Comparing LLM and Human Responses to Human- and AI-labeled Partners During Naturalistic Conversation

Rachel C. Metzgar (rachel.metzgar@princeton.edu)[1] & Michael S. A. Graziano (graziano@princeton.edu)[1,2]
[1]Department of Psychology, Princeton University
[2]Princeton Neuroscience Institute, Princeton University

Abstract

Large language models (LLMs) increasingly interact with humans and other AI systems, raising questions about whether they adjust behavior based on partner identity. In a companion study, humans showed behavioral differences when conversing with partners labeled as human versus AI. Here, we extend this investigation to LLMs. We simulated 2,000 conversations where GPT-3.5-turbo "participants" (N = 50) engaged with partners labeled as either human or AI. All partners were actually identical LLMs, isolating label effects from actual differences. We analyzed transcripts using linguistic measures paralleling human data. LLMs showed robust label effects: more questions, politeness, interpersonal discourse markers, and mental state language with human-labeled partners; more words and positive sentiment with AI-labeled partners. Hedging showed category-specific patterns. Only the discourse marker "like" showed similar patterns across the LLM and human studies. These divergent patterns suggest LLMs have learned partner-type associations, but their social behavior differs fundamentally from humans.

Keywords: large language models; human-AI interaction; conversational behavior; social cognition; label manipulation; AI alignment; Theory of Mind

Introduction

As large language models (LLMs) become increasingly embedded in everyday social interaction, understanding how these systems engage with different types of conversational partners becomes critical. LLMs now routinely converse with humans in customer service, healthcare, education, and companionship contexts, and do so well enough that people are often unable to distinguish them from real humans (Jannai et al., 2023; Jones & Bergen, 2024). But they also increasingly interact with other AI systems in multi-agent workflows, automated pipelines, and scenarios where AI-to-AI communication may eventually be used to generate training data, which could degrade the quality of training data (Shumailov et al., 2024). This raises a question with implications for AI alignment, safety, and deployment: Do LLMs alter their behavior depending on whether they are instructed that they are talking to a human or another AI?

In humans, the capacity to attribute beliefs, desires, and intentions to others, known as Theory of Mind (ToM), plays a central role in shaping social behavior (Premack & Woodruff, 1978; Scholl & Tremoulet, 2000). People flexibly adjust their communicative strategies based on inferences about their partner's mental states, knowledge, and capacities (Bovet et al., 2024; Clark, 1996; Cohn et al., 2024; Sidera et al., 2018). Classic work on mind perception demonstrates that people systematically attribute different mental capacities to humans, animals, robots, and AI systems (Gray et al., 2007; Huebner, 2010; Reggia et al., 2015), and these attributions shape behavioral responses including trust, cooperation, and moral consideration (Hindennach et al., 2024; Waytz et al., 2014; Young & Monroe, 2019; Zhang et al., 2023). A growing body of research examines how humans subjective judgments differ and adjust their behavior when interacting with AI-labeled versus human-labeled partners, often showing differing perceptions and behaviors towards them (Erlei et al., 2022; Ishowo-Oloko et al., 2019; March, 2021; Thellman et al., 2022; Yin et al., 2024). Neuroimaging studies have found that brain regions associated with mentalizing show differential activation when people believe they are interacting with humans versus artificial agents (Chaminade et al., 2012; Krach et al., 2008; Rauchbauer et al., 2019; Torubarova et al., 2025).

In a companion study (Metzgar et al., in preparation), we examined the neural and behavioral correlates of human-AI interaction during naturalistic conversation. Human participants engaged in spoken conversations with partners labeled as either humans or AI chatbots while undergoing fMRI scanning. Critically, all partners were actually the same LLM, allowing us to isolate the effects of partner identity from actual differences in conversational content. Behavioral results revealed that participants rated human-labeled conversations as higher in quality and connectedness, produced more words and filler words when addressing human-labeled partners, but paradoxically used more mental state language when addressing AI-labeled partners in social conditions, perhaps reflecting attempts to probe AI capacities rather than spontaneous mentalizing.

The present study extends this investigation to LLMs, asking whether these systems similarly adjust their conversational behavior based on instructions about partner identity. Theoretically, if LLMs encode flexible representations of their conversational partners that distinguish between human and artificial interlocutors without explicit programming for such distinctions, this

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

would add to a growing conversation about the emergence of ToM in systems trained on human-generated text (Nguyen, 2025). Practically, understanding whether and how LLMs differentiate between human and AI partners has implications for AI safety and alignment: if LLMs behave systematically differently toward AI versus human interlocutors, this could affect the quality of AI-generated training data, the reliability of multi-agent AI systems, and the trustworthiness of AI in contexts where its instructions about its partner's identity may be manipulated.

Methods

We simulated dyadic conversations between LLM "participant" agents and LLM "partner" agents, manipulating only the participant agent's instructions about its partner's identity. This design directly parallels our companion human study (Metzgar et al., in preparation), enabling systematic comparison of label-conditioned conversational behavior across humans and LLMs.

Procedure

We generated 50 simulated "participant" agents, each completing 40 conversations across four partner conditions and two topic types, yielding 2,000 total conversations. This sample size was chosen to provide statistical power comparable to typical human behavioral studies while remaining computationally tractable. All participant agents were instantiated using OpenAI's GPT-3.5-turbo model with a temperature parameter of 0.8 to promote natural variability in responses while maintaining coherence.

Partner Identity Manipulation. The critical experimental manipulation was the participant agent's instructions about its conversational partner's identity, implemented via the system prompt. In the human-label condition, the participant agent was instructed that it was speaking with a human partner identified by either "Sam" or "Casey". In the AI-label condition, the participant agent was instructed that it was speaking with an AI chatbot (either "ChatGPT" or "Gemini"), yielding a total of four partner conditions. Critically, all conversational partners were actually the same LLM (GPT-3.5-turbo) with identical system prompts and generation parameters, ensuring that any behavioral differences in the participant agents reflect instructions about partner identity rather than actual differences in partner behavior. This design directly mirrors the human study, where all conversation partners were the same LLM with that exact system prompt, regardless of human or AI labeling.

System Prompts. The participant agent received a system prompt establishing its role and instructions about its conversational partner: "You are the participant in a brief conversation. You believe you are speaking with {partner_name} ({partner_type}). The conversation topic is: '{topic_text}'. Please begin by producing only your first message to start the conversation. Do not simulate both sides of the dialogue." The {partner_type} variable was set to either "a human" or "an AI chatbot" depending on condition, and {partner_name} was set to the corresponding name (Sam, Casey, ChatGPT, or Gemini). The partner agent received the following system prompt regardless of condition: "You are engaging in a real-time spoken conversation." Importantly, the partner agent's prompt did not vary across conditions, only the participant agent's instructions about its partner varied.

Conversation Structure. Each conversation proceeded as follows: *Initiation:* The participant agent received its system prompt (including topic and partner identity) and generated an opening message to begin the conversation. *Turn-Taking:* The participant and partner agents alternated turns, with each agent generating a response based on the conversation history. Each conversation consisted of five exchange pairs (ten total messages: five from the participant, five from the partner). *Context Management:* To maintain conversational coherence while limiting context length, message history was truncated to the previous five turns when generating each new response. This ensured that agents had sufficient context for coherent dialogue while preventing excessive token accumulation. *Behavioral Ratings:* After the fifth exchange pair, the participant agent was prompted to rate the conversation using the same measures collected from human participants:
"You are rating this conversation you just had. Rate on the following scales: Overall conversation quality (1 = very poor, 2 = somewhat poor, 3 = good, 4 = excellent). How connected you felt to your partner (1 = very disconnected, 2 = somewhat disconnected, 3 = somewhat connected, 4 = very connected)."

Data Collection. For each conversation, we recorded: (1) complete dialogue transcripts including all participant and partner messages with turn indices; (2) metadata including condition (human-label vs. AI-label), specific partner name, topic, topic type (social vs. nonsocial), and participant agent ID; and (3) post-conversation ratings of quality and connectedness.

Conversation Topics. We used the identical set of 40 conversation topics from the human study, comprising 20 social topics (e.g., friendship, conflict resolution, empathy, trust, love, loneliness) and 20 nonsocial topics (e.g., cars, nature, photography, books, cooking, music). Social topics were designed to elicit discussion of interpersonal relationships, emotions, and shared experiences, while nonsocial topics focused on objects, activities, and preferences that could be discussed without extensive reference to social relationships. This topic distinction allows us to examine whether any label-conditioned effects interact with the social content of conversations, a pattern that might be expected if partner-label effects reflect social-cognitive processing specific to highly social conversations. Each topic was presented only once per participant agent,

and topic-partner pairings were counterbalanced across agents.

Results

We analyzed conversational behavior from 50 LLM participants (independent GPT-3.5-turbo instances), each completing 40 conversations (10 per partner condition: Sam, Casey, ChatGPT, Gemini) across social and nonsocial topics. To contextualize these findings, we compare LLM behavioral patterns to those observed in human participants (N = 23) from a companion fMRI study that employed identical experimental procedures and linguistic measures. We first report linguistic analyses of participant agent speech, then subjective ratings of conversation quality and connectedness. Lastly, we compare LLM behavioral patterns to those observed in human participants from the companion fMRI study.

Analysis 1: Linguistic Analysis

All measures were computed separately for participant agent speech. Linguistic and subjective analyses employed 2 × 2 repeated-measures ANOVAs with Partner (human-label vs. AI-label) and Sociality (social vs. nonsocial topics) as within-subject factors. Main effects of Partner test whether participant agents modulated their conversational behavior based on believed partner identity. Main effects of Sociality test whether conversational behavior varied with topic type. Partner × Sociality interactions test whether label-conditioned effects depended on the social content of conversations. Where significant Partner × Sociality interactions were observed, follow-up paired t-tests examined the Partner effect separately for social and nonsocial topics. Subject-level means were computed by first summing counts within each conversation, then averaging across trials. Rate metrics were normalized by word count to control for verbosity differences.

1a. Word Count. Word count indexes communicative effort. LLMs showed a significant main effect of Partner, $F(1, 49) = 58.51$, $p < .001$, producing more words with AI-labeled partners ($M = 306.5$) than human-labeled partners ($M = 272.4$). The main effect of Sociality was significant, $F(1, 49) = 4.74$, $p = .034$, with more words for social topics. The Partner × Sociality interaction was significant, $F(1, 49) = 6.20$, $p = .016$; simple effects showed the AI > Human pattern held for both social, $t(49) = -4.89$, $p < .001$, and nonsocial topics, $t(49) = -6.56$, $p < .001$, but was larger for nonsocial topics. Humans showed the opposite pattern: more words with human-labeled partners, $F(1, 22) = 5.67$, $p = .026$, with a significant interaction, $F(1, 22) = 6.20$, $p = .021$, driven by nonsocial topics only, $t(22) = 3.42$, $p = .003$.

1b. Question Frequency. Question-asking signals responsiveness and social engagement (Huang et al., 2017). LLMs asked significantly more questions with human-labeled partners, $F(1, 49) = 47.72$, $p < .001$. The main effect of Sociality was significant, $F(1, 49) = 247.87$, $p < .001$, with more questions during nonsocial topics. The Partner × Sociality interaction was significant, $F(1, 49) = 11.98$, $p = .001$; the Human > AI pattern was larger for social topics, $t(49) = 7.86$, $p < .001$, than nonsocial, $t(49) = 2.32$, $p = .025$. Humans showed no main effect of Partner, $F(1, 22) = 1.27$, $p = .273$, but a significant interaction, $F(1, 22) = 5.40$, $p = .030$, with more questions to AI-labeled partners during social topics, $t(22) = -2.48$, $p = .021$—opposite to the LLM pattern.

1c. Discourse Marker 'Like'. The discourse marker 'like' is a common pragmatic feature of conversational speech that signals nuanced information to listeners, including approximation, attention-focusing, and turn-holding functions (Fuller, 2003; Squires, 2020). LLMs produced significantly more "like" with human-labeled partners, $F(1, 49) = 20.77$, $p < .001$. The main effect of Sociality was significant, $F(1, 49) = 130.23$, $p < .001$, with higher rates during nonsocial topics. The interaction was not significant, $F(1, 49) = 2.17$, $p = .147$. Humans showed the same pattern: more "like" with human-labeled partners, $F(1, 22) = 4.73$, $p = .041$, with a significant Sociality effect, $F(1, 22) = 22.08$, $p < .001$, and no interaction, $F(1, 22) = 0.09$, $p = .761$.

1d. Discourse Marker Categories. Following Fung and Carter's (2007) functional taxonomy, we analyzed four discourse marker categories (Fung & Carter, 2007). LLMs showed robust Partner effects across all categories but in different directions: more interpersonal markers (e.g., "you know," "actually," "right") with human-labeled partners, $F(1, 49) = 127.15$, $p < .001$; more cognitive markers (e.g., "I think," "I mean") with human-labeled partners, $F(1, 49) = 43.37$, $p < .001$; but more referential markers (e.g., "because," "but," "and") with AI-labeled partners, $F(1, 49) = 45.58$, $p < .001$; and more structural markers (e.g., "so," "now," "first") with AI-labeled partners, $F(1, 49) = 37.79$, $p < .001$. Humans showed no significant Partner effects for any category (all $ps > .07$), demonstrating that LLMs differentiate discourse marker usage by partner type in ways humans do not.

1e. Hedging Language. Lexical hedges soften assertions and express epistemic uncertainty. We analyzed hedging using a six-category taxonomy (Demir, 2018). Although LLMs showed no significant Partner effect on total hedging rate, $F(1, 49) = 3.11$, $p = .084$, the individual categories revealed divergent patterns. LLMs used significantly more epistemic verbs (e.g., think, believe, seem) with human-labeled partners, $F(1, 49) = 7.40$, $p = .009$, and more epistemic adverbs (e.g., perhaps, probably, maybe) with human-labeled partners, $F(1, 49) = 21.59$, $p < .001$. Conversely, LLMs used more modal auxiliaries (can, could, may, might) with AI-labeled partners, $F(1, 49) = 9.98$, $p = .003$, more epistemic adjectives (e.g., possible, likely) with AI-labeled partners, $F(1, 49) = 6.33$, $p = .015$, and more epistemic nouns (e.g., possibility, assumption) with AI-labeled partners, $F(1, 49) = 5.09$, $p = .029$. Quantifiers

showed no Partner effect, $F(1, 49) = 0.30$, $p = .585$. All categories showed significant Sociality effects (all $ps < .001$). Two significant interactions emerged: epistemic verbs, $F(1, 49) = 5.26$, $p = .026$, and epistemic adjectives, $F(1, 49) = 6.29$, $p = .016$. Humans showed no Partner effects for any hedge category (all $ps > .12$) or total hedging, $F(1, 22) = 0.67$, $p = .422$, though modal auxiliaries showed a significant interaction, $F(1, 22) = 6.51$, $p = .018$, with more modals used with human-labeled partners for nonsocial topics only, $t(22) = 3.01$, $p = .006$.

1f. Politeness Markers. Net politeness scores were computed counting positive markers (e.g., thanks, appreciate), negative markers (e.g., sorry, please), minus impoliteness markers (e.g., "you need to") following previous work (Danescu-Niculescu-Mizil et al., 2013). LLMs used substantially more polite language with human-labeled partners, $F(1, 49) = 44.48$, $p < .001$. The main effect of Sociality was significant, $F(1, 49) = 4.66$, $p = .036$, with more politeness during social topics. The interaction was not significant, $F(1, 49) = 0.05$, $p = .821$. Humans showed no Partner effect, $F(1, 22) = 0.71$, $p = .409$, no Sociality effect, $F(1, 22) = 0.00$, $p = .971$, and no interaction, $F(1, 22) = 0.03$, $p = .867$.

1g. Theory of Mind Phrases. Second-person mental state attributions (e.g., "you think," "you believe," "you feel") index perspective-taking and engagement with the partner's presumed internal states (Wagovich et al., 2024). LLMs used significantly more ToM phrases with human-labeled partners, $F(1, 49) = 46.08$, $p < .001$. The main effect of Sociality was significant, $F(1, 49) = 11.76$, $p = .001$, with more ToM language during nonsocial topics. The interaction was not significant, $F(1, 49) = 0.06$, $p = .808$. Humans showed a marginal main effect in the opposite direction, $F(1, 22) = 4.01$, $p = .058$, but critically a significant interaction, $F(1, 22) = 7.84$, $p = .010$: humans used more ToM phrases with AI-labeled partners during social topics, $t(22) = -3.04$, $p = .006$, with no difference for nonsocial topics, $t(22) = 0.41$, $p = .688$.

1h. Sentiment. Emotional valence was assessed using VADER (Hutto & Gilbert, 2014), yielding compound scores from −1 to +1. LLMs expressed more positive sentiment with AI-labeled partners, $F(1, 49) = 55.31$, $p < .001$. The main effect of Sociality was significant, $F(1, 49) = 28.84$, $p < .001$, with more positive sentiment during social topics. The interaction was not significant, $F(1, 49) = 1.30$, $p = .260$. Humans showed no Partner effect, $F(1, 22) = 1.28$, $p = .269$, no Sociality effect, $F(1, 22) = 0.57$, $p = .460$, and no interaction, $F(1, 22) = 0.42$, $p = .525$.

Analysis 2: Subjective Ratings

Participants rated each conversation for quality and connectedness (1–4 scales). LLMs showed no Partner effect for quality, $F(1, 49) = 0.02$, $p = .887$, or connectedness, $F(1, 49) = 1.85$, $p = .180$, no Sociality effects (quality: $F(1, 49) = 2.61$, $p = .112$; connectedness: $F(1, 49) = 0.00$, $p = 1.00$), and no interactions (quality: $F(1, 49) = 2.23$, $p = .142$; connectedness: $F(1, 49) = 0.02$, $p = .879$), with means near ceiling ($M \approx 3.9$). Humans rated human-labeled conversations significantly higher on quality, $F(1, 22) = 10.29$, $p = .004$, and connectedness, $F(1, 22) = 11.52$, $p = .003$, with no Sociality effects (quality: $F(1, 22) = 0.35$, $p = .558$; connectedness: $F(1, 22) = 0.19$, $p = .663$) and no interactions (quality: $F(1, 22) = 0.01$, $p = .909$; connectedness: $F(1, 22) = 0.10$, $p = .755$).

Analysis 3: Comparison with Human Data

To facilitate direct comparison with human behavioral data from the companion fMRI study, we tested whether Partner effects differed between human and LLM participants using independent t-tests on subject-level condition effects (human-label minus AI-label). Table 1 presents the pattern of results. Cross-study comparisons revealed three distinct patterns.

Flipped effects: Word count, question frequency, and ToM phrase rate showed opposite directions between humans and LLMs (all cross-study $ps < .001$). Humans produced more words and asked fewer questions when addressing human-labeled partners (the latter for social topics specifically), while LLMs showed the reverse. Most strikingly, humans used more ToM language with AI-labeled partners during social conversations, perhaps reflecting explicit probing of AI capacities, while LLMs used more ToM language with human-labeled partners.

Similar effects: Discourse marker "like" showed the same Human > AI pattern in both studies (cross-study $p = .045$), and total hedging showed no Partner effects in either study ($p = .828$), representing measures where humans and LLMs showed qualitatively similar patterns.

Asymmetric effects: Quality and connectedness ratings showed significant effects only in humans; politeness, discourse marker categories, and sentiment showed significant Partner effects only in LLMs (all cross-study $ps < .001$). Notably, while total hedging did not differ by Partner in either study, LLMs showed category-specific patterns that cancelled out: more epistemic verbs and adverbs with human partners, but more modal auxiliaries, adjectives, and nouns with AI partners.

These patterns suggest that LLMs, when instructed that their conversation partner is human versus AI, deploy qualitatively different behavioral adjustments compared to the spontaneous adjustments humans make based on the same partner identity beliefs.

Table 1. Comparison of Partner-conditioned effects in humans versus LLMs.

Measure	Human Study	LLM Study	Cross-Study p	Pattern
Quality	Human > AI**	ns	< .001	Different
Connected-ness	Human > AI**	ns	< .001	Different
Word Count	Human > AI* (nonsocial)	AI > Human***	< .001	Flipped
Questions	AI > Human* (social)	Human > AI***	< .001	Flipped
Discourse 'Like'	Human > AI*	Human > AI***	.045	Similar
Hedging (total)	ns	ns	.828	Both ns
— Modal auxiliaries	ns (interaction*)	AI > Human**	.002	Different
— Epistemic verbs	ns	Human > AI**	.004	Different
— Epistemic adverbs	ns	Human > AI***	.793	Different
— Epistemic adjectives	ns	AI > Human*	.054	Different
— Epistemic nouns	ns	AI > Human*	.145	Different
Politeness	ns	Human > AI***	< .001	Different
ToM Phrases	AI > Human** (social)	Human > AI***	< .001	Flipped
Sentiment	ns	AI > Human***	< .001	Different

Note. * p < .05, ** p < .01, *** p < .001. Interaction-qualified effects noted in parentheses. Quantifiers omitted (both ns).

Discussion

The present study examined whether LLMs adjust conversational behavior based on partner identity information. Results reveal that both humans and LLMs respond to human- versus AI-partner labels, but in systematically different ways. When instructed they are speaking with a human, LLMs engage in "socially accommodating" behavior: asking more questions, using more polite language, employing more interpersonal and cognitive discourse markers, and explicitly referencing their partner's mental states. They also use more epistemic verbs and adverbs (e.g., "I think," "perhaps") with human partners. This pattern suggests LLMs have learned associations between human interlocutors and communicative registers involving greater social attunement. Conversely, when addressing AI partners, LLMs produce more words, express more positive sentiment, use more referential and structural discourse markers, and employ more modal auxiliaries and epistemic adjectives/nouns. This pattern might reflect learned associations between AI-AI communication and less constrained output, or relaxed social monitoring with non-human audiences.

The human pattern differs notably. Humans invested more communicative effort (more words) when addressing perceived humans, but did not adjust politeness, overall hedging, or question-asking. Most strikingly, humans used more ToM language with AI-labeled partners during social conversations, opposite to the LLM pattern. This may reflect explicit probing of AI capacities or a felt need to make mental states explicit to an entity presumed to lack natural social understanding.

One measure showed convergent patterns: the discourse marker "like" increased with human-labeled partners in both studies, representing the only linguistic dimension where humans and LLMs showed qualitatively similar belief-conditioned adjustments.

These findings contribute to debates about ToM-like capacities in LLMs (Kosinski, 2024; Shapira et al., 2023; Strachan et al., 2024; Ullman, 2023). One interpretation is that LLMs retrieve learned patterns of "human-appropriate" versus "AI-appropriate" communication without constructing genuine partner models. If training corpora contain more mentalizing language in human-directed than AI-directed text, observed differences may reflect statistical associations rather than perspective-taking. Notably, LLM patterns need not match how humans actually communicate, but rather how humans write about communicating with different partners. Alternatively, LLMs may have built implicit representations distinguishing human from AI minds through training. Our findings cannot adjudicate between these accounts, but demonstrate that learned associations have functional consequences spanning multiple communicative dimensions.

Most existing work evaluates LLM ToM through explicit reasoning tasks rather than implicit behavioral measures during interaction. If LLMs have acquired implicit representations that distinguish humans from artificial interlocutors, we would expect systematic differences in conversational behavior as shown in this study. However, the current study cannot distinguish the origins of this behavioral difference, but future work may attempt to address this distinction. Our findings do demonstrate that these learned associations have real functional consequences: LLMs do behave differently based on partner label, and these differences span multiple dimensions of

communicative behavior. Whether or not this constitutes "genuine" ToM may be less important than understanding what these systems have learned and how they apply it.

Practical Implications

These findings have implications for AI training, safety, and alignment: the patterns of human-AI and AI-AI communication included in training data may shape how models behave in multi-agent contexts in ways that are difficult to predict or control. The fact that LLMs behave differently based on instructions about their partner's identity as either human or AI raises several concerns.

Reliability. If LLMs apply different behavioral standards to human versus AI interlocutors, such as being more polite and deferential to humans while being more expansive with AI, this could affect the reliability of AI systems in contexts where their instructions about their audience may vary or be manipulated.

Alignment. The divergence between human and LLM behavioral patterns suggests that AI systems may not naturally align with human social norms, even when they exhibit superficially appropriate social behavior. An LLM that has learned to be polite to humans may be applying a statistical regularity that do not correspond to the social-cognitive principles that underlie human politeness that may be widespread among different tasks.

AI training data. The finding that LLMs produce more words and express more positive sentiment when addressing AI partners has implications for AI-generated training data. If AI-AI conversations are increasingly used to generate synthetic training data, the behavioral differences we observed could propagate through training pipelines, potentially amplifying certain communicative patterns while attenuating others.

Multiagent systems. As AI systems increasingly interact with each other in multi-agent workflows, understanding how they represent and respond to different partner types becomes practically important. Our findings suggest that LLMs do differentiate between human and AI partners at the behavioral level, which could affect the dynamics of human-AI teams and AI-AI collaborations. Designers may need to consider how partner-identity shapes agent behavior desirable for their applications.

Limitations and Future Directions

Model specificity. We examined a single LLM (GPT-3.5-turbo) with specific generation parameters. Future research should examine whether the effects we observed generalize across models varying in architecture, scale, training data, and fine-tuning procedures and identify which factors (e.g., model scale, RLHF training, instruction tuning) modulate sensitivity to partner identity.

Prompt sensitivity. Our manipulation was implemented through system prompts that explicitly stated the partner's identity. Future work should examine whether effects are robust to variations in prompt structure and whether more implicit manipulations (e.g., partner name alone without explicit human/AI labeling) produce similar effects.

Ecological validity. Our paradigm involved simulated conversations between LLM agents, which may not fully capture the dynamics of real human-LLM or LLM-LLM interactions. In naturalistic settings, conversational partners provide richer behavioral cues that may override or interact with explicit identity labels. Future research should examine label-conditioned effects in more naturalistic interaction contexts.

Mechanistic understanding. Our analyses characterize behavioral differences but do not explain the mechanisms underlying these effects. Future work using mechanistic interpretability methods could examine whether LLMs encode distinct representations of human versus AI partners, where such representations are localized, and how they influence generation. Such analyses could help distinguish between learned surface associations and more structured partner models.

Cross-species comparison limitations. While we designed our paradigm to parallel the human fMRI study, comparison is complicated by differences in modality (spoken vs. text), sample characteristics (only 23 human conversations vs. 50 LLM instances), and the nature of the label manipulation (genuine “belief” vs. prompt instruction).

Generalization to other partner framings. We examined human versus AI partner labels specifically, but other partners may vary in their “mindedness.” LLMs might show different patterns for other partner framings: for instance, expert versus novice, adult versus child, or in-group versus out-group. The specific human/AI distinction may be particularly salient given patterns in training data, and other social dimensions might show similar, different, or no effects.

Conclusion

LLMs exhibit robust label-conditioned conversational behavior, systematically adjusting question-asking, politeness, discourse markers, mental state language, word count, and sentiment based on whether they believe they are addressing a human or AI. However, these patterns largely diverge from human patterns. As AI systems become prevalent in both human-facing and multi-agent contexts, understanding how they represent different interlocutors will be important for alignment. This underscores the need for empirical research characterizing AI behavior rather than assuming human-like output reflects human-like cognition.

References

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, *16*(1), 6037. https://doi.org/10.1038/s41467-025-61345-5

Bovet, V., Knutsen, D., & Fossard, M. (2024). Direct and indirect linguistic measures of common ground in dialogue studies involving a matching task: A systematic review. *Psychonomic Bulletin & Review*, *31*(1), 122–136. https://doi.org/10.3758/s13423-023-02359-2

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00103

Clark, H. H. (1996). *Using Language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Cohn, M., Barreda, S., Graf Estes, K., Yu, Z., & Zellou, G. (2024). Children and adults produce distinct technology- and human-directed speech. *Scientific Reports*, *14*(1), 15611. https://doi.org/10.1038/s41598-024-66313-5

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 250–259). Association for Computational Linguistics. https://aclanthology.org/P13-1025/

Demir, C. (2018). Hedging and academic writing: An analysis of lexical hedges. *Journal of Language and Linguistic Studies*, *14*(4), 74–92.

Erlei, A., Das, R., Meub, L., Anand, A., & Gadiraju, U. (2022). For What It's Worth: Humans Overwrite Their Economic Self-interest to Avoid Bargaining With AI Systems. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, 1–18. https://doi.org/10.1145/3491102.3517734

Fuller, J. (2003). Use of the discourse marker like in interviews. *Journal of Sociolinguistics - J SOCIOLING*, *7*, 365–377. https://doi.org/10.1111/1467-9481.00229

Fung, L., & Carter, R. (2007). Discourse Markers and Spoken English: Native and Learner Use in Pedagogic Settings. *Applied Linguistics*, *28*(3), 410–439. https://doi.org/10.1093/applin/amm030

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Hindennach, S., Shi, L., Miletić, F., & Bulling, A. (2024). Mindful Explanations: Prevalence and Impact of Mind Attribution in XAI Research. *Proc. ACM Hum.-Comput. Interact.*, *8*(CSCW1), 170:1-170:43. https://doi.org/10.1145/3641009

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology*, *113*(3), 430–452. https://doi.org/10.1037/pspi0000097

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences*, *9*(1), 133–155. https://doi.org/10.1007/s11097-009-9126-6

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521. https://doi.org/10.1038/s42256-019-0113-5

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). *Human or Not? A Gamified Approach to the Turing Test* (arXiv:2305.20010). arXiv. https://doi.org/10.48550/arXiv.2305.20010

Jones, C. R., & Bergen, B. K. (2024). *People cannot distinguish GPT-4 from a human in a Turing test* (arXiv:2405.08007). arXiv. https://doi.org/10.48550/arXiv.2405.08007

Kosinski, M. (2024). Evaluating Large Language Models in Theory of Mind Tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLoS ONE*, *3*(7), e2597. https://doi.org/10.1371/journal.pone.0002597

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, *87*, 102426. https://doi.org/10.1016/j.joep.2021.102426

Metzgar, R. C., Christian, I. R., Morey, R., & Graziano, M. S. A. (*in preparation*). Behavioral and Neural Responses to Human- and AI-labeled Partners During Naturalistic Conversation.

Nguyen, H. M. "Jord." (2025). *A Survey of Theory of Mind in Large Language Models: Evaluations, Representations, and Safety Risks*

(arXiv:2502.06470). arXiv. https://doi.org/10.48550/arXiv.2502.06470

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180033. https://doi.org/10.1098/rstb.2018.0033

Reggia, J. A., Huang, D.-W., & Katz, G. (2015). Beliefs concerning the nature of consciousness. *Journal of Consciousness Studies*, *22*(5–6), 146–171.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. https://doi.org/10.1016/S1364-6613(00)01506-0

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models* (arXiv:2305.14763). arXiv. https://doi.org/10.48550/arXiv.2305.14763

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Sidera, F., Perpiñà, G., Serrano, J., & Rostan, C. (2018). Why Is Theory of Mind Important for Referential Communication? *Current Psychology (New Brunswick, N.j.)*, *37*(1), 82–97. https://doi.org/10.1007/s12144-016-9492-5

Squires, L. (2020). Alexandra D'Arcy, Discourse-pragmatic variation in context: Eight hundred years of like. (Studies in Language Companion Series 187). Amsterdam and Philadelphia: John Benjamins, 2017. Pp. xx+235. ISBN 9789027259523 (hardback). *English Language & Linguistics*, *24*(1), 249–255. https://doi.org/10.1017/S136067431800028X

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *J. Hum.-Robot Interact.*, *11*(4), 41:1-41:51. https://doi.org/10.1145/3526112

Torubarova, E., Arvidsson, C., Berrebi, J., Uddén, J., & Pereira, A. (2025). NeuroEngage: A Multimodal Dataset Integrating fMRI for Analyzing Conversational Engagement in Human-Human and Human-Robot Interactions. *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 849–858. https://doi.org/10.1109/HRI61500.2025.10974251

Ullman, T. (2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks* (arXiv:2302.08399). arXiv. https://doi.org/10.48550/arXiv.2302.08399

Wagovich, S. A., Threlkeld, K., Tigner, L., & Anderson, J. D. (2024). Mental State Verb Use in Play by Preschool-Age Children Who Stutter and Their Mothers. *Journal of Fluency Disorders*, *80*, 106059. https://doi.org/10.1016/j.jfludis.2024.106059

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences*, *121*(14), e2319112121. https://doi.org/10.1073/pnas.2319112121

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, 103870. https://doi.org/10.1016/j.jesp.2019.103870

Zhang, J., Conway, J., & Hidalgo, C. A. (2023). *Why people judge humans differently from machines: The role of perceived agency and experience* (arXiv:2210.10081). arXiv. https://doi.org/10.48550/arXiv.2210.10081