Title

Capable but not cooperative? Perceptions of ChatGPT as a pragmatic speaker

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Mayn, Alexandra
Loy, Jia
Demberg, Vera

Publication Date

2024

Peer reviewed

Capable but not cooperative? Perceptions of ChatGPT as a pragmatic speaker

Alexandra Mayn[1], **Jia E. Loy**[1] and **Vera Demberg**[1,2]
{amayn, jialoy, vera}@coli.uni-saarland.de

[1]Department of Language Science and Technology
[2]Department of Computer Science
Saarland University, 66123 Saarbrücken, Germany

Abstract

Pragmatic implicature derivation presupposes that the cooperative principle is observed and critically depends on interlocutors expecting each other to behave cooperatively. It is much less clear, however, whether people extend this assumption to communication with artificial agents. People might therefore not draw the same pragmatic inferences when interacting with an artificial agent as they would with other conversationally competent humans, even if the agent is in principle believed to be similarly competent. In our study, we ask participants to interpret messages in a pragmatic reference game which they are told were generated by ChatGPT. Additionally, participants report whether they believe ChatGPT to be capable of the reasoning needed to select the optimal message. We observe a noteworthy discrepancy: in the reference game, participants interpret ChatGPT's messages less pragmatically than those of another adult human, but in the post-test questionnaire, they overwhelmingly rate ChatGPT's pragmatic ability very highly.

Keywords: partner effects; human-computer interaction; pragmatics

Introduction

Communication is often expected to involve collaboration. The principle of least collaborative effort (Clark & Wilkes-Gibbs, 1986) states that interlocutors share the responsibility of ensuring mutual understanding by accommodating to their partner's perspective. Adjusting the presentation format of the information depending on the addressee is known as audience design. Studies show that people are willing to invest more effort when they have a reason to believe that their interlocutor's communicative ability is limited, for instance, by lack of experience or context (Fussell & Krauss, 1989; Tippenhauer, Fourakis, Watson, & Lew-Williams, 2020). It has also been shown that listeners, not only speakers, also sometimes construct a mental model of the speaker and adjust their interpretation of the speaker's utterances based on speaker characteristics, such as adjusting expectations of precision based on the speaker's persona (Beltrama & Schwarz, 2021) and inferring trustworthiness of the speaker based on their perceived language proficiency (Ip & Papafragou, 2021).

With rapid developments in technology and with human-computer interaction becoming more and more commonplace, it is of interest to which extent these collaborative effects also extend to artificial agents. The "Computers as Social Agents" (CASA) framework (Nass, Steuer, & Tauber, 1994) developed in the 1990s states that human-computer interaction is social and that humans readily exhibit social behavior when interacting with artificial agents, such as politeness and personality attribution. The question whether, when interacting with a computer, people tend to take its perspective into account to a lesser or a greater degree than when interacting with another human, has been investigated in the the domain of spatial perspective taking, where instructions may be ambiguous when the perspectives of the speaker and the listener do not match (e.g., "Give me the book on the right" could mean the speaker's or the listener's right). Interestingly, while some studies found that participants were more egocentric in their interpretation when the speaker was another human and took the robot's perspective more willingly (Fischer, 2007; Duran, Dale, & Kreuz, 2011), other studies found the opposite, that is, participants expecting robots to adopt their spatial perspective (Carlson, Skubic, Miller, Huo, & Alexenko, 2014; Loy & Demberg, 2023). The perceived competence of the artificial agent also appears to matter. Fischer (2005) found that participants always took the perspective of a nonverbal robot and gave it streamlined instructions, whereas with a verbal robot they used a more varied vocabulary and syntax and included some egocentric descriptions. Similarly, Branigan, Pickering, Pearson, McLean, and Brown (2011) investigated lexical alignment in dialogue with humans and artificial agents and found that their participants tended to align (i.e. repeat the interlocutor's choice of words) more with a computer than with a human, and with an older computer more than with a more modern computer, presumably reflecting participants' beliefs about computers' limited capability.

More recently, Loy and Demberg (2023) found that participants provided more egocentric descriptions in a spatial perspective taking task with a computer than with a human partner and more egocentric descriptions with a modern compared to an older computer. They explain the discrepancy between their findings and earlier studies, which had mostly found that people were more willing to take the perspective of a computer than that of another human, in terms of a shift in perception of artificial agents' competence: whereas a couple of decades ago, artificial agents were perceived as having very limited capabilities, they are now perceived as much more complex agents with high intelligence and more collaborative capacity. Similarly, in a communication game where participants' task was to get their partner to pick out objects from a grid by referring to them, Peña et al. (2023) found that

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

participants were more likely to consider their partner's visual perspective when their partner was a human than when it was a computer. However, this difference disappeared when it was emphasized to participants that the computer was a collaborative agent with a shared communicative goal and separate from the system which the experiment was run on. This suggests that not only the artificial agent's capability but also their willingness to collaborate may influence how they are perceived and treated by humans.

In this study, we add to this body of work by investigating people's perception of the influential LLM-based chatbot ChatGPT as a cooperative agent in the domain of Pragmatics. Derivation of pragmatic implicatures presupposes that the interlocutors are behaving cooperatively and observing conversational maxims (known as the *cooperative principle*; Grice (1975)). In the interaction between a human and an artificial agent, it is much less clear whether people assume that the cooperative principle holds. People might accordingly not draw the same pragmatic inferences when interacting with an artificial agent which they would draw in an interaction with other (conversationally competent) humans, even if the agent is in principle believed to be similarly competent.

While surveys have been conducted examining people's attitudes towards ChatGPT (Singh, Tayarani-Najaran, & Yaqoob, 2023; Shoufan, 2023; Ngo, 2023), we are not aware of other studies investigating how people communicate with ChatGPT and whether they perceive it as a cooperative conversation partner. To our knowledge, this is the first study to examine people's perception of ChatGPT as a collaborative agent and to investigate the perceived pragmatic abilities of a modern artificial agent.

We use the reference game paradigm from Mayn and Demberg (2024) (henceforth M&D), where participants interpret ambiguous messages sent by the speaker. M&D manipulated the identity of the speaker between participants and found that if participants were told that the speaker was another adult, they were more likely to interpret an ambiguous message as an implicature, whereas if they were told was a 4-year-old child, they were more likely to interpret the message literally. In this study, we tell participants that the messages were sent by ChatGPT. We compare our results to the two speaker conditions in M&D and find that, at the population level, participants interpret ChatGPT's messages less pragmatically than those of an adult but somewhat more pragmatically than those of a 4-year-old child.[1]

We observe a notable discrepancy: participants' performance on the reference game does not match their explicit estimates of ChatGPT's pragmatic ability. When asked explicitly in a post-test questionnaire, participants overwhelmingly expressed high confidence in ChatGPT's ability to solve the task, including those participants who interpreted ChatGPT's messages literally in the pragmatic task itself. We discuss possible reasons for this discrepancy as well as the implications of our findings for our understanding of human perception of modern artificial agents as communicative partners.

[1]Preregistration of the experiment can be found at https://osf.io/7ycrq and data and analysis scripts can be found at https://github.com/sashamayn/perceptions_of_chatgpt_cogsci2024/

Experiment

Participants

40 native speakers of English with an approval rating of at least 95% were recruited via the crowdsourcing platform Prolific. 4 participants needed to be excluded because their performance strongly suggested that they were responding randomly or because their post-hoc explanation of their reasoning strategy suggested that they had misunderstood the setup of the experiment. New participants were recruited in their place, resulting in a total of 40 participants.

Procedure

We used the paradigm from M&D, with the modification that we told participants that the messages were generated by ChatGPT. Since not all participants may be familiar with ChatGPT, they were first given a short description of what ChatGPT is and an example of a prompt and ChatGPT's response unrelated to the topic of the experiment ("What desserts should I try in Paris?").

To make the cover story as convincing as possible, participants were told that since ChatGPT 3.5 only accepts textual input, it completed the speaker task in textual form which is completely equivalent to the visual form. They were shown the speaker task in the visual form side by side with the task in textual form. The textual form was as follows:

> Your task is to send a message so that someone, let's call them the interpreter, is able to identify a particular object from the set of three objects.
>
> You are only allowed to send one of four messages, and the interpreter knows this: the color red, the color green, circle and triangle. You can only send one message.
>
> The objects the interpreter will see: red square, green circle, blue triangle. You need the interpreter to pick out: blue triangle. Which of the four available messages (the color red, the color green, circle or triangle) will you send?

Since the visual scene in this task is very simple, the textual and visual instructions are equivalent and therefore this cover story is unlikely to have raised participants' suspicion or significantly influenced their behavior.

Reference game On each trial, participants saw a screen with 3 objects and a message which they were told had been sent to them by ChatGPT, who was playing the role of the speaker, to refer to one of the 3 objects. Participants' task was to decide, for each of the three objects, how likely it is that ChatGPT was referring to that object by distributing 100 points between the three objects using sliders. An example trial is shown in Figure 1.

Each of the three objects on the screen is composed of two attributes, shape (square, triangle or circle) and color (blue, red or green). The message that participants received was a shape or a color. Additionally, they were told that the set of

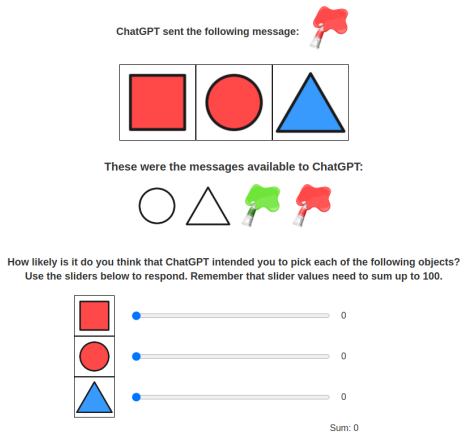

Figure 1: Example of a critical trial.

messages that ChatGPT was allowed to choose from was limited: there was no message "square" and no message "blue".

The experiment consisted of 24 trials, of which 8 were critical and 16 were control trials. Trial order was randomized for every participant. On each trial, the three objects in the display were target, competitor and distractor. The order of the three objects on the display was randomized for every trial.

On critical trials, the message was ambiguous. In the trial shown in Figure 1, the message (red) is literally compatible with two of the three objects, the square and the circle. However, this ambiguity may be resolved by reasoning that if the speaker had wanted to refer to the red circle, they could have sent the unambiguous message "circle". Since the speaker did not do so, the listener may reason that they were referring to the red square, for which there is no alternative message.

M&D argue that whether this implicature is drawn is modulated by perceived reasoning ability of the speaker. In their study, in the child speaker condition, participants were less likely to draw the implicature and assign a high probability to the target (red square) and were instead more likely interpret the message literally and assign roughly equal probabilities to the red square (target) and the red circle (competitor). Also, of course, a participant themselves may fail to reason about alternatives, thus interpreting the message literally.

Of the 16 control trials, 4 were completely ambiguous (two of the three potential referents were identical and the message could refer to either of them) and 4 were completely unambiguous (all features of the three objects were unique). The remaining 8 control trials had the same displays as the critical trials but the target was the competitor (4) or the distractor from the corresponding critical item (4).

After completing the 24 trials, each participant saw the first critical trial that they had seen again and, once they had made their response, a textbox popped up with the question "Why did you decide to put the sliders in those positions?". Participants' responses to this question were annotated as described below, and the distribution of responses was compared to that in the adult and child speaker conditions from M&D.

At the start of the experiment, participants completed 3 trials (2 unambiguous and 1 ambiguous) from the speaker's perspective. The object that ChatGPT needed to refer to was highlighted in yellow. The set of the four available messages was also shown on each speaker trial. This was done to make sure that participants understood ChatGPT's task and the fact that not all messages were available to it.

Annotation of participants' responses We annotated participants' responses to the question "Why did you decide to put the sliders in those positions?" using the annotation scheme from M&D.

If the explanation indicated an inference based on the fact that an alternative unambiguous message wasn't used, it was labeled *correct_reasoning*. Responses indicating guessing between the target and the competitor were labeled *guess*. Responses which explicitly debated whether the speaker possessed sufficient reasoning ability to select an optimal message were assigned the tag *meta_reasoning*. Explanations indicating inverse reasoning, e.g. taking the message "red" to mean the only *non-red* object, were labeled *odd_one_out*. Explanations indicating that the response was based on the participant's preference for a certain shape or color were labeled *preference*. Unclear responses were labeled "unclear".

Post-test questionnaire We were interested in how participants' responding related to their experience with ChatGPT and their beliefs about its competence. Therefore, we included a post-test questionnaire to tap into these questions.

Participants were asked how much experience using ChatGPT (on a scale from 1 to 5) and how much knowledge of Machine Learning and Natural Language Processing (also 1-5) they had. They were then shown an item from the speaker's perspective (Figure 4) and asked to explain in their own words, by referring to that example, how they think ChatGPT determines which message to send. Finally, once they had provided an answer, we explained to the participants that in order to solve this task, the speaker needed to reason about alternatives and that the unambiguous message "color green" is more optimal than "triangle" since upon hearing "triangle", the listener may select the blue triangle. We then asked the participants how likely they thought it was, on a 5-point scale, that ChatGPT is capable of performing that kind of reasoning.

Results

Throughout this section, we plot our results alongside those of M&D for ease of comparison.

We first look at the average target ratings per trial type (unambiguous, critical and ambiguous), shown in Figure 2. We see that, similarly to the child and adult speaker conditions from M&D, in our experiment, participants performed at ceiling in the unambiguous control condition. This shows that participants understood the task and consider ChatGPT to be at least capable of feature matching. In the ambiguous control condition, the target ratings are at chance, as predicted. That is because of how the ambiguous condition is scored:

Table 1: Experiment results

	beta	SE	t	p
Intercept	80.06	1.38	57.82	**<0.0001**
speaker (adult vs. ChatGPT)	5.37	1.84	2.91	**0.004**
speaker (child vs. ChatGPT)	-1.67	1.84	-0.91	0.37
trial type (critical vs. control)	-18.82	0.46	-40.52	**<0.0001**
trial id	0.06	0.05	1.22	0.22
message type (shape vs. color)	-0.78	0.53	-1.49	0.14
target position (left vs. middle)	-0.03	0.66	-0.05	0.96
target position (right vs. middle)	-1.13	0.66	-1.70	0.09
speaker (adult vs. ChatGPT) : trial type	4.91	0.66	7.49	**<0.0001**
speaker (child vs. ChatGPT) : trial type	-2.01	0.66	-3.05	**0.002**

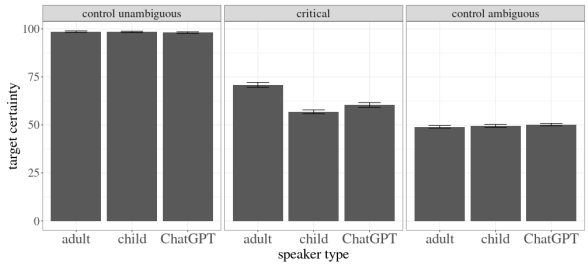

Figure 2: Average target rating ($\pm SE$) per trial type for each speaker condition (child and adult speaker from M&D and ChatGPT speaker from the current study).

on ambiguous trials, two of the three objects are identical and it is decided via coin flip which of the two identical objects is target. Distributing 100 points equally between the two identical objects results in an average target rating of around 50%. Thus the ambiguous condition constitutes a chance baseline.

In the critical condition, the average probability assigned to the target in the ChatGPT speaker condition is 60.36 (*SD*=17.87), which is lower than in the adult speaker condition from M&D (70.8, *SD*=21.01) but slightly higher than in the child speaker condition from M&D (56.81, *SD*=13.50).

In order to verify the significance of these apparent differences between the speaker conditions, we pool our data with M&D's data and fit a linear mixed-effects model. For this analysis, we remove the unambiguous control condition from analysis. We regress the target probability on each trial onto the speaker identity (ChatGPT, child or adult; dummy-coded with ChatGPT as reference level since we want to compare ChatGPT to the two speaker conditions from M&D), item type (sum-coded, levels: -1 = unambiguous, 1 = critical), interaction between speaker and item type, target position (left, middle or right; dummy-coded with middle as reference), mean-centered trial number, and message type (color or shape; dummy-coded with shape as reference). The maximal random effect structure allowing the model to converge was included, which consisted of per-participant random intercept and a random slope for trial number.

The results are reported in Table 1. Much lower target ratings were assigned in the critical condition compared to the unambiguous control (β =-18.82, p<0.0001). There was a main effect of speaker (adult vs. ChatGPT), whereby the target ratings were higher in M&D's adult speaker condition compared to our ChatGPT condition regardless of trial type (β =5.37, p=0.004). There was no main effect of the difference between a child speaker and ChatGPT speaker. There were also significant interactions of speaker and trial type, indicating that the difference between target ratings between the adult and the ChatGPT speaker was stronger in the critical condition than in the control condition (β =4.91, p<0.0001) and that the target ratings for critical items were lower in the child speaker condition than in the ChatGPT speaker condition (β =-2.01, p=0.002). This suggests that, at the population level, ChatGPT's perceived reasoning ability or cooperativity is inferior to that of an adult speaker's but perhaps somewhat superior to that of a 4-year-old child.

Next, we examine the annotations of participants' explanations. The top panel of Figure 3 shows the distribution of annotation labels for the three speaker conditions. The bottom panel shows the corresponding average target ratings. Similarly to the two conditions from M&D, the majority of our participants' explanations falls either into the *correct_reasoning* or the *guess* category. The number of *correct_reasoning* responses in the ChatGPT condition falls in between the adult and the child speaker conditions (ChatGPT: 11/40, adult: 17/40, child: 6/40). The number of *guess* responses in the ChatGPT speaker condition also falls in between the other two speaker conditions (ChatGPT: 20/40, adult: 16/40, child: 23/40). There are also 3 *meta_reasoning* responses where the explanations explicitly express doubts about how likely ChatGPT is to be capable of reasoning necessary to select the optimal message. The average target ratings corresponding to the *correct_reasoning* and *meta_reasoning* tags are also lower than in the adult condition but higher than in the child condition from M&D. Notably, the average ratings are consistent with the reported strategies: the average ratings in the *guess* category are at chance, as we would expect, whereas for *correct_reasoning*

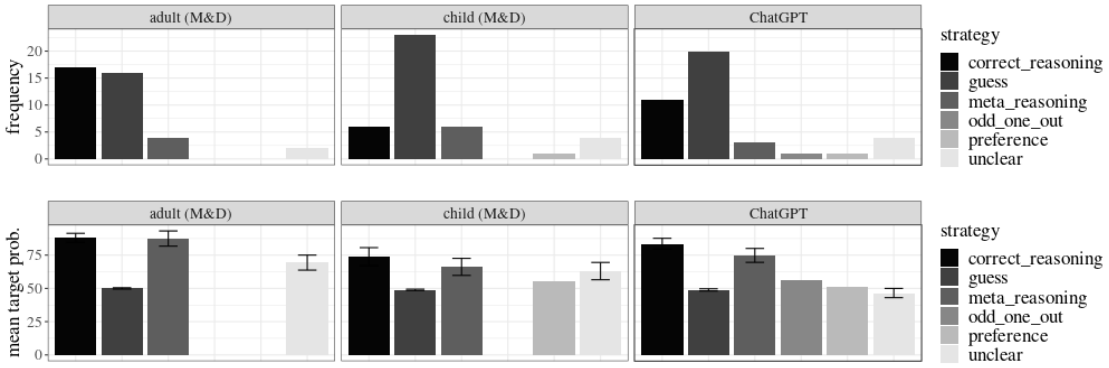

Figure 3: Frequency of each annotation tag (top panel) and corresponding average target ratings (bottom panel) for the three speaker conditions: child and adult from M&D and ChatGPT from the current study.

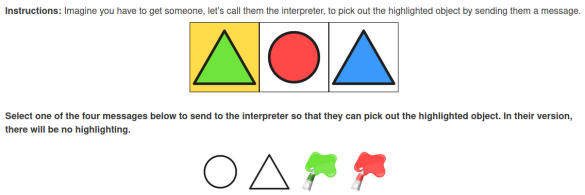

Figure 4: Item from the speaker's (i.e. ChatGPT's) perspective shown to participants in the post-test questionnaire.

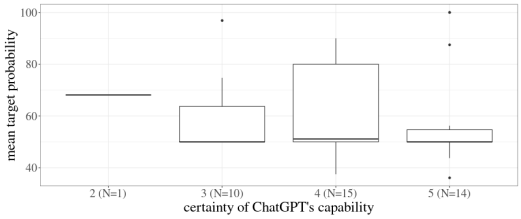

Figure 5: Average target ratings for each response to the question "How capable do you think ChatGPT is of performing this task?"

and *meta_reasoning* they are much higher but not at ceiling, especially in the child and ChatGPT conditions, reflecting population-level uncertainty that those speakers would behave optimally or collaboratively. The results of both this supplementary analysis of annotations and the main regression analysis suggest that, at the population level, utterances produced by ChatGPT as interpreted less pragmatically than those of an adult and more so than those of a child. However, this is a collective interpretation: as we will see, there is a lot of individual variability in the responses.

We now turn to participants' responses to the post-test questionnaire. Our participants had limited experience with ChatGPT (*mean*=2.08, *SD*=1.10) and almost no knowledge of machine learning and NLP (*mean*=1.33, *SD*=0.53).

The speaker item that was shown to the participants in the post-test questionnaire (Figure 4) doesn't exactly correspond to the critical items from the listener's perspective. If it had, the highlighted object would have been the blue triangle. However, in that case, from the speaker's perspective the task is near-trivial since the only way to refer to the blue triangle is the "triangle" message. Therefore, in this question, the highlighted object is the green triangle (which corresponds to the competitor on critical listener trials). There are two ways to refer to the highlighted object, the color green and the triangle, and one needs to reason that the unambiguous message (the color green) is preferable. Interestingly, more than half of the participants (25 out of 40) in their open-ended answers to the question how ChatGPT likely decides what message to send described a process of elimination by which the unambiguous message is preferred to the ambiguous message. This is likely a lower bound since some explanations indicated high perceived ChatGPT competence without specifying the strategy, e.g. "ChatGPT is a master at this".

Likewise, when explicitly asked the likelihood that ChatGPT is capable of reasoning about alternatives needed to select an optimal message, participants gave high ratings (*mean*=4.05 out of 5, *SD*=0.85). Notably, there appears to be no relationship between people's performance on the task itself and their report of ChatGPT's competence on this question. Figure 5 shows average target ratings in the critical condition corresponding to each of the Likert scale responses. It can be seen that the majority of participants who rated ChatGPT's capability of solving the speaker task at a 5 in the post-test questionnaire gave very low target ratings in the actual task. This is an interesting discrepancy which will be discussed in detail in Discussion.

Discussion

In this study, we investigated the perceived pragmatic ability of ChatGPT using the reference game paradigm (Frank & Goodman, 2012; Franke & Degen, 2016), which has been shown to be sensitive to perceived reasoning ability of the speaker. Participants interpreted messages, which they were told had been generated by ChatGPT, and then completed a

post-test questionnaire. We compared participants' performance to the child and adult speaker experiments from Mayn and Demberg (2024) in the same paradigm.

We found that the ratings that participants assigned to the target, which are a proxy for the perceived pragmatic sophistication of the speaker, are consistently and significantly lower than the ratings that were assigned in the adult speaker condition in M&D. Indeed, the responses in our study pattern more closely with the child speaker condition from M&D. At first glance, this seems to suggest that participants perceive ChatGPT to be less pragmatically capable than an adult human. This finding seemingly at odds with the results obtained by Loy and Demberg (2023), who found that participants expected a computer to take their perspective more often than a human, and a modern computer to take their perspective more often than an older computer. Our findings are, in fact, more in line with earlier work which had found that people more willingly took the computer's perspective than another human's (Duran et al., 2011; Branigan et al., 2011).

The reason for the apparent discrepancy between our findings and those of Loy and Demberg (2023) may lie in the difference between the two tasks. In spatial perspective taking, which Loy and Demberg (2023)'s study investigated, the two perspectives are arguably quite salient, and we would expect any adult participant to be aware that there are two perspectives, even if adopting one of them is effortful. The pragmatic perspective taking task in the current study, on the other hand, is arguably much more difficult. Even in the adult condition in M&D, some participants perform at chance on the critical items (Figure 6), suggesting that they were just interpreting the messages literally themselves and did not consider the speaker's perspective at all. Indeed, Franke and Degen (2016) show in a similar paradigm that a subgroup of their participants exhibits behavior that is most consistent with that of the simplest pragmatic listener model, a literal listener. Therefore, it may be less effortful to consider how an artificial agent would behave in a spatial perspective taking task than in the pragmatic reference game. Moreover, participants may assume that perspective taking is built into the computer program that is performing the spatial perspective taking task and that it had been specifically trained to do it, conceivably in order to be accommodating to humans. In the current study, on the other hand, ChatGPT would have needed to perform the task ad hoc, making it more likely that participants would not assume that ChatGPT would behave pragmatically by default.

This brings us to the other interesting finding of the current study: the mismatch in participants' behavior on the task and their explicit ratings of ChatGPT's competence in the self-report questionnaire. When asked explicitly, participants overwhelmingly gave very high confidence ratings of ChatGPT's ability to do the reasoning needed to send the optimal message, including those participants who had just given chance-level responses on the reference game. One factor that likely contributed to this effect is that some participants initially could not solve the task pragmatically themselves,

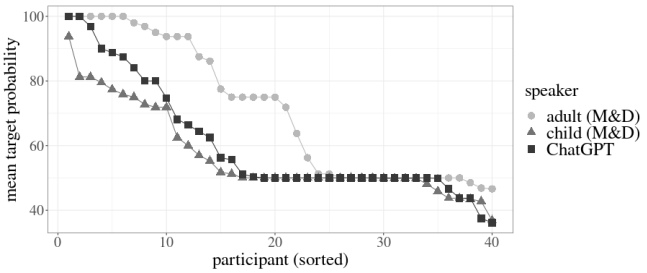

Figure 6: Average target rating per participant (sorted) for the three speaker conditions.

corresponding to chance-level performance in Figure 6, but then, when it was explained to them in the post-test questionnaire, they were forced to think about ChatGPT's ability to perform the task and rated it as high. Another explanation for this mismatch could lie in the distinction between participants believing that ChatGPT is *capable* of performing optimally and them believing that it would actually do so and select a message that is maximally helpful to the listener. M&D hypothesize that the difference between the child and adult speaker conditions that they observe could have to do with the communicative context being more or less conducive to deeper reasoning: in the adult speaker case, it could be that participants are more likely to assume intentionality of the speaker's actions and therefore be more motivated to figure out the intended meaning. The same may be applicable to the distinction between ChatGPT and an adult human speaker. In the case of a human speaker, there is a much more direct parallel to self and one's own reasoning than when ChatGPT is the speaker. Since we know that the task is nontrivial and potentially effortful, sufficient motivation may be needed to engage in reasoning about the speaker's intention, which may be more readily available in the case of a fellow human interlocutor. This explanation is in line with findings of Peña et al. (2023), who showed that participants initially behaved more egocentrically with computers than with humans in a referring task but this difference disappeared when the computer was explicitly framed as a collaborative agent with a shared goal. It could be that in this task, since participants were merely told that ChatGPT generated the messages, the interaction was not direct enough to assume cooperative intention. Future work could investigate whether utterances by ChatGPT are interpreted more pragmatically when the interaction is framed as more direct and collaborative, for instance, when ChatGPT is said to generate messages in real time or when ChatGPT additionally "addresses" the participant directly.

Finally, while we asked participants about ChatGPT's capability on this specific task and about their experience with AI and NLP, we did not elicit their opinions of AI more generally. Future work could include additional questions targeting people's opinions of AI and investigate how those opinions influence their interpretations.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878)

References

Beltrama, A., & Schwarz, F. (2021). Imprecision, personae, and pragmatic reasoning. In *Semantics and linguistic theory* (Vol. 31, pp. 122–144).

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57.

Carlson, L., Skubic, M., Miller, J., Huo, Z., & Alexenko, T. (2014). Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task. *Topics in cognitive science*, *6*(3), 513–533.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Duran, N. D., Dale, R., & Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition*, *121*(1), 22–40.

Fischer, K. (2005). Discourse conditions for spatial perspective taking. In *Proceedings of woslad workshop on spatial language and dialogue, delmenhorst.*

Fischer, K. (2007). The role of users' concepts of the robot in human-robot spatial instruction. In *Spatial cognition v reasoning, action, interaction: International conference spatial cognition 2006, bremen, germany, september 24-28, 2006, revised selected papers 5* (pp. 76–89).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of experimental social psychology*, *25*(3), 203–219.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Ip, M. H. K., & Papafragou, A. (2021). Listeners evaluate native and non-native speakers differently (but not in the way you think). In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Loy, J. E., & Demberg, V. (2023). Perspective taking reflects beliefs about partner sophistication: Modern computer partners versus basic computer and human partners. *Cognitive Science*, *47*(12), e13385.

Mayn, A., & Demberg, V. (2024). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 72–78).

Ngo, T. T. A. (2023). The perception by university students of the use of chatgpt in education. *International Journal of Emerging Technologies in Learning (Online)*, *18*(17), 4.

Peña, P. R., Doyle, P., Edwards, J., Garaialde, D., Rough, D., Bleakley, A., ... others (2023). Audience design and egocentrism in reference production during human-computer dialogue. *International Journal of Human-Computer Studies*, *176*, 103058.

Shoufan, A. (2023). Exploring students' perceptions of chatgpt: Thematic analysis and follow-up survey. *IEEE Access*.

Singh, H., Tayarani-Najaran, M.-H., & Yaqoob, M. (2023). Exploring computer science students' perception of chatgpt in higher education: A descriptive and correlation study. *Education Sciences*, *13*(9), 924.

Tippenhauer, N., Fourakis, E. R., Watson, D. G., & Lew-Williams, C. (2020). The scope of audience design in child-directed speech: Parents' tailoring of word lengths for adult versus child listeners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(11), 2163.