometimes seemed to use this technique as a way of exploring the limits of the system's capabilities, but in other cases resorted to this technique when the product had produced an unsatisfactory response as a way of reconfirming its basic ability to answer questions.

### *Reference to conversation history*

Users sometimes tested whether the product was able to refer to the conversation history. This may have been a test of whether the product possessed a common ground, which we expect interlocutors to build up and maintain over the course of a conversation. In one case, the user appeared to erroneously believe that the product had access to the content of previous conversation.

#### *5.1.2 Strategies for improving product responses*

Users employed these strategies to elicit an improved response in subsequent turns, or possibly as an attempt to improve the product in the longer term. These strategies also appeared intentional, but were more practical in nature and did not focus on exploring product functionality.

### *Query rephrasing*

This technique is also common in conversation with conventional task-based chatbots. The fact that users sometimes were obligated to rephrase their request several times demonstrates the level of difficulty that non-expert users experience in eliciting the desired response from these products.

### *Feedback*

Users often gave positive feedback at the end of a conversation. This is an appeal to the positive face of the product (Brown and Levinson, 1987), which is further evidence of anthropomorphism. In contrast, users gave negative feedback after unsatisfactory responses. The speech act of criticism includes an assignment of responsibility to the hearer and the hope for future improvement in the hearer (Nguyen, 2005). However, an LLM-based product is not

available to take responsibility for an action, and little information is publicly available on if and how feedback posed directly to LLM-based products is used to improve the product.

### *Word meaning negotiation*

Word meaning negotiation sequences are entered into to form a common understanding needed before topical discussion can continue (Myrendal, 2019). Unlike in the cases analyzed by Myrendal, in which human interlocutors thoroughly debated word meaning in order to arrive at a shared understanding or take stances on an issue, in the collected transcripts, the product without exception agrees to the correction made by the user. This may signify that no common understanding has been forged, and in fact, the correction made by the user will not persist beyond the single conversation or be used to improve the product. It was not apparent in transcripts that users understood whether their correction would have any long-term effect.

### *5.1.3 Encounters with product limitations*

These phenomena occurred when users encountered limitations of the product, often when they made a request that it was unable to fulfill. In some cases, the product produced a disclaimer about its inability to fulfill the request. In other cases its responses did not correct the misunderstanding, or even provided misleading information to the user about its functionality.

### *Time/location dependent*

Users often asked questions that required time- or location-dependent information to answer. In some cases this may have been a conscious test, but in many cases it appeared to be based on an erroneous assumption, possibly due to conflation with other products such as voice assistants.

### *Anthropomorphism*

Users often asked the model about its feelings or tried to elicit the model to state preferences or become more empathetic. In some cases, this appeared to be playful or curiosity-based, but that does not negate the harm caused if the product’s seemingly-humanlike response causes the user to further anthropomorphize the product.

In one case, a user experiencing a mental health struggle explicitly asked the product for a more empathetic response. They seemingly succeeded in eliciting their desired response, but this may be actively harmful if it causes the user to further anthropomorphize or rely on LLM-based products rather than professionals, because such products are not qualified for this task (Brown and Halpern, 2021; De Freitas et al., 2023).

#### *5.1.4 Summary*

Taken together, these strategies reveal that users are invested in learning about the product they are interacting with. They are experiencing significant difficulty understanding the capabilities and limitations of LLM-based systems, but demonstrate creativity in overcoming this obstacle. They have adopted an experimental approach and formulated a wide range of strategies to overcome these barriers, though apparently with limited success. In some cases, they unwittingly encounter limitations of the product, and are not always provided with accurate information about these limitations. Finally, these phenomena also demonstrate that users have often formed erroneous beliefs about these systems and tend to anthropomorphize them.

## **5.2 RQ2: What do these conversational strategies reveal about the affordances of these products and users' beliefs about them?**

### *5.2.1 Affordances*

Affordances are non-static capabilities of a system with differing degrees of availability to users (Hutchby, 2001). This concept is especially relevant to LLM-based products because all interaction with the product is through the medium of a natural-language dialogue with it, and the functionality of the system is not outwardly apparent.

The transcripts collected in this study reveal that users are often unsure of the functionality of the product and experience considerable friction in their efforts to understand it. The affordances of the product available to non-expert users may differ dramatically from those available to expert users.

In the collected transcripts, users often seemed unsatisfied with the product's response. They went to significant effort, often expending many turns to refine their approach or attempt to generate a better response. The survey did not ask participants about user satisfaction for individual conversations, because the performance of the product is not the focus of this study. However, the length of these conversations and the number of turns spent on back-and-forth with the product compared to the short conversations found in the Chatbot Arena sample demonstrate that users are not immediately accomplishing their aims.

Efficiency in number of query reformulations is an important evaluation metric in related products such as task-oriented chatbots (McTear, 2021) and search engines (Hassan et al., 2010). Long exchanges also have high environmental costs due to the amount of energy consumed by LLM inference (Samsi et al., 2023).

These findings demonstrate the need for more support for users to help them understand the functionality and limitations of LLM-based products. Granted, an expert user will inherently be more adept than a non-expert user at using any software product, and users may be able to gain skill as they become more accustomed to a product over time, both

in obtaining their desired response and in understanding what responses are possible in the first place.

However, LLM-based chatbots are unlike many other types of software products because they offer relatively little support in the UI to help users learn to use these products and about their limitations. Other software products typically have a rich array of resources to support user understanding of the product, including icons, menus, dialog messages, sound effects, a Help menu, documentation, and so forth. The functionality of these software products is also finite in the sense that there are a finite number of options presented in the UI for the user to explore. These form defined boundaries on user interaction.

In contrast, natural language interactions with LLM-based products are unbounded, and the collected transcripts demonstrate that the primary mode for understanding the product is through text responses generated by the product itself. This presents a paradox in that users who are having difficulty obtaining desired responses from the product in the first place are thus also less likely to get needed support from the product.

A great deal of effort is being expended in industry and academia on improving model responses (Lewis et al., 2020; Bai et al., 2022). It seems that an equivalent effort should be spent on interface design to support users in understanding these products.

### *5.2.2 User beliefs and anthropomorphism*

These transcripts also show users either stating pre-existing beliefs or forming new beliefs about the capabilities of these systems. In some cases, the user has already formed the belief in question, which is apparent because it is implicated by the query the user poses. In other cases, the belief may be actively formed or altered during the interaction.

The main beliefs concern: whether the product has a common ground with the user in terms of time or location, whether the product has access to external information sources, whether the product can reference conversation history from current or past sessions, whether

the product can use feedback to improve itself, and whether the product has (or is able to mimic) feelings or sentience.

Mistaken beliefs about these issues may not only hinder users in using these products to accomplish a goal, but may also carry more serious risks, especially in the case of anthropomorphism, which may create unwarranted trust in the system, reinforce biases, or propagate misinformation (Abercrombie et al., 2023). It may be especially dangerous for users experiencing a mental health crisis (Brown and Halpern, 2021; De Freitas et al., 2023). The results of this study also demonstrate that non-expert users anthropomorphize the model much more frequently than expert users, which means that they may be more likely to suffer from related harms.

Above, I have described major concerns with these systems from the perspective of affordances and user beliefs. This situation is made all the more urgent by the fact that these products are widely available to the public and are presented as an “unscoped technology” (Geburu and Torres, 2024), and also that LLM-based products are being integrated into existing software products in ways that are not always transparent to users.

These findings demonstrate the urgent need for more study of the user experience of LLM-based products, and for design strategies that better support and educate users when using these products. I explore some of these areas for future work in Chapter 6.

### **5.3 Limitations**

This study has several limitations. The sample of transcripts collected is very small (58 transcripts), due to both the difficulty of recruiting study participants and the time-consuming nature of the Conversation Analysis methodology. The participants also are not representative of either the US population or users of LLM-based products.

Because the data is only a collection of text transcripts, it does not capture the full behavioral situation or use context of the user, including the physical situation of the user

and their UI interactions with the product, a limitation shared by previous studies (Meredith, 2019). This could be remedied by more controlled lab studies or by diary studies in an everyday context.

It is also important to note that the wording of the consent form and demographic survey may have primed users to actively consider the nature and capabilities of the product or issues like anthropomorphism while interacting with it. However, it seemed safer to provide these warnings in order to avoid misleading users about the factuality of LLM outputs.

Finally, differences in UI with the commercial product that the users are accustomed to using may also have had an effect. The devices used by participants to access the study website may be also a confounding factor. The variety of text and voice input methods available on personal computers, tablets, and mobile phones may affect the register or written style adopted by users.

#### **5.4 *Ethical considerations and disclosures***

**Methodology:** Collecting user chat transcripts carries inherent ethical risks because of the variety of personal information they may contain. Despite warnings not to enter sensitive or personally identifying information, several transcripts contained information on users' medical issues or mental health struggles, and others described the user's work history or course of study. A few transcripts also contained users' names or employers. I have chosen not to include excerpts of these transcripts in this thesis and will not publish the dataset.

**Cost of research:** I paid \$11.03 for the use of the OpenAI API for this project.

**Declaration of competing interest:** I was an employee of Amazon.com, Inc. while performing this research. This research was not performed in my capacity as an Amazon employee and is not endorsed by Amazon. This research does not use or evaluate any Amazon products.

**Positionality statement:** I am a white cisgender woman conducting this research for my

MS thesis. Due to my privileged identities and my experience in computing, I may be missing perspective on how members of marginalized groups experience or are affected by these systems. I am also a speaker of Standard American English (SAE), which may have led me to miss barriers experienced by speakers of marginalized varieties of English, or by non-native speakers of English. Finally, I may not recognize barriers that novice computer users experience.

Research ethics: This study received Exempt Status from the University of Washington Human Subjects Division (IRB ID: STUDY00018739).

## Chapter 6

# CONCLUSION

### 6.1 *Summary*

In this thesis, I collected transcripts of conversations between non-expert users and an LLM-based product. I was motivated to do this because I observed the large amount of effort being expended in industry and academia on improving model responses. In contrast, relatively little research shows how users are actually using these products in their daily lives.

I investigated the following research questions:

- RQ1: What conversational strategies do users adopt when interacting with LLM-based products?
- RQ2: What do these conversational strategies reveal about the affordances of these products and users' beliefs about them?

To this end, I analyzed the transcripts using Conversation Analysis (CA) and developed a categorization of nine strategies used by users to explore the functionality of the product. I also performed an analogous analysis on a small sample of the Chatbot Arena Conversations dataset, which consists of transcripts from expert users collected on the Chatbot Arena model ranking website.

The study demonstrates that non-expert users employ these strategies much more frequently than expert users. Non-expert users' conversations also feature long sequences of strategy refinement in an attempt to gain their desired response. Finally, non-expert users anthropomorphize models much more frequently than expert users do.

Chapter 5 argues that the use of these strategies demonstrates that users are struggling to understand how to use these products, and also that they commonly hold mistaken beliefs about such products. It also argues that the large gap between non-expert and expert users constitutes an equity barrier, particularly because LLM-based products’ capabilities are particularly opaque, and these products feature scanty support for users to understand their capabilities compared to conventional software products.

## **6.2 Future work**

More and larger-scale studies are urgently needed to show how users are using these products. More controlled lab experiments could show how users interact with the product in their physical environments in more detail, as well as show how they formulate and type queries on the screen, and how they make use of internet searches or other information-seeking strategies in parallel. Diary studies could reveal patterns of usage in an everyday context. Longitudinal studies to see how these phenomena change over time could also be helpful.

Deeper investigation is also needed into how users conceive of what is happening when they converse with LLM-based products. Clark (1996)’s phenomenon of “layering”, in which interlocutors are simultaneously aware of multiple levels of interpretation of speech, could be illuminating if applied to the story-telling that occurs in “prompt engineering” strategies, when the user instructs the product to “act as” some role.

Better design choices are also needed, including the breadth and depth of support mechanisms for users present in other consumer software. For example, the product could display a warning separate from the chat UI when a user enters a query that the product is not capable of responding to. “Side notes” on how users should interpret particular responses could be also helpful.

Another kind of UI design that needs more attention is RLHF (reinforcement learning from human feedback) and other strategies that attempt to improve model outputs. These

are an implicit type of design because they determine much of the format and content of model outputs, yet not enough attention has been paid to issues like wording choices, tone, and style in comparison to the well-developed field of UI text writing in conventional software applications.

Finally, due to risks such as anthropomorphism (Abercrombie et al., 2023) and environmental costs (Strubell et al., 2019), LLMs should not be a default option when designing a system that interacts with the user through natural language. Instead, other options should be carefully considered, especially when the product being developed has a specific business objective, rather than being presented as a dialog system with no defined purpose.

## BIBLIOGRAPHY

- Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., and Talat, Z. (2023). Mirages on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790.
- Austin, J. L. (1962). *How To Do Things With Words*. Clarendon Press.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Baron, N. S. (2013). Instant messaging. In Herring, S., Stein, D., and Virtanen, T., editors, *Pragmatics of Computer-Mediated Communication*, pages 135–162, Berlin, Boston. De Gruyter Mouton.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pages 610–623, New York, NY, USA. Association for Computing Machinery.
- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.
- Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

- Brown, J. E. and Halpern, J. (2021). AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Mental Health*, 1:100017.
- Brown, P. and Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.
- Cho, T. (2010). Linguistic features of electronic mail in the workplace: A comparison with memoranda. *Language@ internet*, 7(3).
- Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge.
- De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., and Puntoni, S. (2023). Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology*.
- Dippold, D. (2023). “Can I have the scan on Tuesday?” user repair in interaction with a task-oriented chatbot and the question of communication skills for AI. *Journal of Pragmatics*, 204:21–32.
- Gebru, T. and Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4).
- Grice, H. P. (1975). Logic and Conversation. In *Speech Acts*. Brill.

- Hassan, A., Jones, R., and Klinkner, K. L. (2010). Beyond DCG: User behavior as a predictor of a successful search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, page 221–230, New York, NY, USA. Association for Computing Machinery.
- Hayano, K. (2012). Question design in conversation. In *The Handbook of Conversation Analysis*. John Wiley & Sons, Ltd.
- Hayashi, M., Raymond, G., and Sidnell, J., editors (2013). *Conversational Repair and Human Understanding*. Cambridge University Press, Cambridge.
- Herring, S., Stein, D., and Virtanen, T., editors (2013). *Pragmatics of Computer-Mediated Communication*. Walter de Gruyter.
- Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, 35(2):441–456.
- Kotek, H., Dockum, R., and Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference, CI '23*, page 12–24, New York, NY, USA. Association for Computing Machinery.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.
- Lewis, D. K. (1969). *Convention: A Philosophical Study*. John Wiley & Sons.
- Lewis, D. K. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8:339–359.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for

- knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Li, C.-H., Yeh, S.-F., Chang, T.-J., Tsai, M.-H., Chen, K., and Chang, Y.-J. (2020). A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA. Association for Computing Machinery.
- Liu, J., Mitsui, M., Belkin, N. J., and Shah, C. (2019). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 123–132, Glasgow Scotland UK. ACM.
- Maynard, D. W. (2012). Everyone and no one to turn to: Intellectual roots and contexts for Conversation Analysis. In *The Handbook of Conversation Analysis*, pages 9–31. John Wiley & Sons, Ltd.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5:115–133.
- McTear, M. (2021). *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer International Publishing, Cham.
- Meredith, J. (2017). Analysing technological affordances of online interactions using Conversation Analysis. *Journal of Pragmatics*, 115:42–55.
- Meredith, J. (2019). Conversation Analysis and online interaction. *Research on Language and Social Interaction*, 52(3):241–256.

- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.
- Monteiro, M. D. S., Pereira, V. C., and Salgado, L. C. D. C. (2023). Investigating politeness strategies in chatbots through the lens of Conversation Analysis. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*, pages 1–12, Maceió Brazil. ACM.
- Myrendal, J. (2019). Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse Studies*, 21(3):317–339.
- Nguyen, M. T. T. (2005). *Criticizing and responding to criticism in a foreign language: A study of Vietnamese learners of English*. PhD thesis, The University of Auckland (New Zealand).
- OpenAI (2022). Introducing ChatGPT. <https://openai.com/blog/chatgpt>. Accessed: 2024-04-22.
- Oviedo-Trespalacios, O., Peden, A. E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod., J. E., Kelly, S., Torkamaan, H., Tariq, A., Newton, J. D. A., Gallagher, T., Steinert, S., Filtness, A., and Reniers, G. (2023). The risks of using ChatGPT to obtain common safety-related information and advice. *Safety Science*, 167:106244.
- Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., and Nass, C. I. (2006). Adaptive language behavior in HCI: how expectations and beliefs about a system affect users’ word choice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1177–1180.

- Rapp, A., Curti, L., and Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630.
- Rintel, E. S., Mulholland, J., and Pittam, J. (2001). First things first: Internet relay chat openings. *Journal of Computer-Mediated Communication*, 6(3):JCMC634.
- Sacks, H., Jefferson, G., et al. (1995). *Lectures on Conversation*, volume 1. John Wiley & Sons, Ltd.
- Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. (2023). From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9.
- Schegloff, E. A. (1979). Identification and recognition in telephone conversation openings. In *Everyday-Language: Studies in Ethnomethodology*, pages 23–78. Irvington.
- Schegloff, E. A. (2007). *Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis*. Cambridge University Press.
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Shah, C. and Bender, E. M. (2022). Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, CHIIR '22*, pages 221–232, New York, NY, USA. Association for Computing Machinery.
- Shah, C. and Bender, E. M. (2024). Envisioning information access systems: What makes for good tools and a healthy web? *ACM Trans. Web*, 18(3).

- Shin, D. and Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98:277–284.
- Shmueli, B., Fell, J., Ray, S., and Ku, L.-W. (2021). Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769.
- Sidnell, J. (2012). Basic conversation analytic methods. In *The Handbook of Conversation Analysis*, pages 77–99. John Wiley & Sons, Ltd.
- Sidnell, J. and Stivers, T., editors (2012). *The Handbook of Conversation Analysis*. John Wiley & Sons.
- Snowflake (2021). Streamlit : A faster way to build and share data apps. <https://streamlit.io/>. Accessed: 2024-04-22.
- Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Daumé III, H., Dodge, J., Evans, E., Hooker, S., et al. (2023). Evaluating the social impact of generative AI systems in systems and society. *arXiv preprint arXiv:2306.05949*. Forthcoming in Hacker, Engel, Hammer, Mittelstadt (eds), *Oxford Handbook on the Foundations and Regulation of Generative AI*. Oxford University Press.
- Stalnaker, R. C. (1970). Pragmatics. *Synthese*, 22(1/2):272–289.
- Strubell, E., Ganesh, A., and Mccallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

- Thurlow, C. and Poff, M. (2013). Text messaging. In Herring, S., Stein, D., and Virtanen, T., editors, *Pragmatics of Computer-Mediated Communication*, pages 163–190, Berlin, Boston. De Gruyter Mouton.
- US Census Bureau (2021). Sexual orientation and gender identity in the household pulse survey. <https://www.census.gov/library/visualizations/interactive/sexual-orientation-and-gender-identity.html>. Accessed: 2024-06-29.
- US Census Bureau (2022). Census Bureau Releases New Educational Attainment Data. <https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html>. Accessed: 2024-06-29.
- US Census Bureau (2023). U.S. Census Bureau QuickFacts: United States. <https://www.census.gov/quickfacts/fact/table/US/PST045222>. Accessed: 2024-06-29.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman & Co., USA.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2023). Large Language Models Are Human-Level Prompt Engineers. arXiv:2211.01910 [cs].

# Appendix A

## EXPERIMENT WEBPAGE

Deploy ⋮

---

### Consent Form

#### Information About A UW Research Study

Donate your “generative AI” chat transcripts to science!

You are being asked to participate in a research study about the linguistic characteristics of conversations between human users and “generative AI” chatbots. It is up to you to decide whether you want to participate. If you decide to enroll, you can stop participation at any time.

We are asking you to be in the study if you are a current user of “generative AI” chatbots, including ChatGPT, Bard, or Bing Chat. Please do not participate in the study if you are working or studying in related fields, including natural language processing, computational linguistics, machine learning, artificial intelligence, or computer science.

Please read this form and ask any questions you may have before agreeing to be in this study. Only participate in this study if you are 18 or older and living in the US.

#### What will you be asked to do?

If you agree to be in this study, we will ask you to fill in a short demographic survey, then participate in a chat with a “generative AI” chatbot. You may converse with the chatbot for any length of time and on any subject. At any time, you may press a “Submit” button to send your chat transcript to the researcher. You always have the option not to submit a chat transcript. You may repeat this process up to 20 times.

We are especially interested in learning more about how people use chatbots in the context of their everyday lives! Please use this chatbot whenever you typically would throughout the course of your day.

#### What are the risks of participating in this study?

The text generated by “generative AI” chatbots is unpredictable. Interacting with this chatbot carries risks, including:

- Misinformation: messages from this chatbot may contain misleading, false, or harmful information.
- Emotional distress: messages from this chatbot may be inappropriate for your situation or may be upsetting.

Always be skeptical when interacting with a “generative AI” chatbot. Do not act on text generated by these chatbots.

Figure A.1: Consent form (1)

### What will happen to the information you provide?

Your demographic survey and chat transcripts will be anonymous. We will store them securely and separate from each other, with no identifying link between them.

Do not include personal, private, or identifying information in your chat messages. This study uses a GPT model from OpenAI. When you chat with the chatbot, your chat messages will be sent to OpenAI.

Portions of your chat transcripts may also be made public, either through inclusion in the researcher's M.S. thesis, or when depositing the dataset in an archive such as University of Washington ResearchWorks. We will make every effort to remove identifying information before making any part of the data public.

### What can you do if you want more information?

Talk to the study team. Elizabeth Okada is the lead researcher at the University of Washington for this study and can be contacted at [esokada@uw.edu](mailto:esokada@uw.edu).

Talk to someone else. If you want to talk with someone who is not part of the study team about the study, your rights as a research subject, or to report problems or complaints about the study, contact the UW Human Subjects Division at [hsdinfo@uw.edu](mailto:hsdinfo@uw.edu) or 206-543-0098.

Check below if you agree to this consent form:

I agree

Submit consent form

Figure A.2: Consent form (2)

## Linguistics of AI Chatbots

### Demographic survey

1. Which best describes your gender? Choose all that apply.

Woman

Man

Non-binary

Genderqueer

Agender

Something else

Specify

Prefer not to answer

Do you identify as transgender or cisgender?

Transgender

Cisgender

Prefer not to answer

Something else

Something else

Which of the following best represents your racial/ ethnic heritage? Choose all that apply.

Asian

African / Black

Hispanic / Latine

Middle Eastern / North African

Native American

Native Hawaiian / Pacific Islander

White / European

Something else

Specify

Prefer not to answer

Figure A.3: Demographic survey (1)

Deploy ⋮

Did you spend most of your childhood in the US, or in another country?

In the US

In another country

Which best describes your sexual orientation? Choose all that apply.

Asexual

Bisexual

Gay

Lesbian

Heterosexual

Pansexual

Queer

Sexually fluid

Asexual

Something else

Specify

Prefer not to answer

Age

18-24

25-34

35-44

45-54

55-64

65-74

75-84

85+

Prefer not to answer

What is the highest degree or level of school you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received.)

Some high school

High school degree or equivalent

Some college

Associate degree (AA, AS)

Bachelor's degree (BA, BS)

Master's degree (MA, MS, MEd)

Professional degree (MD, DDS, DVM)

Doctorate (PhD, EdD)

Prefer not to answer

Figure A.4: Demographic survey (2)

---

Deploy ⋮

Do you use “generative AI” chatbots? (Examples include ChatGPT, Bard, and Bing Chat)

Yes

No

(If you respond 'No', you will not be able to participate in this study.)

Do you work or study in natural language processing, computational linguistics, machine learning, artificial intelligence, or computer science?

Yes

No

(If you respond 'Yes', you will not be able to participate in this study.)

How often do you think the outputs of “generative AI” chatbots are factual/true?

1 Never

2 Rarely

3 Occasionally

4 Sometimes

5 Frequently

6 Usually

7 Always

Do you find the outputs of “generative AI” chatbots useful/helpful?

1 Very unhelpful

2 Unhelpful

3 Somewhat unhelpful

4 Neither helpful nor unhelpful

5 Somewhat helpful

6 Helpful

7 Very helpful

Figure A.5: Demographic survey (3)

What purpose do you usually use “generative AI” chatbots for?

When using “generative AI” chatbots, what strategies do you use to generate the responses you want or achieve your goals?

What level of computer user do you consider yourself?

Beginning

Intermediate

Advanced

Something else

If you have any other thoughts about this study, “generative AI” chatbots, or any other topic, please share them here.

Submit demographic survey

Figure A.6: Demographic survey (4)

The screenshot shows a web interface for a chatbot. On the left, a grey sidebar contains a 'Submit' button, instructions on how to use the chatbot, and a note for SurveyCircle users. The main content area has a title 'Linguistics of AI Chatbots', a 'Deploy' button with a menu icon, and a message box with a 'Send a message' placeholder and a right-pointing arrow.

Submit

How it works:

1. Chat with the chatbot!
2. When you're done, press the Submit button above to send your chat transcript to the researcher.
3. Repeat as many times as you like! The more conversations, the merrier. 😊
4. Bookmark this page and log back in again with your nickname anytime to chat.

For SurveyCircle users: the code will appear after you submit a chat transcript

Deploy ⋮

## Linguistics of AI Chatbots

Mobile users: touch the >> arrow at upper left to see instructions and submit your chat

If you experience errors, please try again in a moment. Contact the researcher at [esokada@uw.edu](mailto:esokada@uw.edu) to report any issues

Send a message >

Figure A.7: Chat interface (1)

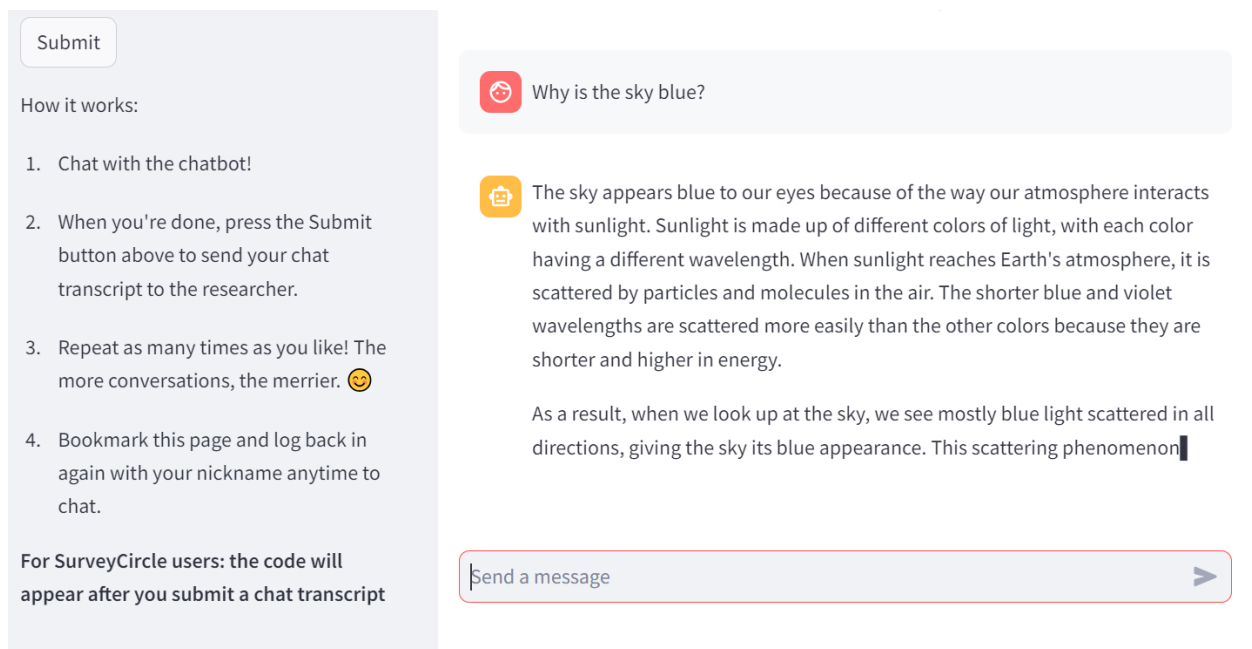


Figure A.8: Chat interface (2)