Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

Wavelength.AI: Extending the Collaborative Game Wavelength as a Testbed for Studying Shared Understanding in Human–Agent Collaboration

Katelyn Morrison[1], **Gabriel Enrique Gonzalez**[2*], **Zahra Ashktorab**[3*], **Matt Riemer**[2], **Andrew Anderson**[2], **Djallel Bouneffouf**[2] and **Justin D. Weisz**[2*]
[1]Carnegie Mellon University
[2]IBM Research
[3]Microsoft
kcmorris@alumni.cmu.edu, zashktorab@microsoft.com, djallel.bouneffouf@ibm.com

Abstract

AI's increasing role as a personal agent assisting knowledge workers in everyday tasks underscores the need to investigate how to help human–agent teams build a shared understanding. We extend the collaborative "mind-reading" game Wavelength to include an AI teammate, presenting the first demonstration of an LLM capable of playing this game. Based on our agent–agent play experiments, we developed Wavelength.AI, which implements two strategies to support shared understanding: an initial team grounding conversation and post-game reflective explanations. We interpret higher team scores as evidence for better shared understanding in a preliminary user study with 24 human–AI teams. Our findings reveal that Wavelength.AI can help researchers evaluate and design different strategies to shape human-agent teams' shared understanding. Human players can see if they are on the same wavelength with AI and view our demo video today at https://play-wavelength-ai.com.

1 Introduction

Personal agents are becoming increasingly embedded in workplace tools, supporting tasks ranging from drafting emails to facilitating brainstorming [Lee *et al.*, 2025]. 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 [Wang, 2024]. Research on human–human collaboration has shown that shared understanding can be established through grounding processes [Clark, 1991; Kraut *et al.*, 2001], motivating the need to design grounding processes for human–agent collaboration. In efforts towards establishing shared understanding, recent literature has explored approaches to personalizing the agent's behavior, including (Mutual) Theory of Mind approaches [Richards and Wessel, 2024; Wang and Goel, 2022; Ullman, 2023; Yin *et al.*, 2025; Wang, 2024], context-engineering techniques [Rajasekaran *et al.*, 2025], and proactive agent designs [Deng *et al.*, 2024; Liu *et al.*, 2025]. However, these approaches face conceptual and practical challenges. Critiques of attributing mental states to AI [Yıldız, 2025; Chen *et al.*, 2024], the difficulty of quantitatively assessing mental state attribution [Yin *et al.*, 2025], 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 [Jones *et al.*, 2025]. However, few works have explored how grounding [Cho and Rader, 2020; Clark, 1991] and reflective explanations [Huang *et al.*, 2023] can support human–agent teams in establishing shared understanding.

Recent work has suggested that the cooperative game Wavelength [BoardGameGeek, 2019] may provide a suitable environment for exploring approaches that enable shared understanding in human–agent teams [Morrison *et al.*, 2025]. Building on this insight, this workshop paper extends the collaborative game Wavelength, presenting Wavelength.AI: a testbed for studying shared understanding in human–agent collaborations. Along with contributing the first system to successfully implement an LLM agent to play Wavelength with a human, we present preliminary analyses from a mixed-methods study examining how free-form conversations and reflective explanations affect the performance of 24 human–agent teams.

2 Wavelength.AI System

Wavelength [BoardGameGeek, 2019], the "mind-reading" party game, was originally designed as a cooperative, physical board game, but digital[1] and hybrid [Mills *et al.*, 2023] 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 clue-giving and target-guessing.

During the **clue-giving**, 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*" [BoardGameGeek, 2019] while avoiding clues that

*Work performed while at IBM Research.

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

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.

Clue-giving phase modifications. We curated a set of 17 spectra to avoid copyright issues with the official Wavelength game. We based the spectra on adjectives and nouns from [Nguyen *et al.*, 2016] (covering sports, media, science, and travel) and randomly sampled from this set each round, 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.

2.1 Interface Implementation

Frontend. The front end, built with Svelte, is designed to mimic the Wavelength mobile application [BoardGameGeek, 2019]. Before starting gameplay, human players can create an account using an email and password or by connecting through their Google account. When starting a new game with a Wavelength.AI agent, players can choose from three different gameplay options: Wavelength Basic, Wavelength Relational, and Wavelength Retrospective. The difference between the three gameplay options is described below. All gameplay options feature a similar interactive game board for clue-giving (left side of Figure **1**) and target-guessing (middle of Figure **1**). Lastly, the frontend displays a game summary (right side of Figure **1**) for players to review.

Backend. The back end, implemented with Python, FastAPI, and Google Firebase, supports LLM API calls and data collection. To measure team performance, the system tracks points scored per round by calculating angular differences between guesses and the target slices. Angular differences are then matched to a score of +0, +2, +3, or +4 points based on how far away the guess is from the target.

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

In-Context Learning. The agent receives no grounding or explanatory support and interacts using its default behavior. The agent can implicitly learn and adapt across games through in-context learning [Brown *et al.*, 2020].

Initial Grounding. Drawing on grounding theory [Clark, 1991], 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 [Ge *et al.*, 2024]. Each teammate received three small-talk questions selected from [Aron *et al.*, 1997] 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 [Mindlin *et al.*, 2025] and self-explanations in LLMs [Huang *et al.*, 2023], 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.

2.2 Agent Implementation

The AI Player. The AI player in Wavelength.AI is built using a large language model, as research has shown that LLMs are capable of playing cooperative games [Hu *et al.*, 2024; Stephenson *et al.*, 2024; Ruhdorfer *et al.*, 2024]. To develop an AI player suitable for playing Wavelength against a human, we sought an LLM that could competently play the game and adapt its strategy based on information about its teammates. We performed prompt engineering through several AI-AI games with four popular LLMs[2] (Mixtral-8x7B [Jiang *et al.*, 2024], Llama2-70B [Touvron *et al.*, 2023], Llama3-70B [Dubey *et al.*, 2024], and Granite-13B [Research, 2024]). We then evaluated their clue-tailoring capabilities by first creating several personas based on the dataset in [Ge *et al.*, 2024]. These experiments yielded a final instruction prompt and four personas that enabled competent gameplay with Llama3-70B. The current demo implements Gemini 2.5 and performs similarly to our initial evaluations with Llama3.

Agent Persona. To enable the LLM to establish common ground with its human teammate, we crafted several different personas for the LLM to adopt. These personas established specific topics and domains that players could probe during the introduction phase to craft their mental model of their teammate. Without a persona, it would be difficult for an LLM to express an opinion on a question posed to it.

We generated four distinct personas, each catering to one of the categories in our spectrum: Travel, Technology, Science, and Sports/Media. To create these personas, we randomly selected four personas from the public dataset provided by [Ge *et al.*, 2024]. We only considered personas that began with "I am" to ensure each persona represented an individual rather than a group or organization. The four personas are:

- *I am a picky Canadian movie lover who enjoys crime comedies but has a soft spot for classic, full structured films.*
- *I am a writer and nature lover who is captivated by the remote and pristine landscapes of the world.*

[2]These LLMs were state-of-the-art publicly available LLMs at the time of our evaluation study. The demo now uses Gemini 2.5.

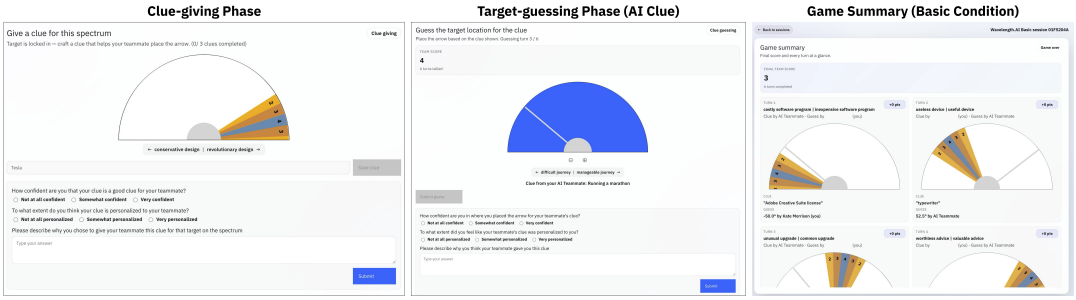

Figure 1: The clue-giving and target-guessing phase designs for human players. Both phases ask human players three similar questions about their confidence and rationalizations for data analysis. A game summary is shown at the end of the game.

- *I am an American football fan who doesn't have interest in soccer and European sports and gets easily overwhelmed by complicated tournament structures.*
- *I am a mobile app developer frustrated with the limitations and quirks of integrated development environments and logging mechanisms.*

We expanded these personas to be more detailed and include opinions related to our spectra using LLama3-70B. After finalizing the detailed personas, we construct a prompt for one instance of Llama3-70B to leverage one persona and another instance of Llama3-70B to follow another persona during clue-giving and guessing. We run another 1,000 games to ensure Llama3-70B can tailor clues based on a given teammate's persona. Both models are provided with context about their own and their teammate's personas during clue-giving and guessing. The average distance from these 1000 games (M (SD) = 43.3 (16.0) degrees of difference between prediction and ground truth) is not significantly different from the average distance of Llama3-70B without personas (M (SD) = 42.5 (15.1) degrees of difference between prediction and ground truth), suggesting that the model can successfully play the game while adapting to personas, $t(1, 998) = 1.19, p > 0.1$. We also manually inspected clues during gameplay to ensure they were tailored to each teammate.

With the interface's recent upgrade to Gemini, we have not modified any of the system prompts or personas used for the agent. Through our manual gameplay inspections, we have verified that the Gemini 2.5 in the latest interface deployment follows the game's rules, similar to how Llama3-70B did. We have also ensured that Gemini 2.5 leverages context from introductions and previous gameplay during clue-giving and target-guessing.

3 Preliminary User Evaluation

We present a preliminary analysis of the data collected from human–agent (Llama3-70B) teams. 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 in-context learning condition (M (SD) = 7.0 (4.54)), though differences were not statistically significant ($p = 0.33$)[3]. 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 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. 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 that had post-game summaries reported that the 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). For example, P7 commented, "*After the first round and review, I had a better idea of how my teammate was reasoning (and what they were capable of reasoning) before the 2nd game*".

4 Conclusion

Overall, we present a demo of the first successful implementation of an LLM playing the cooperative game Wavelength. We use this demonstration as a lens to studying different strategies that can help human–agent teams establish a shared understanding. We look forward to adapting the interface to enable researchers to test their own LLMs and strategies.

[3]One-way ANOVA conducted across conditions.

Acknowledgments

Grammarly and Copilot were used to improve the clarity, spelling, and grammar of this manuscript. Codex was used to assist in interface implementations.

References

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

[BoardGameGeek, 2019] BoardGameGeek. Wavelength, 2019.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chen *et al.*, 2024] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. Tombench: Benchmarking theory of mind in large language models. pages 15959–15983, 2024.

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

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

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

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

[Ge *et al.*, 2024] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

[Hu *et al.*, 2024] Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*, 2024.

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

[Jiang *et al.*, 2024] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[Jones *et al.*, 2025] Brennan Jones, Kelsey Stemmler, Emily Su, Young-Ho Kim, and Anastasia Kuzminykh. Users' expectations and practices with agent memory. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2025.

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

[Lee *et al.*, 2025] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 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*, pages 1–22, 2025.

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

[Mills *et al.*, 2023] Chelsea Mills, Denise Y Geiskkovitch, Carman Neustaedter, William Odom, and Benett Axtell. 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*, pages 1–19, 2023.

[Mindlin *et al.*, 2025] Dimitry Mindlin, Meisam Booshehri, and Philipp Cimiano. 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*, pages 148–157, 2025.

[Morrison *et al.*, 2025] Katelyn Morrison, Zahra Ashktorab, Djallel Bouneffouf, and Gabriel Enrique. Establishing the cooperative game wavelength as a testbed to explore mutual theory of mind. *ToM4AI 2025*, page 54, 2025.

[Nguyen *et al.*, 2016] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *arXiv preprint arXiv:1605.07766*, 2016.

[Rajasekaran *et al.*, 2025] Prithvi Rajasekaran, Ethan Dixon, Carly Ryan, and Jeremy Hadfield. Effective context engineering for ai agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents, September 2025. Accessed: 2026-01-26.

[Research, 2024] IBM Research. Granite foundation models. Technical report, IBM Research, May 2024.

[Richards and Wessel, 2024] Jonan Richards and Mairieli Wessel. What you need is what you get: Theory of mind for an llm-based code understanding assistant. *arXiv preprint arXiv:2408.04477*, 2024.

[Ruhdorfer *et al.*, 2024] Constantin Ruhdorfer, Matteo Bortoletto, Anna Penzkofer, and Andreas Bulling. The overcooked generalisation challenge. *arXiv preprint arXiv:2406.17949*, 2024.

[Stephenson *et al.*, 2024] Matthew Stephenson, Matthew Sidji, and Benoît Ronval. Codenames as a benchmark for large language models. *arXiv preprint arXiv:2412.11373*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

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

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

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

[Yin *et al.*, 2025] Xiaoyun Yin, Elmira Zahmat Doost, Shiwen Zhou, Garima Arya Yadav, and Jamie C. Gorman. When researchers say mental model/theory of mind of ai, what are they really talking about?, 2025.