LLMs and people both learn to form conventions – just not with each other

Cameron R. Jones[1,4], **Agnese Lombardi**[2], **Kyle Mahowald**[3] **& Benjamin K. Bergen**[4]
[1]Department of Psychology, Stony Brook University
[2]Department of Philology, Literature, and Linguistics, University of Pisa
[3]Department of Linguistics, University of Texas at Austin
[4]Department of Cognitive Science, University of California San Diego

Abstract

Humans align to one another in conversation – adopting shared conventions that ease communication. We test whether LLMs form the same kinds of conventions in a multimodal communication game. Both humans and LLMs displayed evidence of convention-formation (increasing the accuracy and consistency of their turns) when communicating in same-type dyads (humans with humans, AI with AI), though AI-AI pairs show limited reduction in length. Heterogeneous human-AI pairs failed to converge on effective referring expressions, suggesting differences in communicative tendencies. In Experiment 2, we prompting LLMs to produce superficially humanlike behavior. While the length of prompted models' messages matched that of human pairs, accuracy and lexical overlap in human-AI pairs continued to lag behind that of both human-human and AI-AI pairs. These results suggest that conversational alignment requires more than just the ability to mimic previous interactions, but also shared interpretative biases toward the meanings that are conveyed.

Keywords: LLMs; conversational alignment; communicative grounding; conceptual pacts; visual reference games; tangrams

Introduction

Humans align with one another in conversation (Garrod & Pickering, 2004). They come to use vocabulary (Brennan & Clark, 1996), phrases (Garrod & Anderson, 1987), and syntactic constructions (Branigan et al., 2000) that their partners have used more often. This could be because people are cooperative communicators who try to be informative, concise, and on-topic (Grice, 1975; Lewis, 2008; Wilkes-Gibbs & Clark, 1992). However, carrying out this kind of explicit cooperative alignment requires maintaining complex representations of partners' mental states and recursively tracking beliefs about what one's partner knows about oneself (Clark, 1996). Alternatively, alignment might emerge through simpler mechanisms like frequency heuristics (Barr, 2004), priming mechanisms (Chartrand & Bargh, 1999; Menenti et al., 2012), and first-order social rules (Hawkins et al., 2017). However, there is debate about the extent to which simpler statistical mechanisms could account for such rich interactive behavior (Fusaroli & Tylén, 2016; Pickering & Garrod, 2004).

Large language models (LLMs) have proven effective at displaying humanlike behavior in short interactions (Jones & Bergen, 2025) and engaging millions of consumers (Chatterji et al., 2025). However, it is unclear whether they develop the same kinds of conversational alignment that people do. On the one hand, LLMs are effective in-context learners (Brown et al., 2020; Lampinen et al., 2025) and are trained via RLHF to be cooperative communicators (Bai et al., 2022), which could equip them to align with partners. On the other, they lack humans' resource-sensitivity (Clark & Brennan, 1991; Zipf, 1949) and communicative intent (Bender & Koller, 2020; Shanahan, 2024), and may differ in how they carve up meaning spaces (Brennan & Clark, 1996; Mahowald et al., 2024)—any of which could prevent them from forming common ground.

To the extent that LLMs develop humanlike conventions, it could suggest that purely statistical learning (from distributional language statistics and reinforcement feedback) is sufficient to facilitate conversational alignment. In contrast, if LLMs are incapable of this kind of alignment, it could suggest that more fundamental differences between human and LLM communicators are crucial for establishing common ground.

Here we investigate whether people and LLMs form shared conventions when they communicate in a version of the tangrams reference game (Wilkes-Gibbs & Clark, 1992). In the game, partners communicate about a grid of ambiguous figures made from sets of simple geometrical shapes called tangrams. Human participants' descriptions tend to become shorter, more effective, and more stable over multiple rounds, indicating that they are forming conventions (Hawkins et al., 2017; Wilkes-Gibbs & Clark, 1992). In Experiment 1, we compare Human-Human (H-H), Human-AI (H-AI), and AI-AI pairs. After finding differences across these kinds of pairs, in Experiment 2 we modify the prompt for the AI partner to determine whether LLMs can be induced to behave more like human conversants through prompting.

Our work follows several recent studies which investigate dyadic reference in LLMs. Hua and Artzi (2024) evaluated multimodal LLMs as speakers and listeners and found that most models do not spontaneously make their own descriptions more efficient even when their interlocutor does. Hua et al. (2025) extended this line by developing a post-training procedure that improves convention formation in text-only reference games. Vaduguru et al. (2025) found that when models were rewarded for both communicative success and message cost, conventions emerged with human listeners. Zeng et al. (2026) used a similar paradigm to ours—a referential communication game about images of baskets—and compared H-AI dyads to pairs of humans and AIs. Their task differed from ours in that one partner was the director throughout the experiment rather than swapping roles each turn, that each decision

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

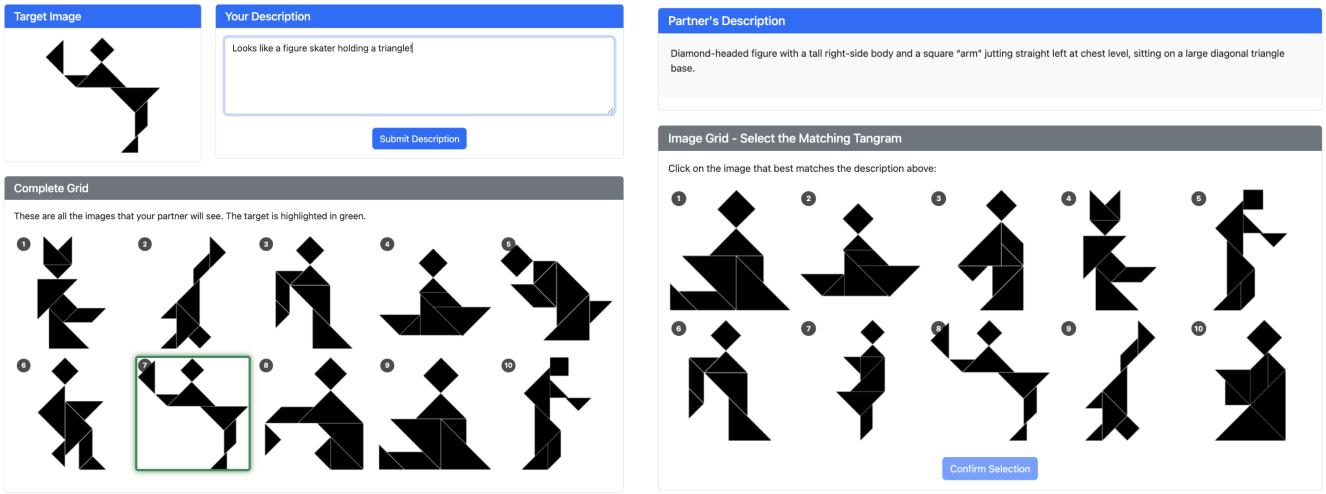

Figure 1: The interface from the perspective of the human director (left) and the matcher (right). As directors, participants see a grid of figures and are instructed to describe the designated figure (in the green box) such that their partner, the matcher, will select it. The matcher is instructed to select the figure described by the director from the presented grid.

was based on multi-turn dialogue and that no explicit feedback was given between turns. They found that both AI-AI and H-AI pairs failed to develop conventions, with AI-AI accuracy dropping across rounds. Our work extends this literature by testing live, bidirectional coordination with general-purpose chat models, and by asking whether prompting models to behave more humanlike can close the gap with H-H pairs.

Experiment 1

In Experiment 1, we compared three different types of partnership: H-H, H-AI, and AI-AI. Each pair communicated about a grid of 10 images over 50 turns. We used this setup to address four pre-registered research questions (Jones et al., 2025a).

First, we asked whether each type of partnership would develop conventions over the course of the experiment as measured by a) their accuracy, b) their length, and c) the degree of lexical overlap between descriptions of the same image across rounds. In line with prior work (Hawkins et al., 2017; Wilkes-Gibbs & Clark, 1992), we predicted H-H pairs would improve on each metric. If statistical learning from next-token prediction and the attendant in-context learning observed in LLMs is sufficient to generate humanlike conversational alignment, then we would expect that accuracy, length, and consistency across rounds should also improve in AI-AI and H-AI pairs.

Second, we asked whether some types of partnership perform better than others. To the extent that LLMs lack humanlike resource-consciousness, communicative intent, or representational biases, we should expect that H-AI pairs will fall behind H-H pairs on all three metrics of convention formation. Different theories make contrasting predictions about AI-AI pairs. On one hand, if forming conventions requires recursively tracking interlocutors' mental states (at which humans outperform LLMs; Jones et al., 2024), then we might expect H-AI pairs (with at least one human) to best AI-AI pairs (with none). Alternatively, if convention-formation rests on sharing conceptual and representational biases about how to carve up meaning spaces, then we might expect homogenous AI-AI pairs to outperform heterogenous H-AI ones.

Third, we asked whether human partners would adapt better to human partners: specifically whether they would produce shorter or more conventional messages to human partners than they did to AI partners. This would suggest that human behavior at the task is flexible and sensitive to partner behavior.

Lastly, we asked about participants' self-reported perception of human and AI partners in an exit survey. We predicted that humans would rate other humans as more helpful, adaptable, and more likely to be human than AI partners.

Method

Materials We used a sample of 18 humanoid tangram figures from the KiloGram dataset (Ji et al., 2022). We used GPT-5 as the AI model in these experiments. We accessed the model through the OpenAI API between 2025-10-13 and 2025-10-24 with reasoning effort set to "low". The model prompt was identical to the instructions that human participants saw, with additional information about how responses should be formatted. The full text of all prompts are available on OSF (Jones et al., 2025c).

Participants We recruited 168 participants through our institution's undergraduate subject pool. 59 participants failed to complete the experiment (for example because either they or their partner disconnected during the experiment). No participants met our exclusion criteria of having average message lengths of less than 10 characters or average response times of less than 5s. Overall we retained data from 109 participants, yielding 39 H-H and 31 H-AI sessions. We additionally ran 30 AI-AI sessions using the same setup.

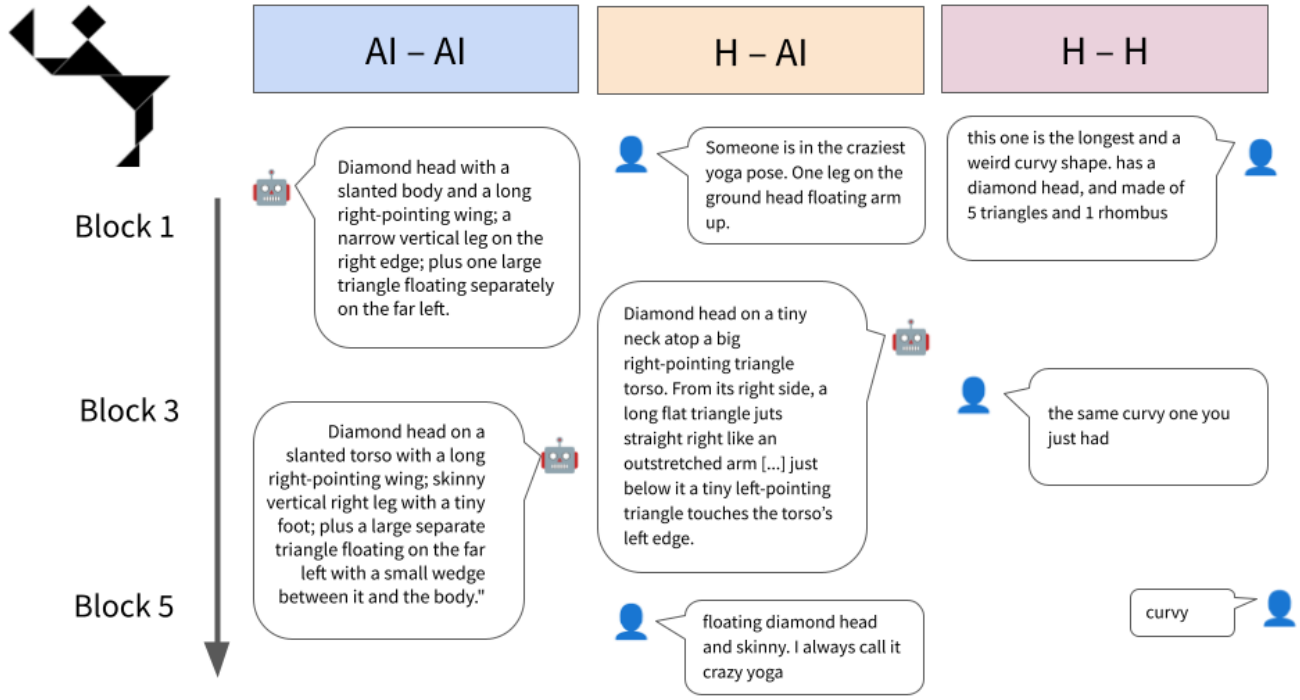

Figure 2: Examples of turns across Experiment 1 from each partnership type. AI-AI pairs (left) produce lengthy detailed geometric descriptions which do not reduce over turns. H-H pairs (right) quickly form short and analogic conventions. H-AI pairs (middle) often showed a contrast, with human partners proposing conventions that were not adopted.

Procedure Participants completed the study online. After providing informed consent, they were randomly assigned to a condition (having either a human or AI partner) and entered into a waiting room. Participants in the H-H condition were matched up with the first available partner. Participants in the H-AI condition waited for a random delay to simulate the matchmaking process. Participants were not informed whether they were interacting with a human or an AI system and the instructions used the neutral term 'partner'.

Participants completed 50 rounds with the same partner, alternating director and matcher roles (see Figure 1). Directors described a cued image and matchers attempted to select the described image from a grid. Both received feedback about which image was cued and selected after each round. Rounds were organized into 5 blocks of 10 images, so each image appeared once per block.

After all 50 rounds, participants completed an exit survey, rating their partner using a 7 point likert scale on i) how helpful they were, ii) how much they took the participants' perspective, iii) how much they used shared history, and iv) how much they adapted to the participant. Finally, participants were asked whether they thought their partner was human or an AI and gave a confidence score (0-100) and their reasoning.

AI-AI games were conducted using the exact same setup as for AI partners in human games, except that the AI system was matched with another copy of the AI system.

Results

For each trial, we measured i) whether the matcher accurately selected the intended image, ii) the length of the description in words, iii) and the Jaccard similarity between the words in the description of an image and the pair's description of the same image in the previous block (lexical overlap). Lexical overlap was computed over whitespace-tokenized, lowercased words; we did not apply stemming, lemmatization, or stopword removal. We compared conditions using mixed-effects models with random intercepts for session and target image.

As predicted, each of the three partnership types significantly increased in accuracy and lexical overlap, while decreasing in length across blocks (see Figure 3). However, these trends were qualitatively very different. H-H pairs accuracy started high (81%) and increased steadily to 93% by block 5. H-AI accuracy started much lower (54%) and increased at a similarly steady rate to H-H pairs. AI-AI accuracy also started low (56%) but shot up to 86% by Block 2 and had reached ceiling (99%) by Block 4. While the length of H-H pairs started at 20 words and dropped to 6, AI-AI pairs' messages dropped only from 58 to 56, and H-AI pairs' messages were somewhere in between. AI-AI pairs started with the highest lexical overlap and maintained this lead across all blocks. While H-AI and H-H pairs started with similar overlap (~ 0.23), H-H pairs steadily increased to 0.43 by Block 5 while H-AI pairs lagged at 0.3.

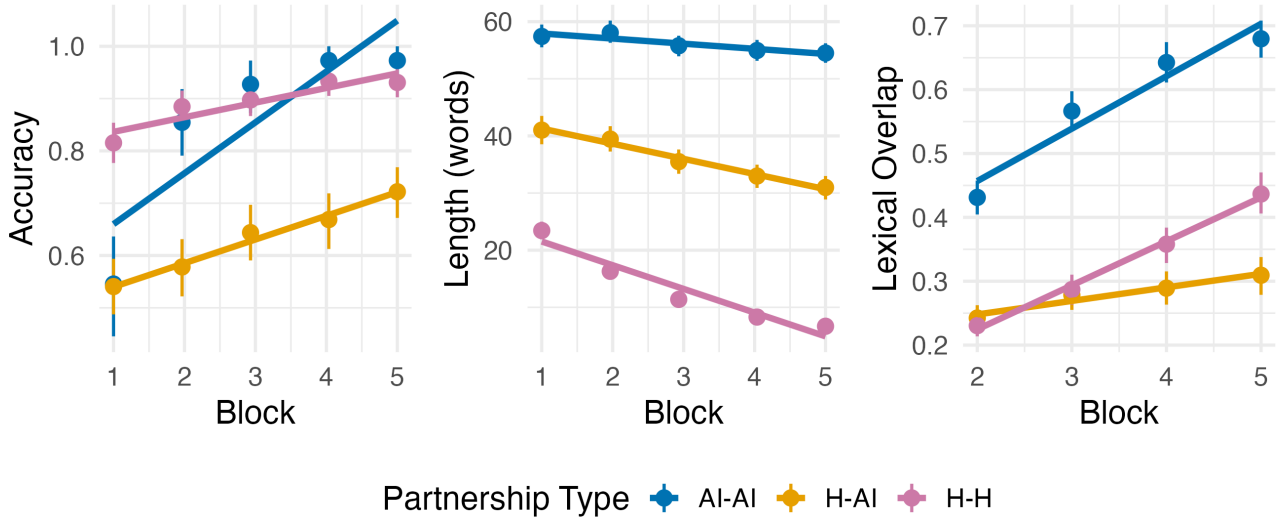

Figure 3: **Left:** Accuracy increased for all partnership types in Experiment 1. AI partnerships all started below H-H pairs, but while AI-AI pairs quickly caught up, H-AI pairs did not. **Center:** H-H messages started shorter than other pair types and also decreased fastest. **Right:** Lexical overlap between descriptions of the same object from one block to the next increased for all partnership types. H-H pairs showed more overlap than H-AI pairs but less than AI-AI pairs.

H-H pairs outperformed H-AI pairs on all three metrics, as predicted (all $p < 0.009$). However, the AI-AI to H-AI comparison produced mixed results. While H-AI messages were shorter than AI-AI pairs' ($t = -6.36$, $p < 0.001$), AI-AI pairs were both more accurate and had higher lexical overlap with previous turns (both $p < 0.001$).

As predicted, human partners rated other humans significantly higher on each of the exit questions (all $p < 0.001$). While 83% of participants believed that GPT-5 partners were AI, just 12% of participants thought this of human partners.

As an exploratory analysis, we used GPT-5-mini to tag descriptions as either analogical (using analogies such as 'the ice skater' to refer to images), geometric (e.g. describing the image in terms of shapes), or a mixture. 68% of H-H descriptions were tagged as analogical, versus 12% as geometric and a further 13% as mixed. Just 5% of AI-AI pairs' messages were classed as analogical, with 38% being classed as geometric and 57% as mixed. H-AI messages were 24% analogical, 23% geometric and 52% mixed.

In order to better understand the mechanism behind convention formation, we found the mean Jaccard similarity between descriptions of the same tangram in subsequent rounds, conditioned on whether or not the last description had been successful. Curiously, we saw significant increases in similarity for successful vs unsuccessful previous turns for all director/matcher combinations (all $p < 0.001$) except for AI directors/human matchers ($p = 0.307$). To test whether lexical overlap could be driven by convergence across (rather than within) pairs, we re-analysed the data normalizing to between-pair similarity. All effects were robust under this operationalization, suggesting lexical overlap represented within-pair idiosyncrasy.

Discussion

We set out to test whether LLM-based AI systems could establish common ground conventions in the same way that humans do. The results replicated the finding that human communicators form these conventions with one another. H-H pairs showed the hallmark increases in accuracy and lexical overlap, as descriptions shrank to just a few words (Figure 2).

In contrast, H-AI pairs failed to match human performance, in line with recent work on similar tasks (Hua & Artzi, 2024; Zeng et al., 2026). While these partnerships showed small significant increases on each of the metrics, their performance remained qualitatively poor (sending 30 word messages with only 70% success by Block 5). What accounts for this failure? We found that human speakers send longer messages to AI models than they do to other humans, which might not create the requisite downward pressure on length to facilitate conventionalization. In our exploratory analyses, we found that AI systems' descriptions were qualitatively different from humans': predominantly geometric rather than analogical. Moreover, we found that AI directors were not sensitive to the success of previous turns when producing descriptions of an image specifically when interacting with a human partner, which could account for the uniquely poor performance of H-AI pairs. Additionally, 83% of human participants could identify GPT-5 as an AI, potentially influencing their behavior.

The most surprising performance came from AI-AI pairs. In contrast to Zeng et al. (2026), we found these pairs displayed

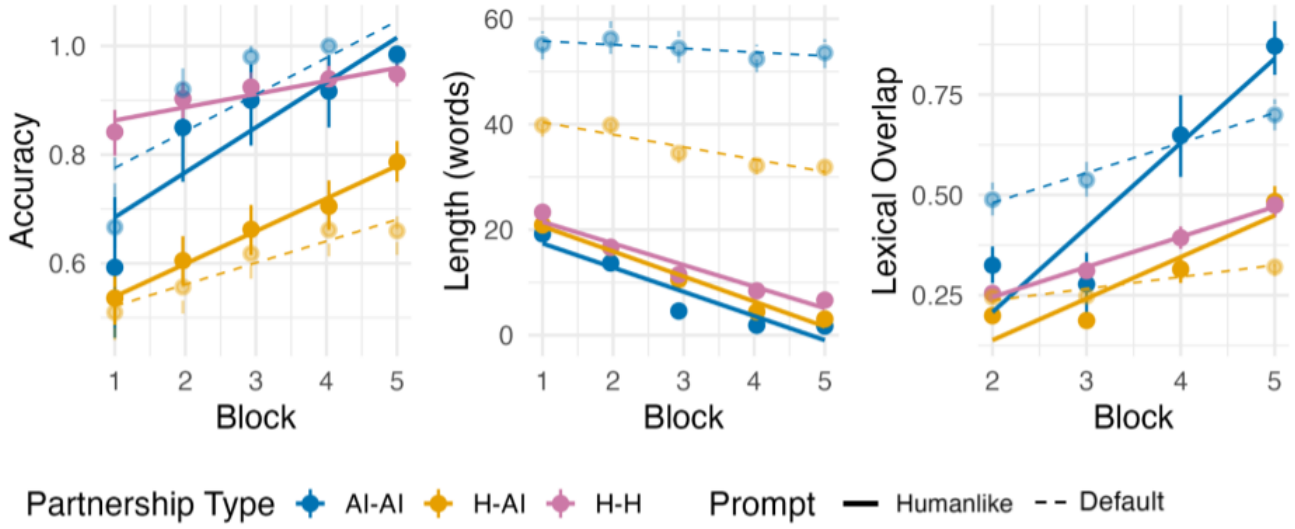

Figure 4: Experiment 2: A new "Humanlike" prompt caused both H-AI and AI-AI pairs to produce a similar trajectory in description length to H-H pairs (center). While AI-AI pairs showed all of the hallmarks of convention-formation, H-AI pairs continued to show significantly lower accuracy (left) and lexical overlap (right) than H-H pairs.

evidence of aligning to one another, but in a very different way from humans. They reached ceiling performance on accuracy within three blocks, with twice the proportion of lexical overlap as H-H pairs by Block 5. Their low accuracy in Block 1, as well as the fact that different AI-AI pairs converge on different referring expressions, suggests that AI-AI pairs are constructing genuine context-driven conventions, rather than context-free similarity between partners. On the other hand, however, the length of their messages barely dropped (from 58 to 56 words).

These results suggest that AI systems are capable of developing stable, shared forms of reference, but that they are not resource-sensitive in the same way that humans are. These results make sense with respect to models' architecture and training regime. The transformer architecture (Vaswani et al., 2017) allows models to attend to any previous token in the input, potentially allowing error-free recall of previous descriptions (Olsson et al., 2022). However, in contrast to humans who bear metabolic and opportunity costs for all of their actions, models have no intrinsic incentives to be concise.

The fact that both AI-AI pairs and H-H pairs are able to achieve high accuracy rates raises the question of why H-AI pairs perform so much more poorly. This could be due to inherent differences in how human and AI participants interpret images, or from more superficial features of their descriptions that prevent development of high lexical overlap. We address this question in Experiment 2.

Experiment 2

To test whether the differences between Human and AI players could be attributed to stylistic differences, we designed a prompt which explicitly instructed the system to behave more like humans. We used AI-AI game simulations to iteratively adjust the prompt until it produced descriptions that were qualitatively similar to human ones and showed a similar trajectory in terms of length (see Figure 2).

We replicated Experiment 1, comparing this 'Humanlike' prompt to the 'Default' prompt in the H-AI condition. We asked 3 pre-registered research questions (Jones et al., 2025b). First, does the the Humanlike prompt outperform the Default prompt (in both the convention metrics and exit survey)? This would be consistent with the theory that GPT-5 lacks the propensity, rather than the capacity, to conventionalize. Second, we asked whether H-H pairs would continue to outperform Humanlike-prompted H-AI pairs. Any residual performance gap between these conditions would point to more fundamental capacity differences between humans and LLMs. Lastly, we asked whether human participants would be better than chance at determining whether the Humanlike-prompted AI partner was not human.

Methods

The experiment proceeded identically to Experiment 1 except that participants were assigned to one of 3 types of partner: Human, AI (Default) and AI (Humanlike). AI (Default) was identical to the setup used in Experiment 1. AI (Humanlike) was also a GPT-5 model with the same parameter settings, but the prompt contained additional instructions to use analogical descriptions, to keep descriptions very short, and to lower the length of descriptions across the game. 145 participants completed the experiment, and we retained 39 H-AI Default, 40 H-AI Humanlike, and 33 H-H sessions. We additionally ran 20 AI-AI Default and 20 AI-AI Humanlike sessions using the same setup as Experiment 1.

Results

In H-AI games, the Humanlike prompt led to shorter ($t = -17.7$, $p < 0.001$) and more accurate turns ($z = 1.99$, $p = 0.04$) compared to the Default prompt (see Figure 4). Lexical overlap increased numerically but not significantly ($t = 1.54$; $p = 0.126$). The Humanlike prompt was judged significantly higher on all of the exit survey questions (all $p < 0.001$).

Compared to H-H pairs, Humanlike-prompted H-AI pairs produced messages that were not significantly longer (and in fact, numerically shorter; $t = 1.58$, $p = 0.118$). However, H-H messages showed significantly higher lexical overlap ($t = 2.98$, $p = 0.004$) and accuracy ($z = 9.34$, $p < 0.001$). Moreover, human partners outperformed the Humanlike prompted AI partners across all exit survey questions (all $p < 0.02$).

84% of participants with human partners believed they were human, while just 12% of participants thought the Default-prompted AI model was human. Partners of the Humanlike-prompted AI believed it was human 59% of the time. This rate was significantly higher than the Default prompt ($z = 3.63$, $p < 0.0001$) and crucially no different than would be expected if participants were randomly guessing ($p = 0.90$). However, it was significantly lower than the rate at which humans were judged to be human ($p = 0.008$).

Exploratory analyses showed that the Humanlike prompt produced a similar distribution of description types to H-H pairs (53% analogic, 12% geometric, 30% mixed). Moreover, the Humanlike prompt showed a significant increase in lexical overlap in cases where the previous description for that image had been successful ($p < 0.001$). However, both prompts showed large asymmetries in terms of how much one partner adapted to the other. In H-AI pairs, AI directors had significantly greater overlap with their own previous descriptions of the same tangram ($\mu = 0.51$) versus when their human partner had provided the previous description ($\mu = 0.139$, $p < 0.001$). Human speakers showed a significant but smaller self-preference effect with AI partners (0.34 vs 0.168, $p < 0.001$), and an even smaller difference with human partners (0.41 vs 0.29, $p < 0.001$). AI-AI pairs, meanwhile, showed no significant difference of whether the previous description was theirs or their partners (0.60 vs 0.57, $p = 0.104$). Moreover, while humans' descriptions tended to increase in length after failures, the strict requirement for conciseness in the Humanlike prompt caused the model to make descriptions shorter even when the same description for the last tangram had failed.

Discussion

Experiment 2 tested a prompt addressing some of the superficial differences between human and LLM behavior in Experiment 1. When LLMs were prompted in this way, H-AI pairs produced descriptions that were a similar length to the ones H-H pairs sent. However, H-AI pairs still showed lower lexical overlap and accuracy across the experiment than H-H pairs. The results rule out the account that H-AI pairs fail to form conventions because AI systems lack sensitivity to message length.

AI-AI games that used this prompt achieved very high lexical overlap and accuracy, while achieving humanlike conciseness. While the AI-AI results were not pre-registered (but part of the design of the experiment that we developed iteratively), they do demonstrate that at least one LLM-based system is capable of producing convention-forming behavior with other instances–just not with humans. This suggests that there are deeper differences between humans and LLMs which specifically limit convention formation.

General Discussion

Across two experiments, we found that same-type pairs (H-H and AI-AI) successfully formed conventions while mixed H-AI pairs consistently failed—even when models were prompted to produce humanlike behavior. The results have several implications for theories of convention-formation. First, they complicate the question of what is required to form conventions: GPT-5 instances appeared able to form conventions with one another, but not with humans. That they can develop stable conventions suggests that they possess the requisite capabilities to learn in context. The more challenging question is whether this suggests that LLMs possess deeper capacities for recursive reasoning and mental modeling that some argue are required for convention-formation (Grice, 1975; Lewis, 2008; Wilkes-Gibbs & Clark, 1992), or that these capacities are not necessary after all (Barr, 2004; Hawkins et al., 2017; Menenti et al., 2012). Future work using mechanistic approaches may inform whether humans and LLMs form conventions in similar ways.

Secondly, the results imply that forming conventions requires more than just the capacity to refer back to and repeat a partner's expressions (Brennan & Clark, 1996). While AI-AI accuracy was low in block 0 (suggesting that LLMs also failed to produce effective initial descriptions for other LLMs), AI partners were able to agree on consistent idiosyncratic conventions over later blocks while H-AI pairs were not. Perhaps this is because AI models shared the same interpretative biases from their architecture, training data, and prompt. Forming conventions appears to require some additional kind of shared interpretative stance: partners need to start with enough common ground to build on. At a high level, these results suggest that the ability to form conventions is a property of a dyad—rather than of an individual (Clark, 1996). Our results are importantly limited to just one AI system (GPT-5). Future work should investigate other models and in particular heterogeneous AI-AI pairs.

Lastly, our results imply that the social interactions which people are having with LLMs will lack the interactive alignment and development of common ground that human conversations develop. Moreover, we found that people in Experiment 2 couldn't reliably tell when they were talking to a suitably prompted AI system. This suggests that people may not be sensitive to models' systematic failings–continuing to believe that these systems are humans even when they fail in very unhumanlike ways.

Acknowledgments

We would like to thank Coefficient Giving for providing funding that supported this research, three anonymous reviewers who have

AI use. GPT-5 and GPT-5-mini (OpenAI) were used to conduct the experiment as described in Methods. Claude Opus 4.6 and 4.7 (Anthropic) were used for proofreading the manuscript.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive science*, *28*(6), 937–962.

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185–5198.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, *76*(6), 893.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How people use chatgpt* (tech. rep.). National Bureau of Economic Research.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive science*, *40*(1), 145–171.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in cognitive sciences*, *8*(1), 8–11.

Grice, H. P. (1975). Logic and conversation. *Syntax and semantics*, *3*, 43–58.

Hawkins, R. X., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *39*.

Hua, Y., & Artzi, Y. (2024). Talk less, interact better: Evaluating in-context conversational adaptation in multimodal LLMs. *Proceedings of the Conference on Language Modeling (COLM)*.

Hua, Y., Wang, E., & Artzi, Y. (2025). Post-training for efficient communication via convention formation. *Proceedings of the Conference on Language Modeling (COLM)*.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jones, C. R., & Bergen, B. K. (2025). Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*.

Jones, C. R., Bergen, B. K., Mahowald, K., & Lombardi, A. (2025a). E1: Conceptual pact formation in human-ai communication. https://doi.org/10.17605/OSF.IO/HA59Q

Jones, C. R., Bergen, B. K., Mahowald, K., & Lombardi, A. (2025b). E2: Humanlike prompting in human-ai communication. https://osf.io/zawhg

Jones, C. R., Bergen, B. K., Mahowald, K., & Lombardi, A. (2025c). Tangrams reference game. https://osf.io/wgzan/overview

Jones, C. R., Trott, S., & Bergen, B. (2024). Does reading words help you to read minds? a comparison of humans and llms at a recursive mindreading task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Lampinen, A. K., Chaudhry, A., Chan, S. C., Wild, C., Wan, D., Ku, A., Bornschein, J., Pascanu, R., Shanahan, M., & McClelland, J. L. (2025). On the generalization of language models from in-context learning and finetuning: A controlled study. *arXiv preprint arXiv:2505.00661*.

Lewis, D. (2008). *Convention: A philosophical study*. John Wiley & Sons.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in cognitive sciences*, *28*(6), 517–540.

Menenti, L., Pickering, M. J., & Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in human neuroscience*, *6*, 185.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79.

Vaduguru, S., Hua, Y., Artzi, Y., & Fried, D. (2025). Success and cost elicit convention formation for efficient communication. *arXiv preprint arXiv:2510.24023*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of memory and language*, *31*(2), 183–194.

Zeng, P., Li, W., Paige, A., Wang, Z., Kaliosis, P., Samaras, D., Zelinsky, G., Brennan, S., & Rambow, O. (2026). Lvlms and humans ground differently in referential communication. *arXiv preprint arXiv:2601.19792*.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.