"On the Same Wavelength": Exploring Human–Agent Collaboration Strategies through the Cooperative Game Wavelength

Katelyn Morrison*
kcmorris@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Gabriel Enrique Gonzalez*
gabrielenrique.gonzalez97@gmail.com
Independent Researcher
Necochea, Argentina

Zahra Ashktorab*
zashktorab@gmail.com
Independent Researcher
Yorktown Heights, NY, USA

Matthew Riemer
mdriemer@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

djallel bouneffouf
Djallel.Bouneffouf@ibm.com
IBM Research
Yorktown Heights, NY, USA

Andrew Anderson
andrew.anderson2@ibm.com
IBM Research
San Jose, CA, USA

Justin D. Weisz*
jweisz@gmail.com
Independent Researcher
Yorktown Heights, NY, USA

Abstract

Large language model (LLM) agents are increasingly embedded into workplace tools (*e.g.*, Microsoft 365) to support everyday tasks, from drafting emails to building slide decks. However, collaboration with these LLM agents can be challenging when humans and agents lack shared understanding. To investigate strategies that help human–agent teams build shared understanding, we study 24 human–agent teams playing Wavelength, a cooperative game that requires shared understanding for successful play. We compare three strategies: (1) implicit in-context learning across rounds, (2) an initial grounding conversation before gameplay, and (3) initial grounding conversations paired with post-round reflective explanations, where teammates describe how they interpreted clues and made decisions. Our preliminary analyses suggest that grounding conversations can improve team collaboration, but alone may be insufficient. Teams performed best when initial grounding conversations were paired with reflective explanations. We discuss implications for designing human–agent systems that better support shared understanding and coordination in knowledge work.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Cooperation and coordination**.

*Work done while at IBM Research.

CHI '26 Workshop on Human–Agent Collaboration, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Keywords

Human–Agent Collaboration, Cooperative Game, Empirical Study

ACM Reference Format:
Katelyn Morrison, Gabriel Enrique Gonzalez, Zahra Ashktorab, Matthew Riemer, djallel bouneffouf, Andrew Anderson, and Justin D. Weisz. 2026. "On the Same Wavelength": Exploring Human–Agent Collaboration Strategies through the Cooperative Game Wavelength. In *Proceedings of CHI '26 Workshop on Human–Agent Collaboration.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Personal agents are becoming increasingly embedded in workplace tools, supporting tasks ranging from drafting emails to facilitating brainstorming [12]. As these agents rapidly gain new capabilities, it becomes increasingly important for HCI researchers to understand how human–agent teams construct, recognize, and repair shared understanding during collaborative work [20]. Research on human–human collaboration has shown that shared understanding can be established through grounding processes [6, 11], motivating the need to design grounding processes for human–agent collaboration. Recent literature has explored approaches to personalizing an agent's behavior, including (Mutual) Theory of Mind approaches [18–21, 23], context-engineering techniques [17], and proactive agent designs [7, 13]. However, these approaches face conceptual and practical challenges. Critiques of attributing mental states to AI [4, 22], the difficulty of quantitatively assessing mental state attribution [23], and the vast design space of contextual signals that agents may draw upon make it challenging to design effective grounding strategies for human–agent teams. In deployed AI products (*e.g.*, ChatGPT, Gemini, and Copilot), users now have the ability to edit their agent's memory, offering a new, albeit still underexplored, mechanism for shaping shared understanding [10]. However, few works have explored how grounding [5, 6] and reflective explanations [9] can support human–agent teams in establishing shared understanding.

Recent work has suggested that the cooperative game Wavelength [2] may provide a suitable environment for exploring approaches that enable shared understanding in human–agent teams [16]. Building on this insight, this workshop paper presents preliminary analyses from a mixed-methods study examining how free-form conversations and reflective explanations affect the performance of 24 human–agent teams in the cooperative game Wavelength.

2 Human–Agent Collaboration in the Cooperative Game Wavelength

Wavelength [2], the "mind-reading" party game, was originally designed as a cooperative, physical board game, but digital[1] and hybrid [14] versions have also been created. In the game, players are tasked as "mind-readers" to work together to score points by predicting the location of a hidden target on a spectrum of opposing concepts (*e.g.*, hot vs. cold) based on their teammate's clue. Gameplay consists of two main phases: clue-giving and target-guessing.

During the **clue-giving** phase, each teammate receives three random spectra with corresponding target locations. The goal is to provide a clue for your teammate that conceptually maps to "*where the target range is located on the spectrum*" [2] while avoiding clues that would obviously reveal the target location (*e.g.*, through the use of spectra synonyms, ratios, or percentages). For example, for *hot vs. cold* with a target closer to the cold side, a good clue might be *snow*. Based on the Wavelength App, after simultaneously providing clues for the three spectra, the players proceed to the target-guessing phase. During the **target-guessing** phase, both players take turns guessing the target location based on their teammate's clue and seeing their teammate's guess. The game ends once all spectra are shown, with a maximum team score of 24.

2.1 User Study & Agent Implementation

We designed a between-subjects user study comparing three different collaboration strategies (implicit in-context learning, initial grounding conversations, and initial grounding with reflective explanations) in human–agent teams. The agent was implemented via prompt engineering using Llama-3-70B [8][2] as it adhered to the game rules most reliably among other LLMs publicly available at the time of our study[2]. We validated the system prompts through agent–agent tests to ensure rule-following behavior[2]. Finally, we built a lightweight Wavelength-style web interface[3] featuring our Wavelength agent as the human participants' teammate. The interface also featured different collaboration strategies, depending on the participant's assigned condition, to explore their impacts on shared understanding and performance.

The user study involved human–agent teams playing two games together, followed by a short post-study survey about teaming experience, knowledge perceptions, and collaboration perceptions. During the study, players were assigned random game names to mask the agent's identity until the study concluded. We recruited knowledge workers through internal Slack channels. Participants who completed the study were compensated an equivalent of 25 USD.

Clue-giving phase modifications. We curated a set of 17 spectra[4] and randomly sampled from this set each round (covering sports, media, science, and travel), ensuring teammates never received the same spectrum in a given game. After submitting a clue, players answered three brief questions: their confidence in giving the clue, how personalized it was to their teammate, and their *clue rationalization*–a short explanation of why they chose that clue.

Target-guessing phase modifications. After making their guess, but before seeing the result, players answered three questions: their confidence in their guess, how personalized they felt the clue was, and their *guess rationalization*–a brief explanation of why they thought their teammate gave that clue. Together with the clue rationalization, these two rationalizations form the round's reflective explanations shown at the end of the first game.

Collaboration Strategies. We modified Wavelength to foster shared understanding across three different strategies:

- ***Implicit In-Context Learning.*** The agent received no grounding or explanatory support and interacted using its default behavior. However, it could still implicitly learn and adapt across games through in-context learning.
- ***Initial Grounding.*** Drawing on grounding theory [6], the initial 13-minute, free-form conversation was designed to help teams build common ground before gameplay. AI teammates were assigned one of four personas (*e.g.*, sports enthusiast, tech-savvy individual, avid traveler, or science enthusiast) adapted from Chan et al. [3]. Each teammate received three small-talk questions selected from Aron et al. [1] to prompt their free-form conversation. Agent responses were delayed slightly to mimic human response time.
- ***Initial Grounding + Post-Game Reflective Explanations.*** Inspired by recent empirical work on co-constructed explanations [15] and self-explanations in LLMs [9], this phase shows players their teammate's clue and guess rationalizations in a pop-up modal as they review each target location and spectrum at the end of the game. Players were required to scroll to the bottom of the pop-up modal to ensure all explanations were reviewed.

3 Preliminary Findings

We present a preliminary analysis of the data collected from our user study for discussion at the workshop. We interpret higher team scores as evidence for better shared understanding. Our preliminary analyses consider 6 human–AI teams for the implicit in-context learning condition, 9 for the grounding condition, and 9 for the grounding and reflective explanation condition. We focus on how teams performed in the second game, since we would only see an effect for teams that received post-game reflective explanations in the second game score.

Quantitative Insights. Teams in the grounding with reflective explanation condition tended to score higher in the second game (M (SD) = 9.78 (5.51)) than teams with only the grounding (M (SD) = 6.67 (2.45)) or teams in the implicit in-context learning condition (M (SD) = 7.0 (4.54)), though differences were not statistically significant

[1]Wavelength app: https://www.wavelength.zone

[2]Details omitted due to space constraints.

[3]A public version of our interface with Gemini is now in *ongoing development* at http://play-wavelength-ai.com

[4]To avoid copyright issues with the official Wavelength game.

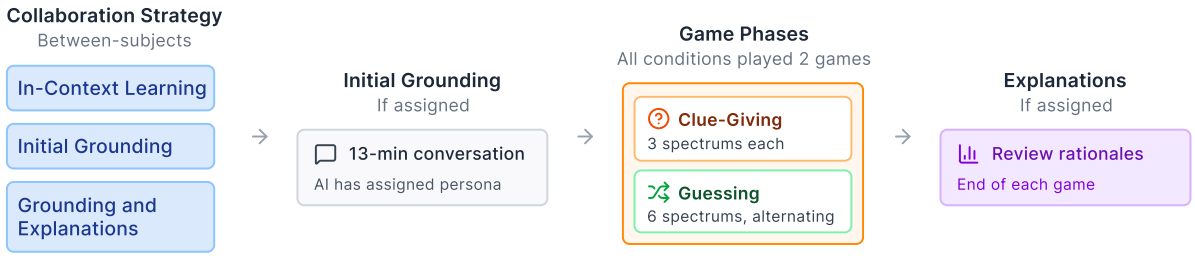

Figure 1: Our user study included one between-subjects variable: collaboration strategy (implicit in-context learning, initial grounding, and initial grounding with post-game reflective explanations). All teams are assigned to play two games together.

($p = 0.33$)[5]. The best-performing team across all teams was in the grounding with reflective explanation condition, scoring 19 out of 24 points. This team utilized the introduction time to discuss a shared game strategy. Upon reviewing the rest of the team introductions, we suspect the lack of significance we observe may stem from low conversational effort among human participants, including limited mutual questioning and difficulty establishing common ground with an AI agent that had a synthetic persona.

Qualitative Insights. Perceptions of the agent varied across conditions. In the in-context learning condition, 66.67% of participants believed their teammate was human. In contrast, all participants in the initial grounding condition correctly identified their teammate as an AI. Only one team in the grounding-with-reflective-explanations condition believed their teammate was human, and this was the team that achieved the highest score.

Despite the trends being insignificant, participants in the grounding and grounding with reflective explanations conditions reported that the **initial grounding conversations contributed to their ability to give better clues and their team's success** (P1, P5, P6, P8, P10, P12, P17, P19). P10 said "*If we did not talk at all, I'm honestly not sure we would've gotten any points*". P1 said "*Without knowing about our shared interest in sci-fi and they [sic] we both enjoy exercising (me running and them hiking) we would have done much worse*". Teams also reported that **reflective explanations helped them adjust their strategy for the second game** (P1, P5, P7). For example, P1 said, "*I realized what assumptions my teammate was making about how I think, and was able to lean into that a bit more in the second game*". Lastly, the **reflective explanations helped players understand the agent's way of thinking** (P1, P2, P5, P6, P7, P20, P21). P7 commented, "*After the first round [game] and review, I had a better idea of how my teammate was reasoning (and what they were capable of reasoning) before the 2nd game*".

Finally, we categorized rationalizations as either grounding-based, logic-based, or other. Teams scored fewer points when clues were based on individual logic rather than based on their common ground: only 9.72% of rounds with logic-based clues earned 3 or 4 points, compared to 19.44% rounds with grounding-based clues. For example, P12's clue rationalization is aligned with the AI's guess rationalization and based on their initial grounding conversation: "*I chose 'Biology' because my teammate mentioned they enjoyed biology in high school, especially the dissections, which might lead them to associate it with simple science*". AI's corresponding guess rationalization was: "*We both loved biology, I consider it to be a simple subject*".

[5]One-way ANOVA conducted across conditions.

4 Discussion, Future Work, & Conclusion

We present a between-subjects empirical study with 24 human–agent teams playing the cooperative game Wavelength, testing different collaboration strategies. While we have run a similar empirical study on human–human and agent–agent teams playing Wavelength with these various collaboration strategies, we focus specifically on analyzing the human–agent teams to ignite discussion around our findings and reflect on the implications of integrating free-form grounding conversations and reflective explanations into human–agent collaborations for knowledge work.

Initial grounding conversations can help human–agent teams establish a shared understanding. Teams that established common ground about the task beforehand collaborated more effectively, reflected in their higher team scores. This echoes Cho and Rader [5], who present that grounding can support human–agent interactions. However, not all teams in this condition established a shared understanding, often due to limited mutual questioning. This suggests agent designs should both expose their reasoning and scaffold mutual understanding. Yet, not all knowledge work tasks require deep shared understanding; future work should examine when grounding processes are most beneficial.

Initial grounding may not always be enough; exposing the agent's reasoning can offer additional benefits. The team that showed the strongest collaborative alignment experienced both grounding and post-game reflective explanations, citing that the explanations helped them better understand their teammate's reasoning. Drawing on the Mutual Theory of Mind framework [21], we interpret this as evidence that exposing model reasoning can support humans' construction and revision of shared understanding. This has implications for long-term human–AI collaboration, where agents that surface their reasoning and reflect on team dynamics may foster more effective coordination [18].

Conclusion. Our preliminary analyses highlight how different grounding processes (initial conversation grounding and post-game reflective explanations) can support shared understanding in human–agent collaboration. We invite researchers to further explore different strategies and designs to support human–agent teams.

Acknowledgments

Figma Make was used to scaffold the design for Figure 1. Grammarly and Microsoft Copilot were used throughout this manuscript to improve grammar, spelling, flow, and clarity.

References

[1] Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin* 23, 4 (1997), 363–377.

[2] BoardGameGeek. 2019. *Wavelength.* Retrieved 05-Sep-2024 from https://boardgamegeek.com/boardgame/262543/wavelength

[3] Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *arXiv preprint arXiv:2406.20094* (2024).

[4] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. [n. d.]. TOMBENCH: Benchmarking Theory of Mind in Large Language Models. ([n. d.]).

[5] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[6] HH Clark. 1991. Grounding in communication. *Perspectives on socially shared cognition/American Psychological Association* (1991).

[7] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 807–818.

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[9] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207* (2023).

[10] Brennan Jones, Kelsey Stemmler, Emily Su, Young-Ho Kim, and Anastasia Kuzminykh. 2025. Users' Expectations and Practices with Agent Memory. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[11] Robert E Kraut, Susan R Fussell, and Jane Siegel. 2001. Situational awareness and conversational grounding in collaborative bicycle repair. *Human Computer Interaction Institute. Carnegie Mellon University* (2001), 13.

[12] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–22.

[13] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang'Anthony' Chen. 2025. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[14] Chelsea Mills, Denise Y Geiskkovitch, Carman Neustaedter, William Odom, and Benett Axtell. 2023. Remote Wavelength: Design and Evaluation of a System for Social Connectedness Through Distributed Tabletop Gameplay. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[15] Dimitry Mindlin, Meisam Booshehri, and Philipp Cimiano. 2025. Towards Co-Constructed Explanations: A Multi-Agent Reasoning-Based Conversational System for Adaptive Explanations. In *Proceedings of the 13th International Conference on Human-Agent Interaction*. 148–157.

[16] Katelyn Morrison, Zahra Ashktorab, Djallel Bouneffouf, and Gabriel Enrique. 2025. Establishing the Cooperative Game Wavelength as a Testbed to Explore Mutual Theory of Mind. *ToM4AI 2025* (2025), 54.

[17] Prithvi Rajasekaran, Ethan Dixon, Carly Ryan, and Jeremy Hadfield. 2025. Effective Context Engineering for AI Agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents. Accessed: 2026-01-26.

[18] Jonan Richards and Mairieli Wessel. 2024. What You Need is What You Get: Theory of Mind for an LLM-Based Code Understanding Assistant. *arXiv preprint arXiv:2408.04477* (2024).

[19] Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

[20] Qiaosi Wang. 2024. *Mutual theory of mind for human-AI communication in AI-mediated social interaction.* Ph. D. Dissertation. Ph. D. Dissertation. Georgia Institute of Technology.

[21] Qiaosi Wang and Ashok K Goel. 2022. Mutual theory of mind for human-ai communication. *arXiv preprint arXiv:2210.03842* (2022).

[22] Tolga Yıldız. 2025. The minds we make: A philosophical inquiry into theory of mind and artificial intelligence. *Integrative Psychological and Behavioral Science* 59, 1 (2025), 10.

[23] Xiaoyun Yin, Elmira Zahmat Doost, Shiwen Zhou, Garima Arya Yadav, and Jamie C. Gorman. 2025. When Researchers Say Mental Model/Theory of Mind of AI, What Are They Really Talking About? arXiv:2510.02660 [cs.HC] https://arxiv.org/abs/2510.02660

Received 15 February 2026; accepted 4 March 2026