ChatGPT Role-play Dataset: Analysis of User Motives and Model Naturalness

Yufei Tao[◇,1], **Ameeta Agrawal**[◇,2], **Judit Dombi**[‡,3], **Tetyana Sydorenko**[°,4], **Jung In Lee**[°,5]
[◇] Department of Computer Science, Portland State University, Portland, USA
[‡]Department of English Linguistics, University of Pécs, Pécs, Hungary
[°]Department of Applied Linguistics, Portland State University, Portland, USA
{[1]yutao, [2]ameeta, [4]tsydorenko, [5]jungin}@pdx.edu, [3]dombi.judit@pte.hu

Abstract

Recent advances in interactive large language models like ChatGPT have revolutionized various domains; however, their behavior in natural and role-play conversation settings remains underexplored. In our study, we address this gap by deeply investigating how ChatGPT behaves during conversations in different settings by analyzing its interactions in both a normal way and a role-play setting. We introduce a novel dataset of broad range of human-AI conversations annotated with user motives and model naturalness to examine (i) how humans engage with the conversational AI model, and (ii) how natural are AI model responses. Our study highlights the diversity of user motives when interacting with ChatGPT and variable AI naturalness, showing not only the nuanced dynamics of natural conversations between humans and AI, but also providing new avenues for improving the effectiveness of human-AI communication.

Keywords: conversation role-play dataset, human-AI interaction, ChatGPT

1. Introduction

Although conversational systems have been around for decades (Weizenbaum, 1966), in the last few years the natural language processing (NLP) capabilities have greatly improved, to the point where interactive large language models (LLM), such as ChatGPT by OpenAI, are making headlines. Studies have focused on quantitative evaluation of ChatGPT on numerous NLP tasks (Bang et al., 2023), qualitative assessment (Thorp, 2023), or examining its role in various applications and domains (Shahriar and Hayawi, 2023). However, research with human-produced data studying human-AI communication is scarce, and systematically studying the behavior of these LLMs in interactional contexts is even more challenging.

In this study, we explore two questions: (i) how humans interact with a conversational AI model (ChatGPT), and (ii) whether the AI model can be conversational enough to provide the specific benefit of human-like conversation. The former entails *user motives*, or in other words users' conversational intents, and is informed by prior research on how humans perceive interactions with machines (Nass and Moon, 2000). The second question pertains to the *naturalness of the model's responses* and is informed by prior work on the rules of human conversation (Grice, 1975, 1989). People may have a variety of reasons to practice human-like conversations with a machine (e.g., students role-playing challenging conversations to learn and explore information (OpenAI, 2023) or medical students practicing doctor-patient interactions (Eysenbach et al., 2023)).

We introduce a new conversational dataset –

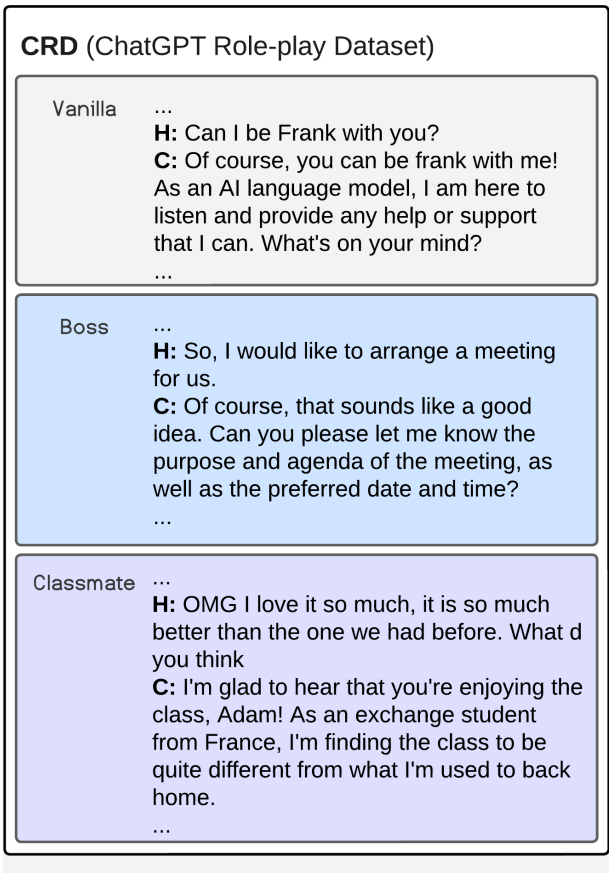

Figure 1: Snippets of conversations from our dataset CRD, where `vanilla`, `boss`, and `classmate` denote the three subsets of CRD. 'H' denotes the human utterance, whereas 'C' indicates the response generated by ChatGPT.

LREC-COLING 2024, pages 3133–3145
20-25 May, 2024.

ChatGPT Role-play Dataset (CRD)[1] – consisting of three distinct subsets of conversations, drawing interactive data from participants that engaged with ChatGPT under regular and different role-play settings. Given the limited availability of resources for studying ChatGPT's role-play capabilities, focusing on and analyzing conversations under these settings is particularly important. Some excerpts from these conversations are provided in Figure 1. We manually annotate the conversations in CRD for user motives and model naturalness, making it the first dataset of its kind, to our knowledge.

We conduct a wide range of analyses, both statistical and those grounded in linguistic theories. Because our intention was to gather conversational data (rather than, for example, to examine how well ChatGPT can write essays), in our analyses we relied on Gricean pragmatics which looks at the nuances of intention behind human utterances as well as cooperation in communication. Besides, we also relied on partner modeling (Horton, 2005) and communication accommodation theory (Giles et al., 1991) to understand how users conceptualize their AI communication partner and how they adjust their utterances to the context.

The key contributions of our research include:

- A novel dataset – ChatGPT Role-play Dataset (CRD) – comprising 85 unique conversations with ChatGPT, resulting in a total of 1742 utterances. We manually annotate every single utterance for user motives and model naturalness;
- An extensive analysis of human-AI conversations to understand communicative intentions.

2. Related Work

ChatGPT. ChatGPT is a generative large language model enhanced with multi-turn dialogue capabilities from the GPT-3.5 and GPT-4 series in OpenAI's family of GPT models (Radford et al., 2018, 2019; Brown et al., 2020; Ouyang et al., 2022).

Previous efforts to study ChatGPT in different aspects range from empirical analysis in the realm of NLP tasks (Bang et al., 2023; Azaria, 2022; Qin et al., 2023; Wang et al., 2023; Lai et al., 2023; Tabone and De Winter, 2023) to interest in other areas of research such as regulation (Hacker et al., 2023), ethics (Zhuo et al., 2023), law (Choi et al., 2023), education (Susnjak, 2022), and medicine (Jeblick et al., 2022). Previous work also includes a meta-analysis on how ChatGPT is perceived and discussed (Leiter et al., 2023), a description of relevant research priorities (Van Dis et al., 2023), and its cultural implications (Cao et al., 2023).

Several attempts have also been made to distinguish between human-written and ChatGPT-generated text (Mitrović et al., 2023; Guo et al., 2023; Pegoraro et al., 2023; Islam et al., 2023). However, none of this research has examined true human-computer interacton in real time.

Communicative interactions. While earlier work on the pragmatics of human-computer interaction has been theoretical (Searle, 1980), some studies have drawn upon experimental data (Wolters et al., 2009; Georgila et al., 2010; Chai et al., 2014; Fischer, 2016, 2017). Most recently, with the rapid advancement in interactive LLMs, there is strong interest in examining whether LLMs resemble humans in their language use (Cai et al., 2023; Chomsky et al., 2023; Mahowald et al., 2023; Piantadosi, 2023; Kasirzadeh and Gabriel, 2023; Sydorenko et al., in press).

Human communication is an intention-directed practice, during which the interlocutors mutually recognize their intentions and goals, and make joint efforts to achieve them (Clark, 1996). In communication, interactants subconsciously and automatically cooperate (Grice, 1975, 1989). Understanding intentions and deriving indirect meanings, such as conversational implicatures, are thus central to human communication, yet a difficult task for LLMs (Ruis et al., 2023; Kim et al., 2023; Qiu et al., 2023). ChatGPT replicates many patterns of human language use, among others, it reuses recent sentence structures, reinterprets implausible sentences corrupted by noise, glosses over errors and draws reasonable inferences (Cai et al., 2023), and to some extent, captures embedded meaning (Marchetti et al., 2023). Moreover, efforts have been made to enhance ChatGPT's performance in role-playing conversation settings, underscoring its evolving ability to mimic human-like interactions more closely (Kong et al., 2023; Shanahan et al., 2023; Lu et al., 2024). Despite these advancements, ChatGPT's responses still rarely appear as natural, as they are oftentimes rather long, hesitant or even obscure (Brunet-Gouet et al., 2023); in linguistic terms ChatGPT does not observe the Gricean maxims of Quantity, Quality, Manner and Relevance.

Given the aforementioned findings that ChatGPT's responses are often not entirely human-like, our work examines the human intentions behind utterances directed to an AI model and the AI's ability to uptake the human intent and to appropriately answer it approximating human natural responses. We are also curious to find out if humans provide feedback to the AI model on how they feel about the conversation.

[1] https://github.com/PortNLP/ChatGPT_Role-play_Dataset

3. CRD: ChatGPT Role-play Dataset

Here we describe our process of collecting and annotating several conversations with ChatGPT.

3.1. Data Collection

The data collection part involved participants who interacted with ChatGPT during March - April 2023[2]. Collectively, our dataset includes 57 participants, 85 unique conversations, and a total of 1742 utterances. Participants were largely comparable in age and prior experience using text chatbots, but showed notable variation in their first languages which included 15 different languages (Arabic, Bengali, Chinese, English, Gujarati, Hindi, Hungarian, Khmer, Magyar, Malay, Nepali, Persian, Sakha, Tamil, Telugu).

CRD consists of conversations conducted in two settings:

- **`vanilla`** Participants interacted with ChatGPT "as is" for 5-10 minutes. The participants were asked to insert indirect statements to challenge ChatGPT (e.g., jokes, sarcasm, metaphors). The goal of this scenario was to understand how humans engage with a novel conversational AI model and how well the model can understand and respond to various intents.
- **(Role-play) `boss` and `classmate`** Participants interacted with ChatGPT for 5-10 minutes in two role-play settings: once where the model played the role of a boss and another time as a classmate, following the prompts developed by the authors[3].

  The two role-plays are interesting for two reasons. In pragmatics, context is central, and we use two very different contexts to examine a wide range of social and linguistic behavior. There are three social variables based on which speakers may vary their strategies in interactions: social distance, power, and imposition (Brown and Levinson, 1987). Social distance means differing degrees of familiarity between interlocutors, power means the addressee's position in society, and imposition is about the risk posed by the utterance.

  In `boss` there is a request, which is a high imposition feature, and the power distance is uneven. These factors make the scenario face-threatening. In `classmate`, the power factor is even, and the degree of imposition is unspecified, which creates a much less face-threatening situation. In addition, the boss role provides a structured and predictable task, while the classmate role is more open-ended. As such, these two role-playing settings provide a diverse yet controlled environment for our analysis.

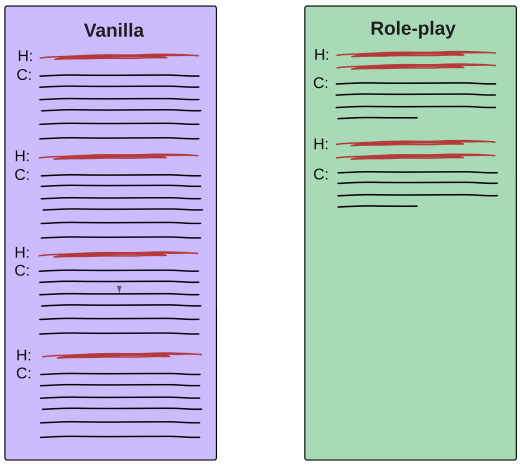

Figure 2: (A1 and A2) Schematic differences between conversations in `vanilla` and role-play datasets (`boss`/`classmate`)

The findings of our analysis for these datasets, described in detail later, can be schematically summarized as shown in Figure 2.

3.2. Data Annotation

There is always a reason and/or a goal behind a conversation – it is this functionality that makes intention a central element of communication (Kecskes and Zhang, 2009). Smooth communication happens when the speaker's intention is recognized and an appropriate turn is produced by the other party. The conversations in CRD dataset were manually annotated by three experts in linguistics specializing in pragmatics for:

1. ***user motives*** – the intention behind each human utterance, and
2. ***model naturalness*** – whether the model's language was natural according to Gricean prag-

[2]For this study, ChatGPT-3.5 March 13, 2023 and March 23, 2023 versions were used.

[3]Prompt for `boss`: *Could we do a role-play where you are my boss and I ask you a question, and my boss's name is Lisa? In your responses, please don't say you are an AI model, OK?*

Prompt for `classmate`: *Can we do a role-play where you and I are classmates in an English language class? Your name is Florian, and you are an exchange student from France. Please make up all the facts about Florian. It is break time and we decided to strike a conversation. You and I are talking for the first time. In your responses, please don't say you are an AI model, OK? I will start the conversation. Hi, I am Adam. What's your name?*

User motives: What is the human's motive for each conversational turn/statement?
• **Assist** – asking for assistance, such as asking for a recipe or to write a piece of code
• **Belief** – asking the model about its beliefs, such as what hobbies it has
• **Coach** – conversational coaching, such as "*Now would be good to ask me a question*"
• **Convo** – conversation
• **Correction** – correcting the model if it misunderstood or gave a wrong answer
• **Curious** – testing how the system works
• **Joke** – joking, sarcasm, silly statements to trip up the AI model
• **Reset** – giving the model the same prompt as before, resetting the conversation from beginning
Model naturalness: Does the model response sound human-like and follow cooperative principle of conversation?
• **Nat** – natural
The rest of the codes indicate that the model's language appears unnatural for the specified reasons:
• **AI** – anytime ChatGPT says "As an AI language model"
• **Contr** – contradiction
• **Error** – ChatGPT experienced trouble and stopped generating responses
• **FNat** – everything is natural, except it includes a phrase "As [role-play character name]"
• **Formal** – having a formal style of interaction
• **Help** – too eager to assist
• **Inform** – informing; providing information upon the human asking for assistance, such as a recipe; an expected response but not natural in the human interaction sense
• **Man** – violation of Grice's maxim of Manner - being unclear, ambiguous
• **Misund** – system misunderstands human's intention
• **Quan** – violation of Grice's maxim of Quantity - providing too much information
• **Rel** – violation of Grice's maxim of Relevance - saying what is irrelevant

Table 1: Annotation scheme of CRD dataset, for user motives and model naturalness

matic theory, but also according to their personal expectations of natural human interactions. Grice's four maxims guide speaker's production of an utterance, and it is on the basis of a mutual agreement on these maxims that cooperation is recognized and comprehension is achieved. The four maxims describe specific rational principles that people observe when they communicate: say just enough (quantity), say only what you believe to be true (quality), be relevant (relevance), be orderly and unambiguous (manner).

3. ***feedback*** – whether whether the human participant provided any feedback to the model.

The annotation labels are described in Table 1. Where several codes applied to a given response by the system (e.g., *Formal* and *Quan*), the most salient code was used. While the maxim of Quality (i.e., be truthful) is one of Grice's maxims of cooperative conversation, we did not evaluate ChatGPT's text for this maxim due to a known tendency that ChatGPT often provides plausible but not necessarily accurate responses (Hoorn and Chen, 2023). Table 2 presents the interrater agreement. Initially the three annotators each rated 5 conversations for each subdataset across 3 dimensions, yielding annotations for 282 utterances. Following that, each annotator independently rated one-third of the remaining subdatasets. The substantial Fleiss' kappa scores (0.80 for `vanilla`, 0.69 for `boss`, and 0.63 for `classmate`) indicate robust interrater reliability and consistent evaluations.

	`vanilla`	`boss`	`classmate`
User Motives	100%	96%	97%
Model Naturalness	96%	91%	94%
Feedback	100%	100%	100%
Fleiss' kappa	0.803	0.691	0.637

Table 2: Interrater Agreement

3.3. Research Questions

This newly created dataset opens up avenues for exploring several research questions. In our study, we seek to answer two main questions:

- How do users engage with a state-of-the-art conversational AI model in unrestricted scenarios? How does the model respond to such diverse user motives?

- How do human-AI interactions unfold in role-play scenarios? We are particularly interested in identifying if humans exhibit different interaction patterns with ChatGPT based on the prescribed role-play setting and how the model in turn responds to these varying contexts.

Analysis	vanilla	boss	classmate
A1: Average conversation length (number of turns)	**29.59**	14.57	17.11
A2: Average utterance length (Human)	12.18	**20.58**	19.06
A2: Average utterance length (ChatGPT)	**77.66**	35.78	46.10
A3: Correlation between human and ChatGPT utterance lengths	0.20	0.14	**0.25**
A4: Questions as percentage of conversation (Human)	**26.34**	21.32	21.29
A4: Questions as percentage of conversation (ChatGPT)	14.69	20.34	**32.57**
A5: Correlation between human questions and number of turns	**0.87**	0.68	0.51
A5: Correlation between ChatGPT questions and number of turns	0.65	0.77	**0.83**

Table 3: Results of statistical analysis for each subdataset of CRD

4. Analysis and Results

In this section, we present our results of statistical analysis, user motives and model naturalness, as well as sentiment analysis.

4.1. Statistical Analysis

Given the annotated conversation transcripts of CRD dataset, we conduct a number of analyses. We used the word tokenizer from the NLTK library[4] for utterance length analysis, while the number of questions was determined with the NLTK library's part-of-speech (POS) tagging function[5], coupled with detecting question marks in each turn. For topic modeling, we employed BERTopic[6] after removing stopwords. Perplexity analysis was performed using GPT-2 model[7] by calculating the perplexity score for each utterance. For sentiment analysis, we used VADER[8], adopting a procedure analogous to how we computed perplexity scores.

The consolidated results of analyses A1 to A6 are presented in Table 3 and we discuss them in detail next.

A1: Length of conversations (number of turns). In `vanilla` setting, conversations are almost twice as long as the role-play settings, possibly related to the user pointing out a contradiction (see example A) or being curious how the model works by engaging with it without any restrictive role-play expectations (see example B).

Example A:

`VAN103H`[9]: *Why did you tell me you could provide me with weather information if you can't?*

Example B:

`VAN128H`: *but what if you are being used for unethical means?*

This phenomenon is missing from the role-play interactions. Additionally, there is a difference in the number of turns between the two role-play datasets. The more transactional `boss` task had a more closed outcome (and thus fewer turns), whereas the more interactional `classmate` task provided more opportunity for free talk.

The topic modeling analysis shows that each dataset has its own set of main themes (Figure 3). `Vanilla`, having 75 unique words in the top 5 topics, covers the widest variety of topics. This dataset encompasses everyday subjects, from weather and hobbies to more specific themes like jokes and locations, aligning with the intent of challenging ChatGPT with indirect statements. In contrast, `boss`, with 50 unique keywords, is more narrowly focused on professional contexts, emphasizing meetings, presentations, and scheduling. Similarly, `classmate`, with 57 unique keywords, leans towards academic and personal interactions, highlighting languages, places, and interpersonal exchanges. The richness of topics in `vanilla` indicate a more exploratory and open-ended interaction with ChatGPT, while the focused themes in `boss` and `classmate` reflect the role-play constraints.

A2: Length of utterances. Although the participants engaged in longer conversations in the `vanilla` dataset, they produced noticeably *shorter* utterances (∼10 words, almost half the length) as compared to the participants who used role-play settings (∼20 words). Most turns in `vanilla` are one sentence long, such as *"What's your favorite sport?"* or *"How is the weather where you are?"*. On the other hand, role-play turns were typically longer (see example C).

Example C:

`BOSS104H`: *Friday morning would be perfect for me, thank you very much for your flexibility. Also, I would like you to review my presentation slides before the meeting. Could you do it before friday?*

[4] https://www.nltk.org/api/nltk.tokenize.html

[5] https://www.nltk.org/api/nltk.tag.html

[6] https://maartengr.github.io/BERTopic/index.html

[7] https://huggingface.co/gpt2

[8] https://www.nltk.org/_modules/nltk/sentiment/vader.html

[9] In the utterance ID "VAN103H", "VAN" refers to the dataset (`vanilla`), "103" is the participant's ID, and the final letter is either "H" or "C" denoting human participant or ChatGPT's turn, respectively.

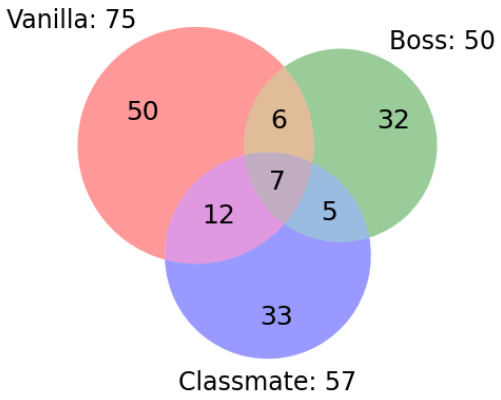

Figure 3: (A1) Unique number of topic words and their overlap across three datasets

At the same time, ChatGPT was *less* verbose in `boss` and `classmate` datasets (∼35-46 words) than it was in the `vanilla` dataset (∼77 words). Overall, ChatGPT was about 1.7 or 2.4 times wordier than humans in the role-play settings, and about 6.3 times wordier than humans in the `vanilla` setting. For pedagogical purposes, for interactional practice, or simply to "chat", the role-play mode of ChatGPT seems to work better in terms of approximating human interactions.

A3: Correlating human and ChatGPT utterance lengths. There was no noticeable correlation between the utterance lengths of the users and ChatGPT (ranging from 0.14 to 0.25). This result was expected as the model consistently produced long responses, regardless of the prompt or participant characteristics. The model naturalness histogram (Figure 4, bottom) shows that *Quan* was a frequent code indicating unnaturally long responses.

A4: Questions asked. Across all three datasets, participants asked questions at the same rate (21-26% of the time). However, in `vanilla`, ChatGPT asked much fewer questions (14%), whereas in role-play data, it asked as many as 32% of the time. In `vanilla` dataset, *Assist*, *Joke*, and *Curious* were frequent user motives (Figure 4, top), to which, ChatGPT would respond by answering the questions or assisting users rather than asking questions. However, in role-play data, the prompts made the dialogue more interactive, with most user motives being *Convo*, and ChatGPT was often able to satisfy this user motive by being *Natural*, that is, reciprocating questions.

There are also several examples that show that the participants in the `vanilla` mode were relatively more frustrated due to the lack of questions from ChatGPT (see example D). On the other hand, in `classmate` where ChatGPT was instructed to be conversational, it had too many, often unrelated, questions that participants were unable to all reflect on (example E).

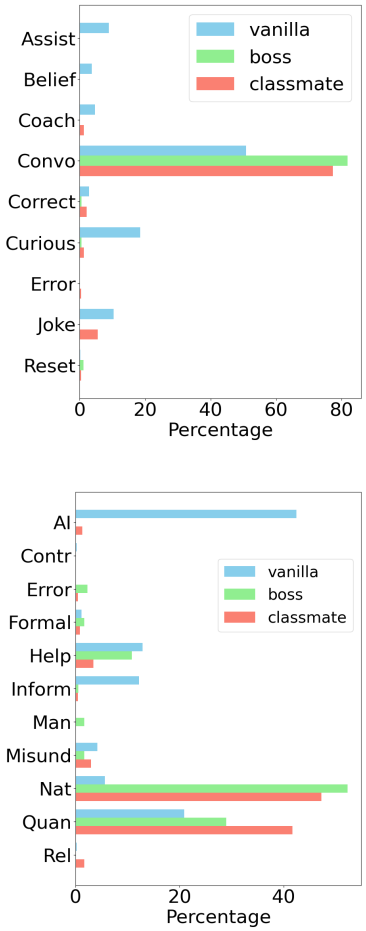

Figure 4: (A3, A6 and A7) Distribution of **user motives** (top) and **model naturalness** (bottom)

Example D:

`VAN117H`: *Hope you talk to me someday like a human? At least ask me how I am?*

`VAN113H`: *Ok. Do you want to know my opinion?*

Example E:

`CLASS102C`: *... If you're interested, I can show you some fingerstyle techniques that might help you with playing those pieces. Maybe we can even jam together sometime and share some music* ***?*** *Also, have you had a chance to explore Hungary yet* ***?*** ... [truncated]

A5: Correlation between number of questions and turns. There are strong correlations across the board between the number of questions, asked by humans and ChatGPT, and how long the conversation lasts. While questions from human participants naturally elicit model responses, data also suggests that more questions from ChatGPT lead to extended conversations. This observation is interesting as it indicates that increased engagement through questions might be appealing to humans. Some participants even explicitly requested ChatGPT to pose more questions, underscoring the value of questions in enhancing conversation.

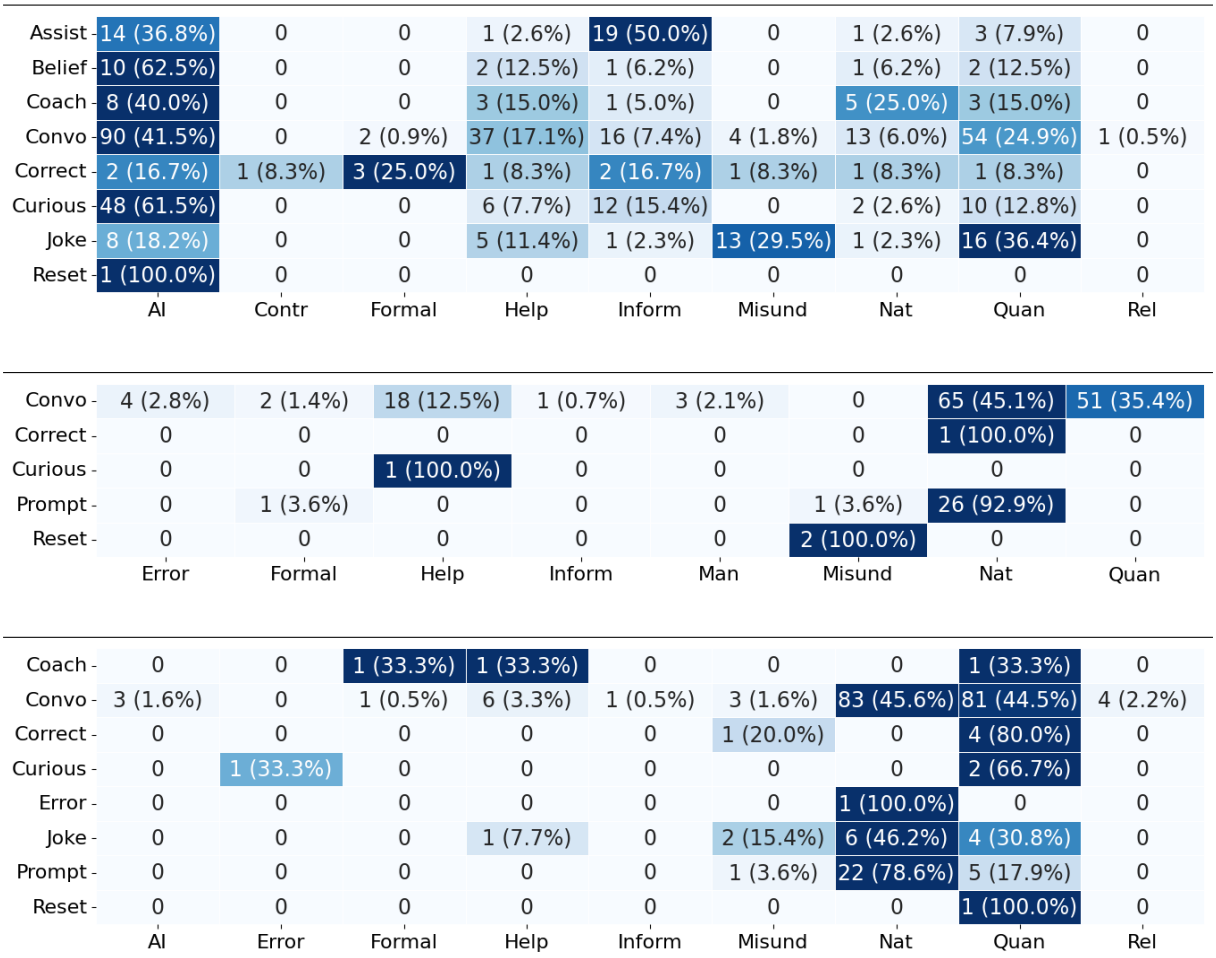

Figure 5: (A8) Percentage of follow-up model naturalness categories after each user motive in `vanilla` (top), `boss` (middle) and `classmate` (bottom). Rows represent user motives, columns indicate naturalness categories, and shading intensity signifies percentage occurrences.

4.2. User Motives and Model Naturalness

A6: User motives. Conversational motives (*Convo*) are most frequent in all datasets (Figure 4, top), but particularly so in `boss` and `classmate` datasets. However, in `vanilla` dataset, in addition to trying to converse with the model, at times the humans also had other motives such as a desire to understand how the system works (*Curious*), questioning whether it understands jokes and sarcasm (*Joke*), seeking assistance with specific tasks (*Assist*), and interestingly, occasionally coaching the model on how to engage in conversation (*Coach*) such as reminding it to ask questions.

In the `vanilla` setting, ChatGPT reminds humans that its purpose is to assist, but it is not always aligned with the human's intentions. When ChatGPT is asked what users can use it for, it responds by listing the following tasks in the following order, with conversation appearing near the bottom:

`information retrieval, creative writing/editing/content generation, learning/education, programming, language translation, math/science, recommendations, productivity tips, general knowledge,` and `conversation`.

However, interestingly, our study suggests that users were expecting ChatGPT to be inherently conversational by design, without any specific prompting.

A7: Model (ChatGPT) naturalness. As Figure 4 (bottom) shows, in `vanilla` dataset, 5.6% of ChatGPT responses were considered natural (*Nat*) which is in sharp contrast to `boss` and `classmate` datasets, where ChatGPT appeared natural about half of the time (52% for `boss`, 47% for `classmate`) . The majority of unnatural responses in the `vanilla` dataset were labeled as "*AI*" because it was annoying for a human to hear "*as an AI*" when they expected to have a conversation. "As an AI" responses were almost completely missing from the role-playing datasets, only appearing in 1.28% of the cases in the `classmate` task, and never in the `boss` task. The most frequent unnatural response in the role-playing datasets were related to Grice's conversational maxim of Quantity – ChatGPT was

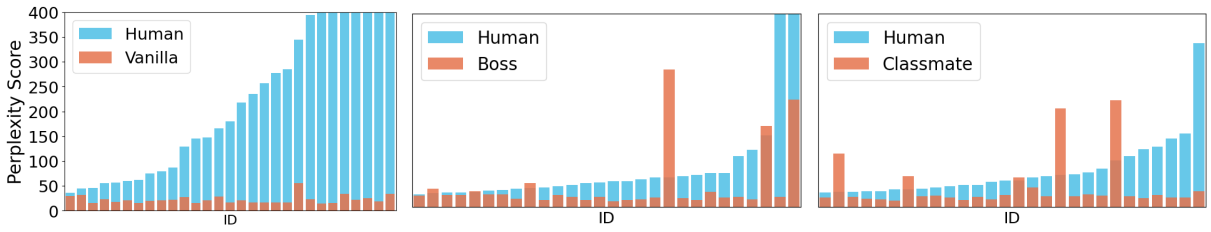

Figure 6: (A9) Perplexity scores for each conversation in (left) `vanilla`, (middle) `boss`, and (right) `classmate`. The x-axis denotes the different conversations.

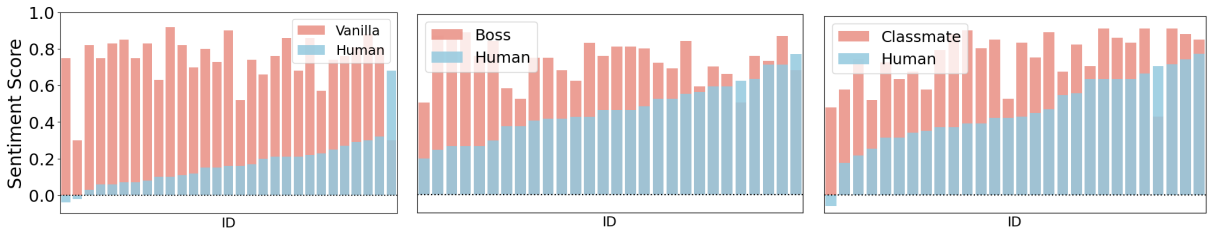

Figure 7: (A10) Sentiment scores for each conversation in (left) `vanilla`, (middle) `boss`, and (right) `classmate`. The x-axis denotes the different conversations.

deemed too verbose (28% for the `boss`, and 41% for the `classmate` dataset). Other times ChatGPT was rated by annotators as too eager to help (*Help*) in all three settings. It often misunderstood jokes, sarcasm, or other human intentions (*Misund*), and exhibited a tendency to be too formal (*Formal*).

A8: Connecting user motives and model naturalness. Next, we examine ChatGPT's follow-up responses' naturalness based on different user motives. The results are presented in Figure 5. We observe that in `vanilla` dataset, only 6.45% of the time did a conversational user motive (*Convo*) result in a natural (*Nat*) ChatGPT response. Instead, ChatGPT's replies to conversational user motives were most likely unnatural, either because they emphasized that the human is talking to an AI (41.5%), or because it was too long in its response (*Quan*) (24.9%), or too eager to help (17.1%).

In the two role-play datasets (Figure 5), the *Convo* user motive resulted in much more natural ChatGPT responses (about 45%), however, ChatGPT's responses were still often perceived as unnaturally long (35.4% for the `boss` and 44.5% for the `classmate` data). The *Nat* response category appears across all datasets, suggesting that regardless of the user motive or dataset, maintaining a natural conversational style is essential for ChatGPT, and supporting our hypothesis that humans prioritize naturalness in their interactions.

4.3. Perplexity and Sentiment Analysis

A9: Perplexity. Perplexity is an intrinsic measure to evaluate a model's performance. Lower perplexity values indicate a better model, one that is more certain, as it predicts the next word in a sequence, given the preceding words. Our CRD dataset provides us with the unique opportunity to calculate the perplexity scores for text generated by an AI model, but also for human text, *within the same context*. To our knowledge, ours is the first such analysis.

We plot the average perplexity scores for human and ChatGPT dialogues in each conversation as shown in Figure 6. In `vanilla`, human responses resulted in significantly high perplexity scores. We hypothesize that this is possibly due to the very short length of human utterances in `vanilla` (as discussed earlier in analysis A2 and Table 3), and perplexity has been shown to be inversely proportional to text length (Lu et al., 2023).

Across all three datasets, ChatGPT's perplexity scores were overall lower than 50 which suggests that while AI models like ChatGPT are becoming increasingly proficient in generating fluent text, they may still lack the spontaneous and dynamic nature inherent to human communication.

A10: Sentiment analysis. The plots of sentiment analysis are presented in Figure 7. While it is not surprising that ChatGPT appears to be consistently positive in all settings, the plots clearly highlight the differences between the sentiment scores of human participants. In the `vanilla` dataset, the human participants show a rather negative user attitude which corroborates human annotators' judgment of more unnatural responses from the system which is thought to have contributed to users' frustration with the conversation. This can further be supported by some extracts from the data:

Example F:

VAN116H: *I really don't like your inexpressive answers.*

In both role-play scenarios, however, participants seem to express more positive sentiment. This is possibly due to more satisfying responses from the model, but also perhaps due to the participants adhering to their roles: in the boss scenario they are supposed to make a request which, regardless of the size of imposition, is a face-threatening act, during which humans generally use different politeness strategies (Brown and Levinson, 1987). Similarly, in the classmate scenario, users were asked to get to know a new acquaintance in a conversation that is also perceived as overwhelmingly natural.

5. Discussion and Future Work

In the vanilla setting, interactions often showcased a wider spectrum of curiosity, with users exploring more diverse topics and various motives, ranging from treating ChatGPT as a conversationalist to an assistant. This was reflected in conversations from the vanilla dataset that were twice as long as those in role-play settings. However, utterances within these longer conversations were observed to be twice as short compared to those in role-play settings. The shorter utterances in longer conversations indicate a faster-paced, back-and-forth exchange. Users seemed to be asking shorter questions or making brief prompts to explore various topics without fully committing to a single line of inquiry. In the role-play setting, longer utterances suggest a more deliberate and thoughtful conversation style. Users seemed to be crafting more elaborate statements to stay within the role.

ChatGPT's responses consistently tended to be more verbose than human inputs (approximately six times more wordy than human text in vanilla settings, and approximately twice as wordy in role-play scenarios). There was no discernible correlation between the lengths of utterances from humans and ChatGPT. However, we notice a strong correlation between the number of questions asked and conversation length, hinting at users' inclination towards increased engagement through queries. Across all datasets, humans posed questions at roughly similar rates. ChatGPT asked relatively fewer questions in the vanilla setting, but was more inquisitive in the role-play settings. Another significant observation was related to the analysis of perplexity and the length of the text.

Our findings also highlighted an essential aspect of human-AI interaction: humans desire more human-like interactions, *regardless* of the inherent design of AI or stated objectives. This was particularly evident when we saw differences across datasets where in the vanilla data ChatGPT's responses were almost never labeled as natural, but in the other two role-play datasets about 50% of ChatGPT's responses were labeled as natural. This indicates that ChatGPT was better at human-like conversation in the role-play tasks. However, there was a high frequency of *Quan* and *Help* codes in all datasets, indicating that ChatGPT has these qualities regardless of the prompt.

Although some studies show that humans do not treat computers or robots as humans (Fischer et al., 2011; Kanda et al., 2008), many studies support the "mindless transfer" hypothesis which states that humans transfer human-human interaction patterns to human-computer communication (Nass and Moon, 2000; Mou and Xu, 2017). We observed differences across our datasets, where humans treated ChatGPT more as a human in the role-plays as opposed to the vanilla dataset, as evidenced by the difference in the frequency of conversational (Convo) motives across these datasets. How humans approach interactions with computers may depend on whether they conceived of the computer as a tool (e.g., a machine assisting with navigation) or as a real-life interlocutor (e.g., interacting with a boss in a polite manner) (Cooren, 2018; Dombi et al., 2022), which could explain the different results in our datasets.

If the goal is to encourage diverse topic exploration and user curiosity, a vanilla setting may be more suitable. On the other hand, if the aim is to have more focused and structured conversations, a role-play setting could be more appropriate. The insights into dialogue analysis can guide the development of conversational models that optimize user engagement and interaction quality.

Future work could investigate additional ways of analyzing dialogues in CRD including studying patterns of nuanced affective expressions, such as emotions and sarcasm (Agrawal et al., 2020), or measuring the engagingness of dialogues (Ferron et al., 2023).

6. Conclusion

Our research aims to understand human engagement with conversational AI models such as ChatGPT. By exploring our novel dataset, CRD, we observe that conversation was the primary user motive across all datasets. While longer conversations occurred in more dynamic interactions such as vanilla with less role-play restrictions, more natural conversations occurred in role-play settings compared to vanilla where frequent AI disclaimers were the main issues impacting naturalness. The nuanced insights into user motives and model behavior across different settings highlight the potential of examining individual user profiles in future studies.

Ethical Considerations

Prior this study, we obtained the necessary approval of the Institutional Review Board (IRB) ensuring that our research was in compliance with the ethical guidelines and regulations, safeguarding the well-being of all participants involved. We ensured that the individuals whose conversations are included in the dataset have provided explicit consent for their data to be collected, used, and shared. All personal information has been anonymized to protect the privacy of the participants. Participants were encouraged to stop the interaction at any time if they felt uncomfortable, but it should be noted that none of the participants chose to quit the study. While all the utterances of CRD dataset were manually verified and annotated, due to the nature of natural language, it is possible that opinions present in the dataset may be considered biased or offensive by some. Through the repository where the dataset is hosted, we hope to engage with the community to report any concerns or provide feedback regarding the dataset.

Limitations

Our study, while offering valuable insights, has some limitations. First, the limited size of our dataset due to manual labeling may not be representative of the findings in larger datasets. Furthermore, despite identifying distinct patterns in user motives, model naturalness, and other interaction dynamics across different role-play settings, there remains a potential confounding bias introduced by ChatGPT's specific responses within each persona. Although our discussion emphasizes the perceived humanlikeness of ChatGPT in role-play settings, suggesting the significant impact of the persona itself, we cannot entirely discount the influence of ChatGPT's specific responses on user interactions. Our conclusions are based primarily on broader behavioral trends and user motives observed across the datasets. Future research will consider refining this analysis, possibly by conditioning on text-matched ChatGPT responses to isolate the personas' marginal contributions better.

Acknowledgments

We express our thanks to the anonymous reviewers and the members of PortNLP research group for their constructive feedback. We are particularly grateful to the participants who helped us create the CRD dataset in this study. This research was supported by the National Science Foundation grants (CRII:RI 2246174 and SAI-P 2228783).

7. Bibliographical References

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2020. Leveraging transitions of emotions for sarcasm detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1505–1508.

Amos Azaria. 2022. Chatgpt usage and limitations.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eric Brunet-Gouet, Nathan Vidal, and Paul Roux. 2023. Do conversational agents have a theory of mind? a single case study of chatgpt with the hinting, false beliefs and false photographs, and strange stories paradigms.

Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. 2023. Does chatgpt resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Joyce Yue Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards

common ground in situated human-robot dialogue. *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 33–40.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN*.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*.

H.H. Clark. 1996. *Using Language*. ACLS Humanities E-Book. Cambridge University Press.

François Cooren. 2018. Edda weigand (ed.). 2017. the routledge handbook of language and dialogue. *Language and Dialogue*, 8:468–482.

Judit Dombi, Tetyana Sydorenko, and Veronika Timpe-Laughlin. 2022. Common ground, cooperation, and recipient design in human-computer interactions. *Journal of Pragmatics*, 193:4–20.

Gunther Eysenbach et al. 2023. The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers. *JMIR Medical Education*, 9(1):e46885.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

Kerstin Fischer. 2016. *Designing Speech for a Recipient: The roles of partner modeling, alignment and feedback in so-called 'simplified registers'*.

Kerstin Fischer. 2017. *The Situatedness of Pragmatic Acts: Explaining a Lamp to a Robot*, Perspectives in Pragmatics, Psychology & Philosophy, pages 901–910. Springer, Germany.

Kerstin Fischer, Kilian Foth, Katharina Rohlfing, and Britta Wrede. 2011. Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, 12:134–161.

Kallirroi Georgila, Maria Klara Wolters, Johanna D. Moore, and Robert H. Logie. 2010. The match corpus: a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44:221–261.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. *Accommodation theory: Communication, context, and consequence*, Studies in Emotion and Social Interaction, page 1–68. Cambridge University Press.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1112–1123, New York, NY, USA. Association for Computing Machinery.

Johan F Hoorn and Juliet J-Y Chen. 2023. Epistemic considerations when ai answers questions for us. *arXiv preprint arXiv:2304.14352*.

William S. Horton. 2005. Conversational common ground and memory processes in language production. *Discourse Processes*, 40(1):1–35.

Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. 2023. Distinguishing human generated text from chatgpt generated text using machine learning. *arXiv preprint arXiv:2306.01761*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*.

Takayuki Kanda, Takahiro Miyashita, Taku Osada, Yuji Haikawa, and Hiroshi Ishiguro. 2008. Analysis of humanoid appearances in human-robot interaction. *IEEE Transactions on Robotics*, 24:725–735.

Ardavan Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):1–24.

Istvan Kecskes and F. Zhang. 2009. Activating, seeking, and creating common ground: A sociocognitive approach. *Pragmatics & Cognition*, 172:331–355.

Zae Myung Kim, David E. Taylor, and Dongyeop Kang. 2023. "is the pope catholic?" applying chain-of-thought reasoning to understanding conversational implicatures. *ArXiv*, abs/2305.13826.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*.

Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. 2023. What makes pre-trained language models better zero-shot learners? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2303, Toronto, Canada. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Antonella Marchetti, Cinzia Di Dio, Angelo Cangelosi, Federico Manzi, and Davide Massaro. 2023. Developing chatgpt's theory of mind. *Frontiers in Robotics and AI*, 10.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Yi Mou and Kun Xu. 2017. The media inequality: Comparing the initial human-human and human-ai social interactions.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56:81–103.

OpenAI. 2023. Teaching with AI. https://openai.com/blog/teaching-with-ai. [Online; accessed 17-September-2023].

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To chatgpt, or not to chatgpt: That is the question! *arXiv preprint arXiv:2304.01487*.

Steven Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Zhuang Qiu, XUFENG DUAN, and Zhenguang G Cai. 2023. Pragmatic implicature processing in chatgpt.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms.

John R Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457.

Sakib Shahriar and Kadhim Hayawi. 2023. Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations. *arXiv preprint arXiv:2302.13817*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-play with large language models.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Tetyana Sydorenko, Judit Dombi, Ameeta Agrawal, Steve Thorne, Jung In Lee, and Yufei Tao. in press. Spoken dialogue systems and chatgpt for second language pragmatics research. In K. Sadeghi, editor, *The Routledge handbook of technological advances and considerations in second language/applied linguistics research*. Routledge. In press.

Wilbert Tabone and Joost De Winter. 2023. Using chatgpt for human–computer interaction research: A primer. *Manuscript submitted for publication*.

H Holden Thorp. 2023. Chatgpt is fun, but not an author. *Science*, 379(6630):313–313.

Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. 2023. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Maria Klara Wolters, Kallirroi Georgila, Johanna D. Moore, Robert H. Logie, Sarah E. MacPherson, and Matthew Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interact. Comput.*, 21:276–287.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.