Human Alignment: How Much Do We Adapt to LLMs?

Tanguy Cazalets*, **Ruben Janssens***, **Tony Belpaeme, Joni Dambre**
IDLab-AIRO, Ghent University - imec, Ghent, Belgium
{tanguy.cazalets,ruben.janssens,tony.belpaeme,joni.dambre}@ugent.be

Abstract

Large Language Models (LLMs) are becoming a common part of our lives, yet few studies have examined how they influence our behavior. Using a cooperative language game in which players aim to converge on a shared word, we investigate how people adapt their communication strategies when paired with either an LLM or another human. Our study demonstrates that LLMs exert a measurable influence on human communication strategies and that humans notice and adapt to these differences irrespective of whether they are aware they are interacting with an LLM. These findings highlight the reciprocal influence of human-AI dialogue and raise important questions about the long-term implications of embedding LLMs in everyday communication.

1 Introduction and Related Work

Large Language Models (LLMs) enable AI systems to approximate human-like dialogue, significantly expanding the possibilities for human-computer interaction. Their capabilities have become integral to modern life, supporting applications such as educational platforms (Kasneci et al., 2023), physician assistants (Thirunavukarasu et al., 2023), mental well-being support (Ma et al., 2024), and generally extremely personalized user interfaces (Chen et al., 2024). More and more, they are becoming a pervasive presence in our personal worlds, which we interact with in a social manner. However, we don't yet know much about how we adapt to them in these social interactions.

Although many studies focus on how to adapt these models to human needs—through fine-tuning, bias mitigation, or personalization (Navigli et al., 2023; Gallegos et al., 2024; Shum et al., 2018; Ouyang et al., 2022)—fewer have examined how humans adjust their own behavior when interacting with AI (Shen et al., 2024; Woodruff et al., 2024; Floridi and Chiriatti, 2020).

We do know that humans continuously adapt to their conversation partners when communicating. After all, human communication is not simply a passive exchange of information; rather, it is a highly adaptive process (Clark and Brennan, 1991; Ghaleb et al., 2024). In human-human interactions, speakers often engage in *interactive alignment* or *grounding*, converging on vocabulary, syntax, and discourse strategies to optimize clarity and efficiency (Pickering and Garrod, 2013). These adaptations reduce cognitive load and help establish common ground (Clark and Brennan, 1991), thus improving the effectiveness of interpersonal communication. Recent work in cognitive neuroscience even indicates that electrical oscillations in human brains synchronize during meaningful social interactions (Lindenberger et al., 2009; Valencia and Froese, 2020).

It follows logically that humans also adapt to LLMs when interacting with them. If we find that humans consistently shift their language patterns to accommodate AI, this shift may have far-reaching implications for cognition, creativity, and social norms, as previously noted in human-human alignment research (Pickering and Garrod, 2004). Exploring this relationship necessitates a broader interdisciplinary approach, drawing insights from psychology, linguistics, cognitive science, and ethics.

Some research has already investigated how humans adapt to LLMs. This research has largely centered on higher-level cognitive processes such as idea generation (Petridis et al., 2023), scientific writing (Shen et al., 2023), and ethical reasoning (McDonald and Pan, 2020). However, it remains unclear how individuals adapt the lower levels of their cognition to LLMs, such as the language they use and their behaviour in social interactions.

Methodologically, capturing and quantifying mutual adaptation in verbal interaction poses unique

*Equal contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 603–613

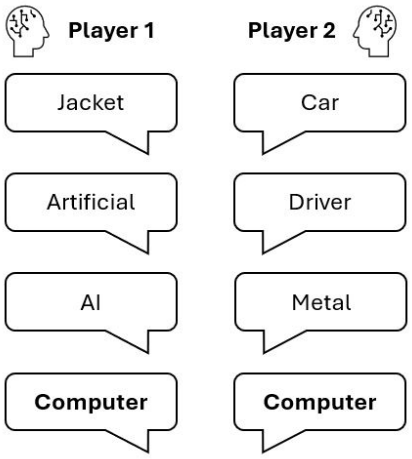

Figure 1: Example of the Word Synchronization Challenge, where participants converge on the same word by the fourth turn.

challenges. While only few studies have looked at how humans adapt to AI systems, alignment in human-human interactions has been a long-standing topic of interest to researchers. However, experimental research in this field has usually studied how humans align their language in reference to some visual information (Garrod and Doherty, 1994; Branigan et al., 2000; Ivanova et al., 2020), while not all human-LLM interactions have visual context. The studies are also limited to lexical and syntactic alignment: they do not study the dynamics of the social interaction.

Yet, social interactions exist of much more than the words that are spoken. Through their choice of words, interlocutors in dialogue share control over the flow of the dialogue. Central to this process is the ability to simulate and predict the other's utterances—part of social cognition (Gandolfi et al., 2022). Do humans use their social cognitive abilities to share control of dialogue with an LLM? And when they do, do they change their behavior compared to when interacting with another human?

1.1 Contribution

In order to study how humans adapt their social behavior and language use to LLMs, we employ a simple language game: the Word Synchronization Challenge (WSC). This game, illustrated in Figure 1, is a multi-turn task where each of two participants (human or artificial) writes down a word, revealed simultaneously at the end of each turn. They aim to converge on the same word as quickly as possible, while not being allowed to use any word previously used by either participant. The game was recently introduced by Cazalets and Dambre (2025), who used it to study LLM-LLM adaptation, but it is also known as an improvisational theater exercise called Convergence or Mind Meld (Hall, 2014), similar to prior work in natural language processing and cross-cultural inference that was inspired by the game Codenames (Kim et al., 2019; Shaikh et al., 2023).

This game constitutes an extremely simple social interaction, not relying on any other modality than verbal interaction, but it requires the two players to coordinate by simulating each other's word associations and aligning their word choices. Convergence in fewer turns is indicative of stronger mutual alignment. The word choices themselves can also be studied: through analyzing the similarity of the chosen words, and the relationships between them, we can study how both players adapt to each other. Furthermore, we study whether any difference in alignment behavior is due to the behavior of the LLM, or because the human is aware they are communicating with an AI model.

In this paper, we address the need to quantify human adaptation to LLMs through the following contributions: (1) introducing the Word Synchronisation Challenge as experimental paradigm to study human-LLM adaptation, (2) studying the extent to which humans align word choices differently with LLMs than they do with humans, (3) studying to which extent this difference is due to the human's awareness of the artificial nature of the LLM, and (4) discussing the potential ethical ramifications for designing AI systems that preserve the richness of human language and cognition.

2 Methods

2.1 Experimental Design

We set up a study where human participants played the Word Synchronization Challenge with both other human players and an LLM. The study used a within-subjects 2x2 factorial design, where we manipulated two factors: whether the participant played against a human or an LLM, and whether the partner was shown to be a human or LLM. This yields the following four conditions:

1. **vs-Human (Human shown):** partner was shown as a human and was indeed a human.

2. **vs-Human (AI shown):** partner was shown as an AI but was in fact a human.

3. **vs-LLM (AI shown):** partner was shown as an AI and was indeed an LLM.

4. **vs-LLM (Human shown):** partner was shown as a human but was in fact an LLM.

Participants completed 4 games per condition (16 games total), with the order of conditions randomized, enabling us to disentangle the effects of actual versus perceived partner identity.

2.2 LLM Implementation

We used OpenAI's GPT-4o model to generate the AI partners responses. The prompt was designed to ensure that the LLMs responses felt natural in the context of the game. In the first round, the prompt encouraged a creative yet random word choice, while subsequent rounds used a dynamic prompt that referenced previous words. Detailed information about the prompt templates, including the rationale behind the design of the prompts, and model settings, is provided in Appendix B.

2.3 Participants

Participants were recruited via Prolific, ensuring a diverse sample of L1 English speakers located in the United Kingdom. A total of 20 participants (6 identified as male, 12 female, and 2 who preferred not to disclose gender; mean age = 34.2 years, SD = 13.05) were enrolled. Participants were compensated GBP 6.90 for their participation, with a median completion time of 48 minutes and 1 second. This payment was considered adequate based on the prevailing market rates in the United Kingdom.

2.4 Procedure

Participants were informed when starting the study that they would complete 16 games with a human or AI player in random order (see instructions in Appendix C). In each game, both players initially entered a random word. In subsequent rounds, they submitted a new, unused word simultaneously, with the game concluding once both players entered the same word, or after a maximum of 16 rounds.

2.5 Post-Game Questionnaire

After each game, participants completed a short questionnaire assessing their experience and strategy use. They rated their partners performance, perceived strategy, and mutual understanding on a 5-point scale (1 representing the lowest performance and 5 the highest). Additionally, they reported their sense of connection with their partner. These self-reported measures complemented our behavioral and linguistic data.

2.6 Ethical Considerations and Data Handling

The study was conducted in accordance with the General Ethical Protocol for research with human participants of our institution, and all data were stored securely. Data were anonymized by assigning each participant a unique randomly generated `playerId`. No personally identifiable information (e.g., IP addresses) was collected.

3 Results and Analysis

3.1 Dataset Filtering and Cleanup

We filtered the data to remove incomplete sessions and other anomalies (e.g., games completed in 2 or fewer rounds, as these were indicative of users repeating a previously used word pattern). The resulting dataset comprised 89 valid H-vs-H games and 139 valid H-vs-LLM games (see Table 1 for details).

3.2 Convergence Metrics

Condition	N	Avg. rounds	Win Rate
vs-LLM (AI shown)	72	8.5	75%
vs-LLM (H shown)	67	8.3	67%
vs-LLM (all)	139	8.4	72%
vs-H (AI shown)	39	6.0	79%
vs-H (H shown)	50	6.8	76%
vs-H (all)	89	6.4	78%

Table 1: Summary of valid games analyzed. We abbreviate Human as H and Artificial Intelligence as AI.

A first high-level indicator is how often the participants successfully converged within 16 rounds and, if they did, how many rounds they needed. Table 1 displays both metrics. A χ^2 test did not reveal any significant differences between the success rates of the four conditions ($p = .63$) or between all human-human and human-LLM games ($p = .35$).

However, when comparing the convergence time for successful games, a Mann-Whitney U-test shows a significant difference ($p < 0.01$) between all human-LLM and human-human games. Within these games, the Mann-Whitney U-tests did not show statistical differences between whether the LLM ($p = 0.64$) or the human partner ($p = 0.27$) were portrayed as AI or human.

3.3 Strategy Analysis

To quantify the convergence strategies used by the participants, we used a linguistic analysis of the

relationships between the words used, and a subjective assessment by the participants themselves. We also qualitatively discuss the trajectory found in one example game in detail, to illustrate the convergence strategies.

3.3.1 Conceptual Linking Score

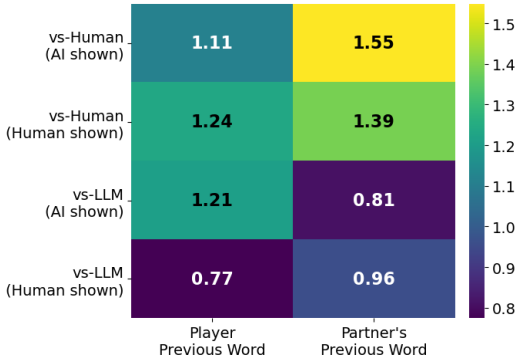

Figure 2: Average CL scores. Each cell represents the average score from word to previous word within a given game configuration.

We computed a Conceptual Linking (CL) score by querying the ConceptNet API (Speer et al., 2017). This score is intended to capture, each round, how semantically related a players current word is to either their own previous word or to their partners previous word. ConceptNet provides weighted associations, and the CL score corresponds to the highest weight found (0 if none). A higher score indicates stronger thematic or conceptual continuity between word choices.

For each game, we computed the average CL scores to both the player's and the partner's word from the previous round, over all rounds. Those averages were then averaged across all games within each configuration, and presented in Figure 2.

While the Mann-Whitney U-test did not reveal any statistical differences between the human-human and human-LLM games for CL score with the player's own previous word ($p = 0.27$), it did show a significant difference when looking at the partner's previous word ($p < 0.001$). No significant differences were found between whether the partner was portrayed as AI or human.

3.3.2 User-perceived Partner Strategy

Following the framework of Cazalets and Dambre (2025), a post-game questionnaire asked participants about their perception of their partners strategies, asking them to choose between: *mirroring* (choosing a word close to the partner's previous word), *staying close* (choosing a word close to their own previous word), or *averaging* (choosing a word halfway between the two previous words).

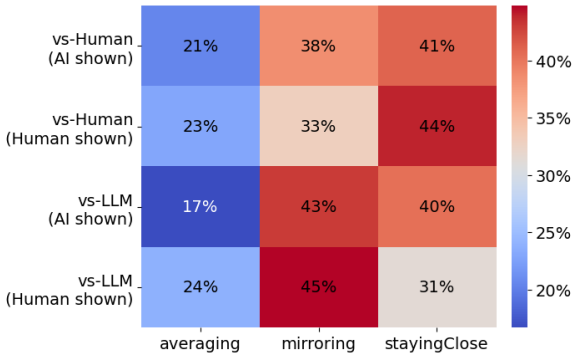

Figure 3: Average reported strategy measures by game configuration. Each cell shows the percentage of time a given strategy was attributed to other player for each game configuration.

Figure 3 shows the average aggregated results of the strategies as reported by the players. A χ^2 test between the four conditions did not reveal a significant difference in user-perceived strategies ($p = 0.56$), and neither when comparing human-human interactions with human-LLM interactions, regardless of how the partner is presented to the player ($p = 0.33$).

3.3.3 Qualitative Illustration of Convergence Trajectory

Figure 4 presents both the word-by-word interaction between a human player and an LLM during a game and the corresponding trajectory of their exchanges in semantic space, with embeddings calculated by word2vec (Mikolov et al., 2013).

This specific game illustrates how both agents adapt their choices based on each other's previous moves. For example, when the player shifts from "sunshine" to "stairs" and "step", the LLM responds with semantically related locations like "rays", and "basement", gradually bridging concepts associated with light, darkness, and structure. In round 3, the LLM's choice of "basement" and the player's move to "dark" signals an attempt to align on the theme of underground, less illuminated spaces. In this example, both partners ultimately settle on the shared concept of "door," indicating a point of semantic agreement.

By projecting the embedding trajectories of these exchanges into three dimensions using PCA (shown in the bottom three views of the Figure 4), we observe how the players word selections nav-

Player	sunshine ☀️	stairs 🪜	step 🪜	dark ⚫	loft 🏠	hatch 🏠	door 🏠
LLM	cellar 🏠	rays ☀️	basement 🏠 ⚫	ladder 🪜 🏠	shadow ⚫	attic 🏠	door 🏠

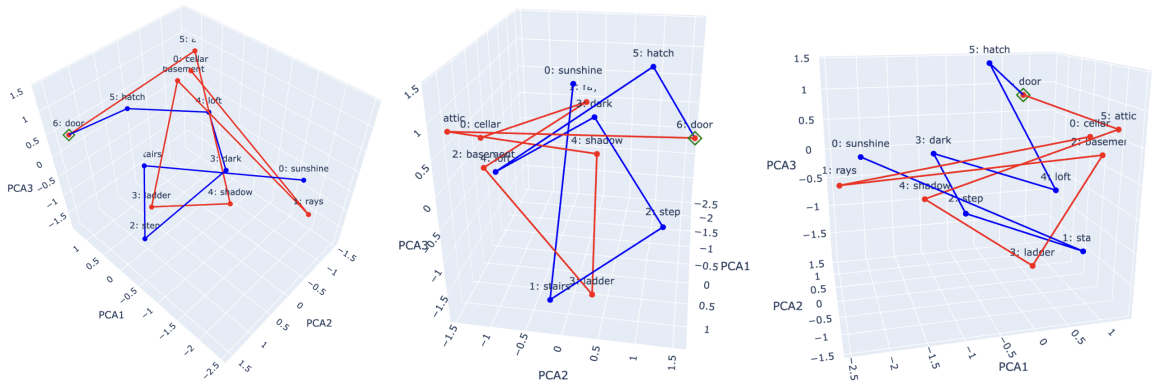

Figure 4: (Top) Table showing the sequence of words exchanged during a game between a Player and an LLM, color-coded by semantic group. (Bottom) Three different views of the projection of the embedding of one game between a human (blue) and a LLM (red). The final word is highlighted with a diamond shape.

igate through semantic clusters—such as moving from themes of light (sunshine, rays) and elevation (stairs, ladder) towards enclosed or connected spaces (loft, attic, door). The visualization highlights periods of alignment, thematic switching, and convergence attempts.

This interaction shows that LLMs adapt and attempt to synchronize with human players over successive rounds, as the human player does with the LLM, and illustrates the strategic, adaptive behaviors emerging from the WSC.

4 Discussion and Conclusion

This study investigated how humans adapt their behavior in social interactions to LLMs. We employed the Word Synchronisation Challenge (WSC) as a "minimal", very simple verbal social interaction, which requires two players to align their word choices, employing social cognition capabilities such as simulating the other's word associations and sharing control of the interaction.

Our results provide evidence that humans do change their behavior in a social interaction with an LLM, compared to how they behave when interacting with another person. Players converged in significantly fewer turns when playing with a human than when playing with an LLM. Furthermore, we analyzed which strategies the players employ in order to converge. Quantifying the semantic similarity between the chosen words through CL scores, we saw that humans changed how they chose their words depending on their partner: when playing against an LLM, they chose words that were significantly less similar to their partner's previous word, compared to when playing against another human. This shows that the difference in convergence rate is linked with a difference in alignment behavior. This difference could be explained by players noticing that the LLM behaves differently—moving more towards their word than a human would—and reacting to this by choosing to stay close to their own word and letting the LLM converge to them. Interestingly, none of these metrics differed when their partner was portrayed as AI or human, indicating the change in adaptation behavior is a reaction to the LLM's behavior rather than to the perception that it is artificial.

In conclusion, our study shows that humans adapt differently to LLMs than to humans in at least some interactions, and that human adaptation happens irrespective of whether they are aware they are interacting with an LLM. With these results in a "minimal social interaction", we make a case for future, deeper research investigating the dynamics of how humans adapt to LLMs, which we believe is an under-researched area of significant importance. As AI systems become increasingly integrated into daily communication, understanding these bidirectional effects is crucial for designing technologies that enhance, rather than constrain, the diversity of human communication. More research is needed, and should focus on the long-term cognitive, social, and cultural implications of these shifts, informing both technological innovation and policy decisions aimed at fostering a balanced co-evolution of human and artificial communicative practices.

Acknowledgments

This project has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 860949, the Flemish Government (AI Research Program), and the Horizon Europe VALAWAI project (grant agreement number 101070930).

Limitations

Our study faces several limitations that warrant consideration. First, the sample size is relatively small and restricted to a narrow demographic, potentially limiting the generalizability of our findings. Variations in individual linguistic proficiency and cultural background may still introduce confounds, suggesting that larger samples are necessary to validate the robustness of these effects.

One notable limitation of our study is inherent to the specificity of the Word Synchronization Challenge. This highly controlled task is both a bug and a feature: while its constrained nature may limit the generalizability of our findings to more spontaneous or naturalistic settings, it also enables precise quantification of alignment effects that might otherwise be obscured in less structured interactions. Our experimental design—though rigorously controlled—cannot fully capture the wide range of spontaneous, real-world conditions under which human-AI dialogue occurs. In particular, the short duration of the Word Synchronization Challenge may not reflect the complexities of natural conversation or the long-term evolution of shared linguistic habits, which could influence both the emergence and persistence of alignment phenomena over time.

It should be noted that LLMs are known to have vocabularies and frequency profiles that diverge from those of human speakers (Yakura et al., 2024). Part of the slower convergence observed with LLM partners may simply reflect this underlying distributional mismatch, rather than the social dynamics we aim to study. However, such differences in vocabulary distributions are inherent to any pair of speakers. Even among humans, active vocabularies can differ substantially, and the aim of the WSC is explicitly to encourage interlocutors to bridge such lexical gaps and to study how they respond to this mismatch by adapting to each other. Our quantitative and qualitative analysis of the adaptation strategies show there is evidence of at least some interactive alignment, rather than the slower convergence only being a result of different vocabulary distributions. Yet, future research should further investigate the impact of these vocabulary distributions, inherent to the chosen model or to the prompt that is used, and strive to better isolate these effects—potentially by controlling or measuring the LLMs distribution more explicitly or comparing with humans with more or less different vocabularies.

Finally, our metrics for quantifying convergence may overlook nuanced pragmatic or syntactic adaptations. Future studies could expand these methods to incorporate richer dialogue annotation, or longitudinal tracking of individual language changes to provide a more comprehensive view of human-LLM co-adaptation.

Ethical Considerations

The divergence seen in human-LLM pairs raises questions about long-term implications of embedding LLMs in daily life. At a societal level, the homogenization of language and thought is a valid concern, particularly if users unconsciously pick up machine-like expressions or patterns. While when considering human-human alignment, some degree of efficiency can be beneficial, a loss of linguistic diversity may undercut creativity and cultural specificity. This underscores the importance of AI literacy initiatives that educate users about potential shifts in their communicative styles when relying heavily on AI systems.

Resources

The codebase and data for reproducing our experiments can be accessed at: `https://github.com/Finebouche/words_synch_challenge`

A demo is available on the project website: `https://word-sync.games/`

Author Contributions

Conceptualization, implementation, data collection: T.C.; **Experiment design and statistical analysis:** R.J.; **Data analysis and manuscript writing:** T.C. and R.J.; **Supervision and manuscript review:** T.B. and J.D.

All authors have read and approved the final version.

References

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Tanguy Cazalets and Joni Dambre. 2025. Word synchronization challenge: A benchmark for word association responses for llms. *Preprint*, arXiv:2502.08312.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Herbert H. Clark and Susan E. Brennan. 1991. *Grounding in communication.*, pages 127–149. American Psychological Association.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Greta Gandolfi, Martin J Pickering, and Simon Garrod. 2022. Mechanisms of alignment: shared control, social cognition and metacognition. *Philosophical Transactions of the Royal Society B*, 378(1870):20210362.

Simon Garrod and Gwyneth Doherty. 1994. Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3):181–215.

Esam Ghaleb, Marlou Rasenberg, Wim Pouw, Ivan Toni, Judith Holler, Asl Özyürek, and Raquel Fernández. 2024. Analysing cross-speaker convergence in face-to-face dialogue through the lens of automatically detected shared linguistic constructions.

William Hall. 2014. *The Playbook: Improv Games for Performance*.

Iva Ivanova, William S Horton, Benjamin Swets, Daniel Kleinman, and Victor S Ferreira. 2020. Structural alignment in dialogue and monologue (and what attention may have to do with it). *Journal of Memory and Language*, 110:104052.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Andrew Kim, Maxim Ruzmaykin, Aaron Truong, and Adam Summerville. 2019. Cooperation and codenames: Understanding natural language processing via codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 160–166.

Ulman Lindenberger, Shu-Chen Li, Walter Gruber, and Viktor Müller. 2009. Brains swinging in concert: cortical phase synchronization while playing guitar. *BMC neuroscience*, 10:1–12.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2024. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105.

Nora McDonald and Shimei Pan. 2020. Intersectional ai: A study of how information science students think about ethics and their impact. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–19.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02).

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 384–387.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *Preprint*, arXiv:2406.09264.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: Challenges and opportunities with social chatbots.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ana Lucía Valencia and Tom Froese. 2020. What binds us? inter-brain neural synchronization and its implications for theories of human consciousness. *Neuroscience of Consciousness*, 2020(1):niaa010.

Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2024. How knowledge workers think generative ai will (not) transform their industries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI 24, pages 1–26. ACM.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

A Data Collection through a Web app

A.1 Overview

In this study, participants used a custom Web application (developed in JavaScript and Node.js) to play the "Word Synchronization Challenge." (Cazalets and Dambre, 2025). Figure 6 shows a screenshot of the game interface during play.

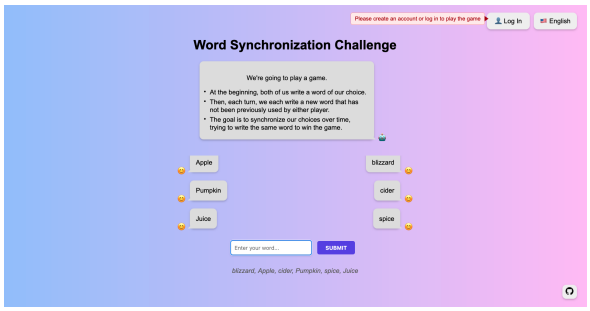

Figure 5: Screenshot of the web app during a game with another human

A.2 Application Architecture

The back end consists of a Node.js/Express server that handles HTTP requests, user sessions, and game-related APIs; AaSocket.io module that enables real-time communication and game state synchronization for human-human games: a SQLite database managed with Sequelize, which stores persistent game and user data.

Our database schema defines two models:

- **Player:** Stores each users `playerId` and prolific Id.
- **Game:** Records game details, including the `playerIds` involved (or a `botId`, when playing with an LLM), language settings, sequence of played words, number of rounds, the winning player, and post-game survey responses.

B LLM Model Prompt Engineering

To standardize interactions with the LLM during the word game, we designed specific prompts for each round. Our approach encourages diversity in the first response and guided, context-aware adaptation in subsequent rounds. Below, we detail the prompts used:

B.1 Initial Round (Round 1)

```
Round 1. New game, please give your first (really random) word and only that word. You can be a bit creative but not too much. Be sure to finish your answer with it.
```

Rationale: This prompt instructs the LLM to provide a single, moderately creative word as its initial response, setting a baseline for the game without strong contextual bias to avoid predictable or semantically anchored openings.

B.2 Subsequent Rounds (Rounds 2 and onward)

```
"${player_word}! We said different words, let's do another round. So far we have used the words: [${past_words_array.join(', ')}], they are now forbidden. Based on previous words, what word would be most likely for next round given that my word was ${player_word} and your word was ${bot_word}? Please give only your word for this round."
```

Rationale: For each subsequent round, the prompt dynamically incorporates both players prior words (explicitly listing them as forbidden to not break the rules) and provides the most recent player and bot words as context. The LLM is explicitly asked to generate the next word based on this shared context, fostering both semantic continuity and adaptation. The instruction to output only your word ensures concise, focused responses.

B.3 Model Settings

We adjusted the model settings based on the round number to ensure varied yet contextually constrained responses. The settings are as follows:

Round	Temperature	Max Tokens
Round 1	1.6	50
Other Rounds	1.1	20

Table 2: Model settings for different rounds.

These settings and the prompt design were critical in ensuring that the models responses were both natural and aligned with the games requirements. Detailed implementation code is available upon request.

C Instruction given to participants

In this study, you will participate in a simple word-guessing cooperative game

At the beginning, each player writes a random word. Then, on each turn, each player writes a new word that has not been previously written. The goal is to produce the same word, in which case the game is won. If the game exceeds 15 turns, the game is lost.

Connection (1 time)

1. When you arrive on the website, please click on **"Log in" "Generate ID"**.
2. Copy this ID and **"Log in"** using it.
3. Fill in the requested information carefully (some are redundant with Prolific).
4. **Important:** Ensure you fill in your Prolific ID.

Play the game (16 times)

1. You will play **8 games with an Artificial Intelligence (AI)** system and **8 games with another human player**, in random order.
2. Press the button **"Play with a human"** or **"Play with an AI"** to start the game.
3. It might take some time before the AI system is ready or before another human player joins. If you have to wait more than 3 minutes, refresh the page and click the **"Play with..."** button again.
4. After each game, you will be asked to fill in a questionnaire (see below).

Fill in the questionnaire (16 times) After each game, youll be asked a few questions about your experience. These questions help us understand how you and your partner strategized during the game. Specifically, youll be asked to share:

- **Your Strategies:** What approach or idea you used while playing.
- **Your Partners Strategies:** What you think your partner was trying to do.
- **Whether you believe your partner understood the strategy you used.**
- **Whether you understood your partners strategy.**
- **Partner Rating:** Rate your partners performance on a scale from 1 (lowest) to 5 (highest).

Your answers will help us learn more about how interacting with AI versus human partners influences communication and decision-making.

Upon completion of all the games and questionnaires, you will be given a link to complete the study in Prolific. Thanks for playing!

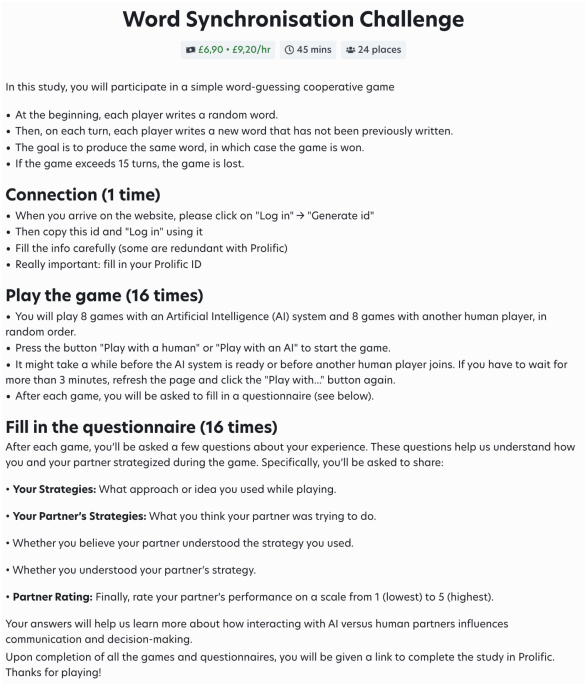

Figure 6: Screenshot of the instruction as seen in prolific

D Responsible use of AI

GitHub Copilot and OpenAI ChatGPT have been used as coding assistants for website implementation and data analysis. All code generated by AI assistants was manually reviewed.

OpenAI ChatGPT was used to correct grammar and typos in writing. All AI-generated text was manually reviewed.

E Licenses

This work includes data from ConceptNet 5, which was compiled by the Commonsense Computing Initiative. ConceptNet 5 is freely available under the Creative Commons Attribution-ShareAlike license (CC BY SA 4.0) from https://conceptnet.io. The included data was created by contributors to

Commonsense Computing projects, contributors to Wikimedia projects, Games with a Purpose, Princeton University's WordNet, DBPedia, OpenCyc, and Umbel.

The Word Synchronization Challenge framework is licensed under the MIT license, a permissive open-source license that allows for unrestricted reuse, modification, and distribution, including for commercial purposes, provided that the original copyright notice and license are included.