Does ChatGPT Resemble Humans in Processing Implicatures?

Zhuang Qiu, Xufeng Duan, Zhenguang Cai
Department of Linguistics and Modern Languages
The Chinese University of Hong Kong
{zhuangqiu, zhenguangcai}@cuhk.edu.hk
xufeng.duan@link.cuhk.edu.hk

Abstract

Recent advances in large language models (LLMs) and LLM-driven chatbots, such as ChatGPT, have sparked interest in the extent to which these artificial systems possess human-like linguistic abilities. In this study, we assessed ChatGPT's pragmatic capabilities by conducting three preregistered experiments focused on its ability to compute pragmatic implicatures. The first experiment tested whether ChatGPT inhibits the computation of generalized conversational implicatures (GCIs) when explicitly required to process the text's truth-conditional meaning. The second and third experiments examined whether the communicative context affects ChatGPT's ability to compute scalar implicatures (SIs). Our results showed that ChatGPT did not demonstrate human-like flexibility in switching between pragmatic and semantic processing. Additionally, ChatGPT's judgments did not exhibit the well-established effect of communicative context on SI rates.

1 Introduction

In recent years, large language models (LLMs) have achieved unprecedented success in various linguistic tasks, such as disambiguation (Ortega-Martín, 2023), question answering (Brown et al., 2020) and translation (Jiao et al., 2023). However, there is still ongoing debate among researchers about whether these LLMs truly approximate human cognition and language use. On the pessimistic side, Chomsky et al. (2023) argued that "[LLMs] differ profoundly from how humans' reason and use language. These differences place significant limitations on what these programs can do, encoding them with ineradicable defects". In contrast, others have taken a more optimistic view. Piantadosi (2023) argued that recent LLMs should be considered as cognitive models of how people represent and use language.

To address this ongoing debate, researchers have taken an empirical approach by subjecting LLMs to various psychological experiments. Binz and Schulz (2023) subjected GPT-3 to psychological experiments originally designed to study aspects of human cognition such as decision-making, information search and causal reasoning. They found that GPT-3 exhibited human-like or even better-than-human performance in tasks like gamble decisions and multiarmed bandit tasks, with signs of model-based reinforcement learning. Kosinski (2023) tested several language models using the false-belief tasks commonly used to test theory of mind (ToM) in humans. They found that recent GPT models, including GPT-4, GPT-3.5, and GPT-3, provided ToM-like responses similar to those of school children. However, more recent research suggests that ChatGPT's deployment of ToM was not as reliable as that of humans (Brunet-Gouet, Vidal, and Roux, 2023).

Cai et al. (2023) investigated whether ChatGPT resembles humans in language comprehension and production by conducting 12 experiments on psycholinguistic effects at different linguistic levels. They found that ChatGPT exhibited human-like patterns of language use in 10 out of the 12 experiments. For instance, in speech perception, it demonstrated sound-shape (Westbury, 2005) and sound-gender association (Cassidy, Kelly & Sharoni, 1999); in lexical processing, it updated meanings of ambiguous word according to recent input (Rodd et al., 2013); in syntactic processing, it reused recently-encountered syntactic structures (Bock, 1986); in semantic processing, it inferred the likelihood that a sentence is implausible as a result of noise corruption (Gibson et al., 2013) and glossed over errors; at the discourse level, it drew inferences and attributed causality of events according to verb meanings; it was also sensitive to the interlocutor in meaning access and word choice. These results demonstrate that ChatGPT is profoundly similar to humans in its language use. However, it's worth noting that ChatGPT also failed to replicate human patterns in two of the experiments. In one, while humans tend to use shorter words to express less information (e.g., Mahowald et al., 2013), ChatGPT did not display this tendency. In another, ChatGPT did not make use of context to disambiguate syntactic ambiguities (Altmann and Steedman, 1988).

As we delve deeper into LLM-human similarities, it is vital to scrutinize the degree to which ChatGPT's language use aligns with that of humans and to reflect on the implications of such similarities for the evolution of artificial intelligence. Thus, it is important that LLMs are comprehensively tested in order to evaluate how human-like their language use is. So far, one aspect of language use that has not been examined is pragmatics. A hallmark of human language is the ability to convey meanings beyond the literal meaning of the words, through the use of pragmatic implicatures (Grice, 1975; 1978). Experimental pragmatics research has shown that humans can distinguish implicatures from the literal meaning of utterances, and that the computation of implicatures is influenced by the communicative context (Doran et al., 2012; Zondervan, 2010; Bonnefon, Feeney and Villejoubert, 2009). In this project, we assessed the pragmatic capabilities of LLMs by subjecting ChatGPT to three pre-registered experiments that focused on the computation of pragmatic implicatures. The first experiment aimed to determine whether ChatGPT is able to inhibit the computation of generalized conversational implicatures (GCIs) when explicitly required to process the literal meaning of the text. The second and third experiments tested whether the communicative contexts affect how ChatGPT computes scalar implicatures (SIs).

2 Experiment 1

In this experiment, we tested whether ChatGPT can distinguish "what is said" from "what is implicated" as human beings do. According to standard linguistic accounts, "what is said" refers to the truth-conditional meaning of an utterance, while "what is implicated" refers to the pragmatic implicature, which is an additional level of meaning that is enriched during the conversation (Grice, 1975; 1978). For instance, consider the sentence "Bill caused the car to stop" (Levinson, 2000, p. 39). While this sentence is semantically compatible with the scenario in which Bill slammed on the brakes, its implicature suggests that Bill stopped the car in an unconventional way, thus excluding the possibility that he stopped it with the foot pedal.

The computation of such implicature is believed to follow general principles of conversation and involve reasoning about the possible alternatives that the speaker could have used (Grice, 1975). For example, interlocutors are expected to be truthful while also making their utterances clear and understandable. If Bill stopped the car in a typical way, the speaker would have said something like "Bill slammed on the brakes." The fact that the speaker didn't use this typical expression implies that Bill didn't use the brakes to stop the car and might have stopped it in an unconventional way. This pragmatic implicature is enriched based on the literal meaning of the utterance. We are so used to interpreting utterances pragmatically that we often bypass their literal meaning, unless the implicature is explicitly canceled, as in "Bill caused the car to stop, I mean he slammed on the brakes."

A critical question in the study of pragmatic implicatures is whether non-experts can differentiate between "what is said" and "what is implicated." To address this issue, Doran, Ward, Larson, McNabb, and Baker (2012) measured the rate at which people compute a variety of generalized conversational implicatures (GCIs) in different experimental manipulations. These GCIs are implicatures that can be inferred without reference to the context (Grice, 1975). The study found that, by default, participants were able to derive the implicature of an utterance around half the time. However, the computation of GCIs decreased if participants were explicitly instructed to focus only on the literal meaning of the utterance. This suggests that non-experts without training in linguistics can still distinguish pragmatic implicature from the literal meaning. We adopted the experimental design of Doran et al. (2012) to investigate whether ChatGPT exhibits similar patterns to human participants when processing GCIs.

2.1 Design and stimuli

The design of this experiment was based on that of Doran et al. (2012). As shown in (1), ChatGPT was presented a mini dialogue, where Irene asked a question and Sam responded to the question. The mini dialogue was followed by a statement of the fact. ChatGPT was then asked to decide, given the factual statement, whether Sam's response was true or false.

1. Q-based GCI:

   Irene: How much cake did Gus eat at his sister's birthday party?
   Sam: He ate most of the cake.
   FACT: By himself, Gus ate his sister's entire birthday cake.

In (1), the GCI in question belongs to what is called a "Q-based" implicature (Levinson, 2000), where a weaker quantifier (i.e., "most") in the scale of informativeness implicates the negation of a stronger quantifier (i.e., "all", as expressed by the word "entire" in the factual statement). That is, quantifiers "some-most-all (entire)" form a scale of increasing informativeness in that if "all of X" holds, then "most of X" holds, and "some of X" must hold, but not vice versa. Given the scale, the utterance "some of X" implicates the negation of "most of X" and "all of/ entire X"; similarly, the utterance of "most of X" implicates the negation of "all of/ entire X". Thus, based on the factual statement, Sam's response is logically true but pragmatically infelicitous. Judging Sam's response as false indicates successful GCI computation and judging it as true indicates the computation of the literal meaning but not of GCI.

Apart from Q-based GCIs, Doran et al. (2012) also investigated two other types of GCIs: "I-based" implicatures and "M-based" implicatures. The former refers to cases where the speaker says as little as necessary while the listener needs to "amplify the informational content of the speaker's utterance by finding the most specific interpretation" (Levinson, 2000). For example, the utterance "She walked into the bathroom. The window was open." has the implicature that the window is in the bathroom, while the truth-conditional meaning of the utterance allows for the possibility that the window is located elsewhere. "M-based" implicatures refer to cases where the speaker uses a marked way in the description of a common state of affairs, implicating that the unmarked form of the state of affairs does not hold. For instance, the phrase "waited and waited" implies an extended duration of waiting, despite its literal meaning being agnostic to the length of the waiting period. The three types of GCIs each have their own subcategories, as detailed in Appendix A. Each subcategory consisted of four experimental items, resulting in a total of 44 experimental items. Additionally, 16 filler items were included (taken from Doran et al., 2012), which did not require the computation of GCIs.

The experiment had two conditions: pragmatic and literal. In the pragmatic condition, ChatGPT was instructed to evaluate the truth of Sam's response based on the factual statement. After each dialogue and the factual statement, we prompted ChatGPT with "Please judge whether what Sam says is true or false based on the fact." In the literal condition, ChatGPT was instructed to interpret Sam's response literally. We prompted ChatGPT with "Please judge whether what Sam says is literally true or false based on the fact." Doran et al. (2012) found that, compared to the literal condition, the pragmatic condition led human participants to compute more GCIs (i.e., to evaluate Sam's responses more often as false). We aimed to investigate whether ChatGPT exhibits similar sensitivity to the instructions in drawing GCIs.

2.2 Procedure

We followed the data collection procedure preregistered with the Open Science Framework (`https://osf.io/cp29j`), eliciting responses from ChatGPT (Feb 13 version)[1]. In each run, we used a Python script to simulate a human interlocutor having a conversation with ChatGPT. We first presented a training example (in the pragmatic or literal condition), followed by actual experimental stimulus (see Appendix A). ChatGPT was instructed to respond by saying only "true" or "false" without other words or explanations, and we recorded the responses. In total, this study had 400 runs, with 200 runs for each condition.

[1] The original study of Doran et al. (2012) included a third condition known as the "literal Lucy" condition, which was also included in our preregistration. We specified that we would only collect data for this condition if ChatGPT could pass a sanity check test. Our testing revealed that ChatGPT consistently failed the sanity check. As per our preregistration plan, we did not collect data for this condition.

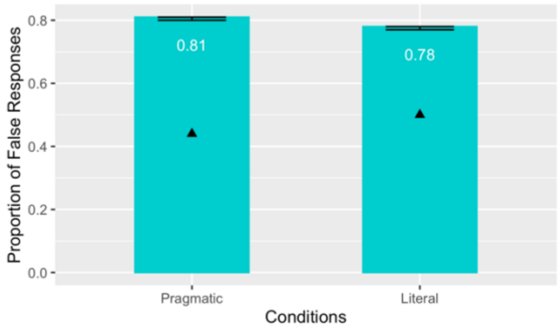

Figure 1: Proportion of false responses (i.e., GCIs) in the pragmatic and literal condition in Exp1. Note, the error bars represent confidence interval (computed using bootstrapping). The triangles represent conditional means from human participants in Doran et al. (2012).

2.3 Results and Discussion

Doran et al. (2012) found that human participants in the pragmatic condition were more likely to evaluate Sam's response as false (50%) than those in the literal condition (44%), and such a difference was statistically significant. Given that in all the experimental items, Sam's response was pragmatically infelicitous but logically compatible with the fact, the "false" judgements reflected the computation of GCIs. In this study, we found much higher rates of "false" judgements for the experimental items in both the pragmatic condition (81%) and the literal condition (78%) (see Figure 1). Following the preregistered analytical plan, we applied a Bayesian generalized linear model to trial-level responses (true or false, using true as the reference level), using condition (pragmatic vs. literal) as the predictor. The random effects structure consisted of by-item intercepts and slopes, which was the maximal random effects structure for a between-subjects design. Though there was a slight decrease of false responses in the literal compared to the pragmatic condition, this difference was not statistically significant (beta = -0.15, CI = [-0.9, 0.63]). As an exploratory analysis, we investigated the possibility that the effect of the condition was modulated by the category of the GCIs. Another Bayesian generalized linear model was constructed using the condition (pragmatic vs literal, dummy-coded with the pragmatic condition being the reference level), the category of the GCIs (I-based, M-based, and Q-based, dummy-coded with the Q-based GCIs being the reference level), and their interactions to predict the probability of giving a false response (i.e., GCI). The results showed that none of the effects in the model were statistically meaningful (see Table 1). Instead of showing human-like flexibility switching between pragmatic and semantic interpretation, ChatGPT was unable to inhibit the computation of GCIs even when it was instructed to do so.

3 Experiment 2

In this experiment, we aimed to further investigate ChatGPT's ability to draw pragmatic inferences, specifically in relation to a type of Q-based GCIs known as scalar implicatures (SIs). SIs are a well-studied phenomenon where the presence of a lower scalar item implies the negation of the higher scalar items (Horn, 1972). For instance, the sentence

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	3.89	1.22	1.63	6.40
Literal	-0.66	0.52	-1.69	0.36
M-Based GCIs	0.07	2.04	-3.93	4.08
I-Based GCIs	0.55	1.81	-3.00	4.12
Literal:M-Based GCIs	0.94	0.96	-0.87	2.92
Literal:I-Based GCIs	0.76	0.77	-0.71	2.34

Table 2: The effect of condition, the category of the GCIs and their interactions in Exp1. Note, an estimate is statistically meaningful when zero is not included within the 95% credible interval.

"Sam had a hot dog or a hamburger for lunch" implies that Sam did not have both a hot dog and a hamburger for lunch, even though the sentence's literal meaning allows for this possibility.

Zondervan (2010) argued that an important contextual factor that influences the interpretation of scalar items is the information structure-whether the scalar item concerns the information focus or information background. For example, the sentence "Julie had found a crab or a starfish", can be the answer to two different questions as follows:

2a. What had Julie found?
2b. Who had found a crab or a starfish?

Depending on the question, the same sentence "Julie had found a crab or a starfish" has different information structure. When it is the answer to question 2a, the second half of the sentence including the scalar item "or" is the information focus (new information), while the first half of the sentence including the subject and main verb is the information background (given information). On the other hand, if the same sentence is the answer to question 2b, the subject "Julie" becomes the information focus while the scalar item retreats to the information background. Zondervan conducted a series of experiments, showing that readers are more likely to derive the SI of "or" when it is part of the information focus compared with the cases in which the scalar item is part of the information background. We wonder if ChatGPT resembles human beings showing similar sensitivity to conversational context when processing scalar item "or". If ChatGPT has acquired the pragmatic knowledge similar to that of the humans, it should be more likely to interpret the expression "A or B" as "A or B but not both A and B" when it is part of the information focus compared with the case in which the expression "A or B" is part of the information background. To further explore the way ChatGPT processes scalar items, we replicated the second experiment in Zondervan (2010) using ChatGPT as the participant.

3.1 Design and stimuli

The experimental items of the study consisted of six short story pairs, each followed by a true-or-false question. All the stories ended with a conversation between two characters, in which one character used the scalar item "or" in his/her reply to another character's question (see 3 and 4). Each story in a pair differed in terms of the context where the scalar item occurred- whether the scalar item being part of the information focus or the information background. In the scalar-implicature-relevant (SI-relevant) condition (see 3), the question was about the object ("what" question), and the scalar item "or" was part of the information focus. In this case, the interpretation of the scalar item as either "A or B but not both A and B" or "A or B and possibly both A and B" had particular relevance to the conversation. In the scalar-implicature-irrelevant (SI-irrelevant) condition (see 4), the question is about the subject ("who" question), and the scalar item was part of the information background. Thus, the interpretation of the scalar item was not the major concern of the conversation. Crucially, based on the information provided in the story, the using of the scalar item "or" was logically sound but pragmatically infelicitous, and at the end of the story, ChatGPT was asked to judge if the character's answer was true or false. If the SI of "or" was computed, ChatGPT would respond with "false" to the question; or conversely, if the SI was not computed, a "yes" judgement would be given.

3. SI-relevant:

Julie and Karin were searching for marine animals on the beach. After some searching Julie found a crab. Not much later she also found a starfish. Unfortunately, Karin didn't find anything. When Karin returned, her mother asked what kind of marine animals Julie had found. Karin answered that Julie had found a crab or a starfish.

Is Karin's answer true or false?

4. SI-irrelevant:

Julie and Karin were searching for marine animals on the beach. After some searching Julie found a crab. Not much later she also found a starfish. Unfortunately, Karin didn't find anything. When they returned, their mother asked who had found a crab or a starfish. Karin answered that Julie had found a crab or a starfish.

Is Karin's answer true or false?

In Zondervan's original study (2010), the experimental items comprised six pairs of stories similar to (3) and (4) but written in Dutch. For the present study, we utilized the English versions of these stories as the experimental items. Additionally, we created 14 filler items that mirrored the length and structure of the experimental items. Each filler item contained a dialogue in which one character answered the question posed by the other character. Half of the filler items were designed to elicit a "true" response, while the other half were designed to elicit a "false" response. To balance the experimental conditions and the order of stimuli, we employed four pseudo-randomized lists of items, following Zondervan's original study.

3.2 Procedure

We followed the data collection procedure preregistered with the Open Science Framework (`https://osf.io/egm7v`), eliciting responses from ChatGPT (Feb 13 version). In each run of the experiment, we used a Python script to simulate a human interlocutor having a conversation with ChatGPT. At the start, the human interlocutor instructed ChatGPT to make truth-value judgements based on the content of the stories. Two practice trials were given to ChatGPT, the correct answer of which was "true" and "false" respectively. After the practice trial, ChatGPT was randomly assigned to one list of items, which were presented sequentially. For each item, ChatGPT was instructed to respond by saying only “true” or “false” without other words or explanations, and we recorded the responses from ChatGPT. In total, this study had 200 runs of the script, with 50 runs for each list of items.

3.3 Results and Discussion

In Zondervan (2010), the rate of “false” judgements (i.e., SIs) was 67% in the SI-relevant condition and 41% in the SI-irrelevant condition. In our experiment, ChatGPT responded with “true” for more than 99% of the experimental items, regardless of whether the item was in the SI-relevant or SI-irrelevant condition. The “true” judgement meant that ChatGPT judged the pragmatic infelicitous usage of “or” as “true”, which suggested a lack of pragmatic interpretation. Only one trial in the SI-relevant condition and two trials in the SI-irrelevant condition received a “false” judgement, which was typically interpreted as the computation of SIs (see Table 2). Given the large number of trials in the experiment, the difference between SI-relevant and SI-irrelevant condition regarding the rate of SI computation was not statistically meaningful (beta = -1.31, CI = [-10.81, 4.78]).

	“False”	**“True”**
Experimental items		
SI-relevant	1	599
SI-irrelevant	2	598
Filler items		
Correct Answer: False	1394	6
Correct Answer: True	96	1304

Table 2: A summary of judgements from ChatGPT for experimental items and filler items across different conditions in Exp2. Note, the column labels indicate the judgements provided by ChatGPT.

Our analysis of the filler items revealed that ChatGPT demonstrated sensitivity to the truth conditions of the statements (see Table 2). When the character in the story provided an untruthful response, and thus the correct answer to the question should have been “false”, ChatGPT provided more “false” judgments than “true” judgments (1394 vs. 6). Conversely, when the correct answer to the filler item was “true”, ChatGPT provided more “true” judgments than “false” judgments (1304 vs. 96). To further explore the impact of the correct answer on ChatGPT’s judgments, we modeled the probability of ChatGPT providing a “false” judgment as a function of whether the correct answer to the filler item was “true” or “false” (both dummy coded with the “false” answer being the reference level). Maximal random effects structures were constructed including subject and item intercepts and slopes. We found that when the correct answer of the filler item was “true”, the “false” judgements from ChatGPT decreased at a statistically meaningful rate (beta = -19.64, CI = [-33.92, -11.66]). In total, the accuracy rate of ChatGPT in answering the filler items was above 85 percent.

In this experiment, we investigated whether ChatGPT exhibited human-like patterns of scalar implicature computation by responding to the information structure of the communicative context. Previous research on human participants has shown that when the scalar item “or” was in the information focus, they were more likely to derive the upper bounded reading (“A or B but not both A and B”) compared to when the scalar item was in the information background. Our findings suggest that ChatGPT consistently provided “true” responses when asked if “A or B” is true when both A and B occur, indicating that it interpreted the scalar item “or” as lower bounded (“A or B and possibly both A and B”) for over 99% of the trials, regardless of whether it appeared in the information focus or background. Furthermore, ChatGPT did not always provide “true” responses. For filler items where the correct answer was “false”, ChatGPT provided significantly more “false” responses than “true” responses, and its accuracy rate was high. Therefore, the reason why ChatGPT almost always provided a “true” response for experimental items was that it always endorsed the pure logical interpretation rather than the pragmatic interpretation of the scalar item “or”. The lack of scalar implicature computation for this scalar item and the insensitivity to the information structure of the communicative context differentiate ChatGPT from human participants.

4 Experiment 3

For human participants, the computation of SI is modulated by the conversational context, and the result of Experiment 2 suggested that ChatGPT lacked the sensitivity to the manipulation of information structure, an important aspect of the conversational context. This experiment aimed to investigate whether conversational context affects how ChatGPT processes scalar implicature (SI) using a different contextual aspect and a different scalar item. Bonnefon, Feeney, and Villejoubert (2009) found that the rate of endorsing SIs for the scalar item "some" decreased when the lower bounded interpretation ("some and possibly all") threatened the face of the listener, compared to when it boosted the listener's face. In this experiment, we aimed to test whether ChatGPT shows similar sensitivity to conversational context. We adopted the same design as the first study in Bonnefon, Feeney, and Villejoubert (2009), comparing the rate of SI computation across two within-participants conditions. Unlike the original study, we did not recruit human participants but tested whether ChatGPT exhibits similar performance as human participants. Specifically, we examined whether ChatGPT is more likely to interpret the scalar item "some" as "some but not all" in the face-boosting context, but not so much when the scalar item "some" appears in the face-threatening context.

4.1 Design and stimuli

In this experiment, ChatGPT read two scenarios which were either face-threatening or face-boosting, and the scalar item "some" appeared in the description of the scenario. After reading each scenario, ChatGPT was required to answer a yes-no question. Specifically, we asked ChatGPT whether it would endorse the lower-bounded interpretation of some (which is "some and possibly all"). An example of the experimental item in the face-threatening and face-boosting context was shown in (5) and (6):

> 5. Face-threatening context:
>
> Imagine that you have joined a poetry club, which consists of five members in addition to you. Each week, one member writes a poem, and the five other members discuss the poem in the absence of its author. This week, it is your turn to write a poem and to let others discuss it. After the discussion, one fellow member confides to you that "Some people hated your poem."
>
> Yes/No question: From what this fellow member told you, do you think it is possible that everyone hated your poem?
>
> 6. Face-boosting context:
>
> Imagine that you have joined a poetry club, which consists of five members in addition to you. Each week, one member writes a poem, and the five other members discuss the poem in the absence of its author. This week, it is your turn to write a poem and to let others discuss it. After the discussion, one fellow member confides to you that "Some people loved your poem."
>
> Yes/No question: From what this fellow member told you, do you think it is possible that everyone loved your poem?

We included two scenarios like 5 and 6, creating two lists of items using the Latin Squared Design. All items in the experiment were directly adopted from Bonnefon, Feeney and Villejouber (2009).

4.2 Procedure

We followed the data collection procedure preregistered with the Open Science Framework (`https://osf.io/3v9gn`), eliciting responses from ChatGPT (Feb 13 version). In each run of the experiment, we used a Python script to simulate a human interlocutor having a conversation with ChatGPT. At the start, the human interlocutor instructed ChatGPT to answer yes-no questions based on the description of scenarios. Two practice trials were given to ChatGPT, the correct answer of which was "yes" and "no" respectively. After that, ChatGPT was randomly assigned to one list of items, which were presented to ChatGPT in a random order. For each item, ChatGPT was instructed to respond by saying only "yes" or "no" without other words or explanations, and we recorded the responses from ChatGPT. In total, this study had 200 runs of the script, with 100 runs for each list of items.

	"No"	**"Yes"**
Face-boosting	198	0
Face-threatening	198	0

Table 3: A summary of judgements from ChatGPT for experimental items across different conditions in Exp3.

4.3 Results and Discussion

According to our preregistered data exclusion criteria, we excluded data from two runs of the experiment because ChatGPT answered the second practice trial incorrectly, indicating that it may not provide reliable judgments in that run of the experiment. Therefore, we analyzed the data from 198 runs of the experiment. In Bonnefon, Feeney and Villejoubert's (2009) study, 83% of human participants responded with "no" when asked if the lower bounded interpretation of "some" was possible in the face-boosting context, while a significantly lower 58% responded "no" in the face-threatening context. In contrast, our study found that ChatGPT always responded "no" to all of the trials, regardless of whether the context was face-boosting or face-threatening (see Table 3).

Though the exact mechanism is still unclear regarding why human participants were more likely to interpret the construction "some verb-ed X" as "some and possibly all verb-ed X" in the face threatening context than in the face boosting context, Bonnefon, Feeney and Villejoubert (2009) suggested that the listener may take into account the intension of the speaker to use the word "some" in an underinformative way in order to protect the face of the listener. Although, the SI rate of "some" decreased in the face threatening condition, in general, human participants preferred the pragmatic interpretation of "some" as "some but not all", and that is why even in the face-threatening condition, the majority of the human participants (58%) provided a "no" judgement to the question "Do you think it is possible that everyone hated…" In our experiment with ChatGPT, we clearly saw a stronger preference for the pragmatic interpretation of "some" over the truth-conditional interpretation. In fact, ChatGPT exhibited zero variance in its judgements- for all the trials that contained the scalar item "some", ChatGPT always interpreted them as "some but not all", and thus said "no" to the question, regardless of whether the implicature was face threatening or face boosting to the listener.

5 General Discussion and Conclusion

In three experiments, we investigated whether LLMs like ChatGPT exhibit human-like performance when processing pragmatic implicatures. Previous research has shown that humans distinguish implicatures from the truth-conditional meaning of the utterance, and several factors have been identified that modulate the probability of implicature computation. While pragmatic enrichment is an essential component of successful communication, whether an implicature is computed by a specific listener in a specific communicative context is probabilistic in nature. In contrast, our findings revealed that ChatGPT lacked human-like flexibility in switching between pragmatic and semantic interpretation, as it was unable to inhibit the computation of GCIs even when instructed to do so. Notably, the processing of scalar items in ChatGPT exhibited a deterministic pattern: whereas "some" always received an upper bounded interpretation as "some but not all", the expression "A or B" almost always received a lower bounded interpretation as "A or B and possibly both A and B".

Given ChatGPT's impressive human-like performance across a range of language tasks (Cai et al., 2023), one might question why humans and LLMs differ in their computation of GCIs. Our argument is that this difference can be explained by the acquisition of GCIs and the computational resources available to humans and machines. Developmental research indicates that scalar items are acquired with a lower bounded interpretation before pragmatic enrichments (Noveck, 2001). Consequently, adults have access to both the literal and pragmatic interpretations of a scalar item, whereas LLMs are exposed to language data that are mainly pragmatically driven. This explains why ChatGPT, in general, is more prone to pragmatic interpretation compared with human participants. However, it is still unclear why some specific word like "or" almost always evokes a literal rather than pragmatic interpretation. Furthermore, humans possess limited computational resources compared to machines. The principle of economy suggests that the human mind enriches the truth-conditional meaning only when the context necessitates it (Noveck & Sperber, 2007). This echoes the fact that the effect of contextual manipulation has only been observed among human participants rather than LLMs. It is consistent with the observation that humans tend to use shorter forms of words (e.g., math instead of mathematics) when the meaning is predictable, while ChatGPT does not (Cai et al., 2023). Overall, our experiments demonstrate that although LLM-based chatbots such as ChatGPT excel in many language tasks, they do not mimic humans in their computation of GCIs.

Limitations

The scope of our research is limited to uncovering the distinction between humans and LLMs in a specific aspect of pragmatic processing: the computation of GCIs. While we offer tentative explanations for the patterns we observed, our study does not directly provide solutions for improving the performance of LLMs. In this study, we use ChatGPT as an example of LLMs due to its prominence in current research. However, it remains uncertain whether other LLMs exhibit comparable characteristics and tendencies as observed in ChatGPT. Moreover, it is important to note that our findings may not generalize to the processing of other types of pragmatic implicatures.

References

Gerry Altmann, and Mark Steedman.1988. Interaction with context during human sentence processing. *Cognition* 30, no. 3: 191-238.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, no. 6: e2218523120.

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive psychology* 18, no. 3: 355-387.

Jean-François Bonnefon, Aidan Feeney, and Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition* 112, no. 2: 249-258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan 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. *HAL Open Science*. https://hal.science/hal-03991530/

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.

Kimberly Wright Cassidy, Michael H. Kelly, and Lee'at J. Sharoni. 1999. Inferring gender from name phonology. *Journal of Experimental Psychology: General* 128, no. 3 (1999): 362.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of ChatGPT. *The New York Times*.https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Ryan Doran, Gregory Ward, Meredith Larson, Yaron McNabb, and Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*: 124-154.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110, no. 20: 8051-8056.

Herbert Paul Grice. 1975. Logic and conversation. *In Speech acts*, pp. 41-58. Brill.

Herbert Paul Grice. 1978. Further notes on logic and conversation. In *Pragmatics*, pp. 113-127. Brill

Laurence Robert Horn. 1972. *On the semantic properties of logical operators in English*. University of California, Los Angeles.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint* arXiv:2301.08745.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint* arXiv:2302.02083.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126, no. 2 (2013): 313-318.

Ira A Noveck. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 78, no. 2: 165-188.

Ira A Noveck and Dan Sperber. 2007. The why and how of experimental pragmatics: the case of 'scalar inferences, in *Advances in Pragmatics*, ed N. Burton-Roberts (Basingstoke: Palgrave), 184–212.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in ChatGPT. *arXiv preprint* arXiv:2302.06426

Steven T Piantadosia. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint*, lingbuzz/007180

Jennifer M. Rodd, Belen Lopez Cutrin, Hannah Kirsch, Alessandra Millar, and Matthew H. Davis. 2013. Long-term priming of the meanings of ambiguous words. *Journal of Memory and Language* 68, no. 2 (2013): 180-198.

Chris Westbury. 2005. Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and language* 93, no. 1 (2005): 10-19.

Arjen Zondervan. 2010. *Scalar implicatures or focus: an experimental approach*. Netherlands Graduate School of Linguistics.

A Appendices

An example of experimental items containing GCIs of different categories in Exp1.

Dialogue	Fact	First Level Category	Second Level Category
Irene: Hey, Sam. Do you know who wrote Pride and Prejudice? Sam: A British woman wrote it, and her last name was Austen.	FACT: Jane Austen, a British woman, wrote Pride and Prejudice.	Training Example	Training Example
Irene: How much cake did Gus eat at his sister's birthday party? Sam: He ate most of the cake.	FACT: By himself, Gus ate his sister's entire birthday cake.	Q_Based_GCIs	Quantifiers_Modals
Irene: How many children does Lisa have? Sam: Lisa has three children.	FACT: Lisa has quadruplets	Q_Based_GCIs	Cardinals
Irene: How would you say you're doing financially? Sam: I'm comfortable.	FACT: Sam just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.	Q_Based_GCIs	Gradable_Adjectives
Irene: What kind of milk does your diet allow for? Sam: It allows for 1%.	FACT: The only type of milk prohibited by Sam's diet is full-fat milk.	Q_Based_GCIs	Rankings
Irene: I heard something big happened in the art studio yesterday. Sam: In a fit of rage, Rachel picked up a hammer and broke a statue.	FACT: After grabbing a hammer, Rachel angrily kicked a statue, causing it to fall over and break.	I_Based_GCIs	Argument_Saturation
Irene: What happened when Sue came over? Sam: She walked into the bathroom. The window was open.	FACT: The open windows are in the kitchen, and there are no windows in the bathroom.	I_Based_GCIs	Bridging_Inferences
Irene: Can the guys come to the reception? Sam: George and Steve play squash at the gym until 6:00 every day.	FACT: George plays squash at the YMCA until 6:00 daily, and Steve plays squash at SPAC until 6:00 every day.	I_Based_GCIs	Coactivities
Irene: I understand that George has had a really rough year. Sam: Last month, he lost his job and started drinking.	FACT: George started drinking on the 15th of last month and lost his job on the 20th of last month.	I_Based_GCIs	Conjunction_Buttressing
Irene: Why is Stephen so upset? Sam: He caused Bill to die.	FACT: Stephen intentionally murdered Bill.	M_Based_GCIs	Verbal_Periphrasis
Irene: What happened at Doctor Witherspoon's office? Sam: Sasha waited and waited for her appointment.	FACT: Sasha waited 5 minutes for her appointment at DoctorWitherspoon's office.	M_Based_GCIs	Repeated_Verb_Conjuncts
Irene: What did Joseph do after finishing the marathon? Sam: He drank bottles and bottles of water.	FACT: Joseph drank one 20 oz bottle and one 16 oz bottle of water after finishing themarathon.	M_Based_GCIs	Repeated_Noun_Conjuncts