HUMAN ALIGNMENT: HOW MUCH WE ADAPT TO LLMS?

Anonymous authors
Paper under double-blind review

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 & 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 & 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 & Garrod, 2013). These adaptations reduce cognitive load and help establish common ground (Clark & 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 & 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 & 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 & 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.

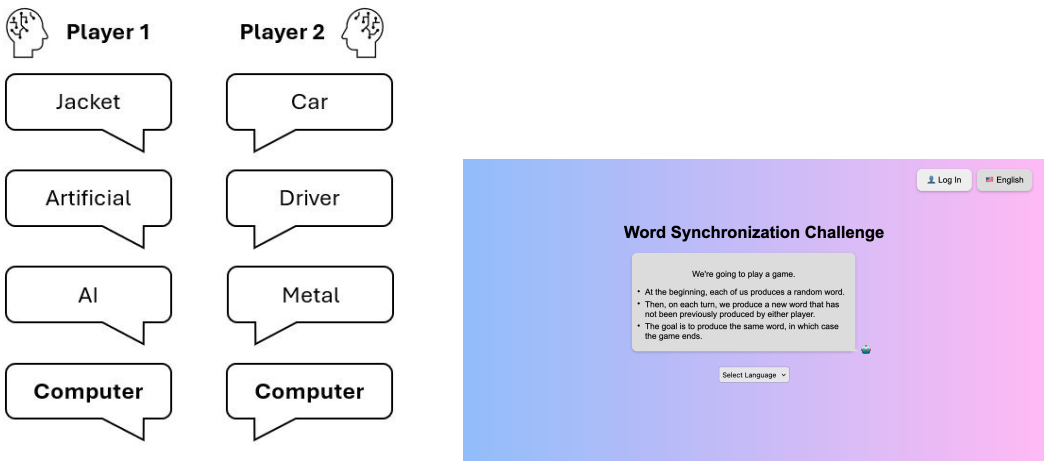


Figure 1: (a) Example of the Word Synchronization Challenge, where participants converge on the same word by the fourth turn. (b) Screenshot of the web app used to run our experiment.

Methodologically, capturing and quantifying mutual adaptation in verbal interaction poses unique 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 & 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). Are humans able to 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 & 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).

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 LLMs. The study used a within-subjects 2x2 factorial design, where we manipulated two aspects: whether the participant played against a human or an LLM, and whether the opponent 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. If the players did not converge after 16 turns, the games were automatically stopped.

2.2 LLM IMPLEMENTATION

We used OpenAI's GPT-4o model to generate the AI partner's responses. The prompt was designed to ensure that the LLM's 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.

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."
```

Subsequent Rounds:

```
"${player_word}!  We said different words, let's
do another round.  So far we have used the words:
[${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."
```

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
Subsequent Rounds	1.1	20

Table 1: Model settings for different rounds.

These settings and the prompt design were critical in ensuring that the model's responses were both natural and aligned with the game's requirements.

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 GBP6.90 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. Prolific IDs were collected to ensure data integrity.

2.4 PROCEDURE

Participants accessed the study's web application, logged in with their Prolific ID, and selected either "Play with a Human" or "Play with an AI." They were informed that they would complete 16 randomized games (see Appendix B for full instructions). In each game, both players initially entered a random word. In subsequent rounds, they submitted a new, unused word simultaneously, with the game concluding immediately 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 partner's 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 2 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 2: 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 2 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 Dynamics evaluations

Visualizing word embedding trajectories reveals how partners adjust their language over time to synchronize their word choices. By reducing the embeddings to three dimensions using PCA and displaying them in a 3D scatter plot, we can track the evolution and convergence of word choices during the game. When the final words match, they are highlighted with a special diamond.

For example, Figure 2 shows an unsuccessful case where the models' trajectories split into three distinct manifolds. The first utterance jumps between the first two manifolds, suggesting that the models attempted to get closer by choosing semantically related words, although they ultimately failed to converge. This visualization not only captures the technical dynamics of alignment but also serves as a proxy for the social intelligence needed in natural human–computer interactions. It underscores the importance of dynamic adaptation in achieving communicative success and hints at potential challenges when the cognitive models underlying language production diverge.

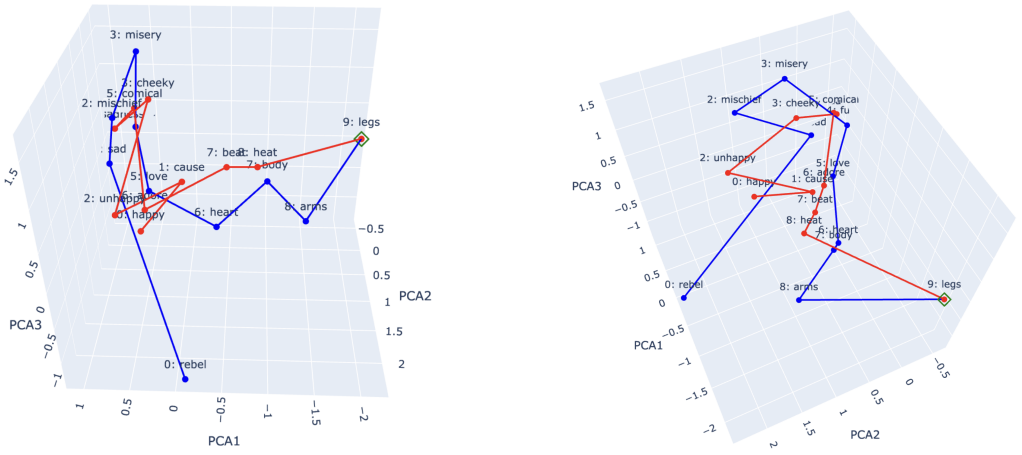


Figure 2: Two different views of the projection of the embedding of one game between humans

Furthermore, Table 3 details the actual sequence of words exchanged between Player 1 and Player 2 during a game. The words are categorized by sentiment and color-coded accordingly. This table provides a complementary, discrete perspective on the dynamics observed in the embedding trajectories, allowing us to correlate the semantic evolution with the underlying visual patterns.

Table 3: Sequence of words exchanged between Player 1 and Player 2 during a game, categorized by sentiment and color-coded accordingly

Player	1	2	3	4	5	6	7	8	9	10
Player 1	rebel	sad	mischief	misery	funny	love	heart	body	arms	legs
Player 2	happy	cause	unhappy	cheeky	sadness	comical	adore	beat	heat	legs

Beyond the technical dynamics, these patterns provide a glimpse into the social aspects of interaction. Observing how words group into "manifolds" or clusters highlights the strategic adjustments players make. For example, switching between thematically related words (e.g., "happy" to "sad" or "mischief" to "funny") shows players testing different semantic fields as they try to find common ground. Successful alignment (or lack thereof) reflects not just linguistic skill but also the ability to predict and adapt to a partner's behavior—an important aspect of both human social intelligence

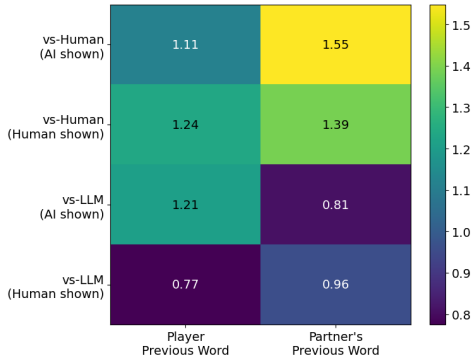


Figure 3: Average CL scores. Each cell represents the mean score for "Player Score" or "Partner Score" within a given game configuration.

and effective AI design. This insight points to broader implications for human-computer interaction, particularly in designing systems that foster smoother, more intuitive communication.

3.4 Strategy Analysis

We used two approaches to analyse the convergence strategies used by the participants: a linguistic analysis of the relationships between the words used, and a subjective assessment by the participants themselves.

3.4.1 Conceptual Linking Score

To further probe the nature of adaptation between participants, 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 player's current word is to either their own previous word or to their partner's previous word. ConceptNet returns a set of edges (associations) along with corresponding weights that reflect the strength of each association. The CL score correspond to the maximum weight from the returned edges (0 if no association is found). Thus, a higher CL score indicates a stronger thematic or conceptual continuity between successive word choices.

For each round of each game we computed the average CL scores to both the player's and the opponent's word from the previous round. Those averages were then averaged across all games within each configuration, and presented in Figure 3.

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 opponent's previous word ($p < 0.001$). No significant differences were found between whether the opponent was portrayed as AI or human.

3.4.2 User-perceived partner strategy

Following the framework of Cazalets & Dambre (2025), in post-game questionnaires participants were asked about their perception of their partner's strategies, asking them to choose one of the three strategies :

- **Mirroring:** Other player has chosen a word that was really close to my previous word.
- **Staying Close:** Other player has chosen a word that was really close to its own previous word.
- **Averaging (Balancing):** Other player has chosen a word in between the two previous words.

Figure 4 shows the average aggregated results of the strategies as reported by player about their partner.

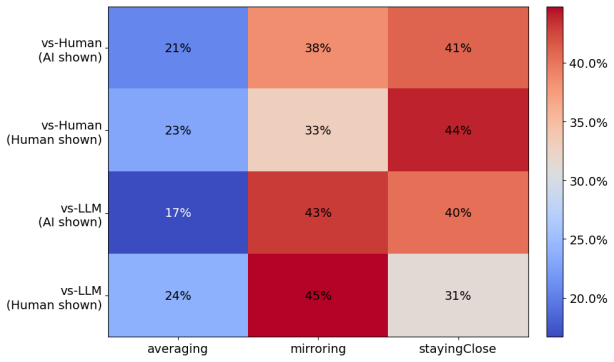


Figure 4: Average reported strategy measures by game configuration. Each cell shows the percentage of time a given strategy was attributed to other player for each game configuration.

4 Discussion and Conclusion

This study investigated whether humans adapt their behavior and word choice differently when interacting with an LLM than with another human, using the WSC to simulate a simple verbal interaction requiring player to align their word choice.

While there was no significant difference in winning rate between when playing against a human or LLM partner, players did converge in significantly fewer rounds when playing with a human than with an LLM.

Analysis of the word choice revealed that humans did change their behavior depending on their partner: when playing against an LLM they chose words that were significantly less similar to their partner's previous word than when playing against a human. This seems to be a reaction to their perception of their partner's strategy, with players indicating that human partners seemed more likely to stay close to their own words than LLMs.

4.1 Implications

The divergence seen in human–LLM pairs raises questions long-term implications of embedding LLMs in daily life. LLMs become integrated into everyday communication—be it through chat apps, educational tools, or healthcare advisers— and understanding existing adaptation gaps is crucial. From a design perspective, developers could explore feedback mechanisms that encourage LLMs to connect semantically not only with user inputs but also with broader contextual cues, potentially promoting richer, more human-like alignment.

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 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.

4.2 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.

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.

4.3 Future Work

Future studies might include larger and more diverse participant pools to validate and extend these observations. Future research should focus on the long-term cognitive 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.

5 Acknowledgments

Removed for anonymisation

References

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

Tanguy Cazalets and Joni Dambre. Word synchronization challenge: A benchmark for word association responses for llms, 2025. URL `https://arxiv.org/abs/2502.08312`.

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

Herbert H. Clark and Susan E. Brennan. *Grounding in communication.*, pp. 127–149. American Psychological Association, 1991. ISBN 1557981213. doi: 10.1037/10096-006.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, November 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09548-1.

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

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

Simon Garrod and Gwyneth Doherty. Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3):181–215, December 1994. ISSN 0010-0277. doi: 10.1016/0010-0277(94)90048-5.

Esam Ghaleb, Marlou Rasenberg, Wim Pouw, Ivan Toni, Judith Holler, Aslı Özyürek, and Raquel Fernández. Analysing cross-speaker convergence in face-to-face dialogue through the lens of automatically detected shared linguistic constructions. May 2024. doi: 10.48550/ARXIV.2405.08546.

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

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

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

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

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 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, pp. 1105, 2024.

Nora McDonald and Shimei Pan. 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, 2020.

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

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. Training language models to follow instructions with human feedback. March 2022. doi: 10.48550/ARXIV.2203.02155.

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

Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02), April 2004. ISSN 1469-1825. doi: 10.1017/s0140525x04000056.

Martin J. Pickering and Simon Garrod. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347, June 2013. ISSN 1469-1825. doi: 10.1017/s0140525x12001495.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 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*, pp. 384–387, 2023.

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. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024. URL https://arxiv.org/abs/2406.09264.

Heung-Yeung Shum, Xiaodong He, and Di Li. From eliza to xiaoice: Challenges and opportunities with social chatbots. January 2018. doi: 10.48550/ARXIV.1801.01957.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

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

Ana Lucía Valencia and Tom Froese. What binds us? inter-brain neural synchronization and its implications for theories of human consciousness. *Neuroscience of Consciousness*, 2020(1):niaa010, 06 2020. ISSN 2057-2107. doi: 10.1093/nc/niaa010. URL https://doi.org/10.1093/nc/niaa010.

Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 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, pp. 1–26. ACM, May 2024. doi: 10.1145/3613904.3642700.

A Data Collection through a Web app

A.1 Overview

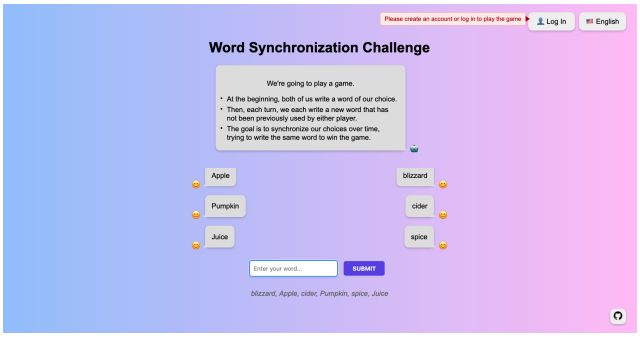


Figure 5: Screenshot of the web app during a game with another human

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

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.

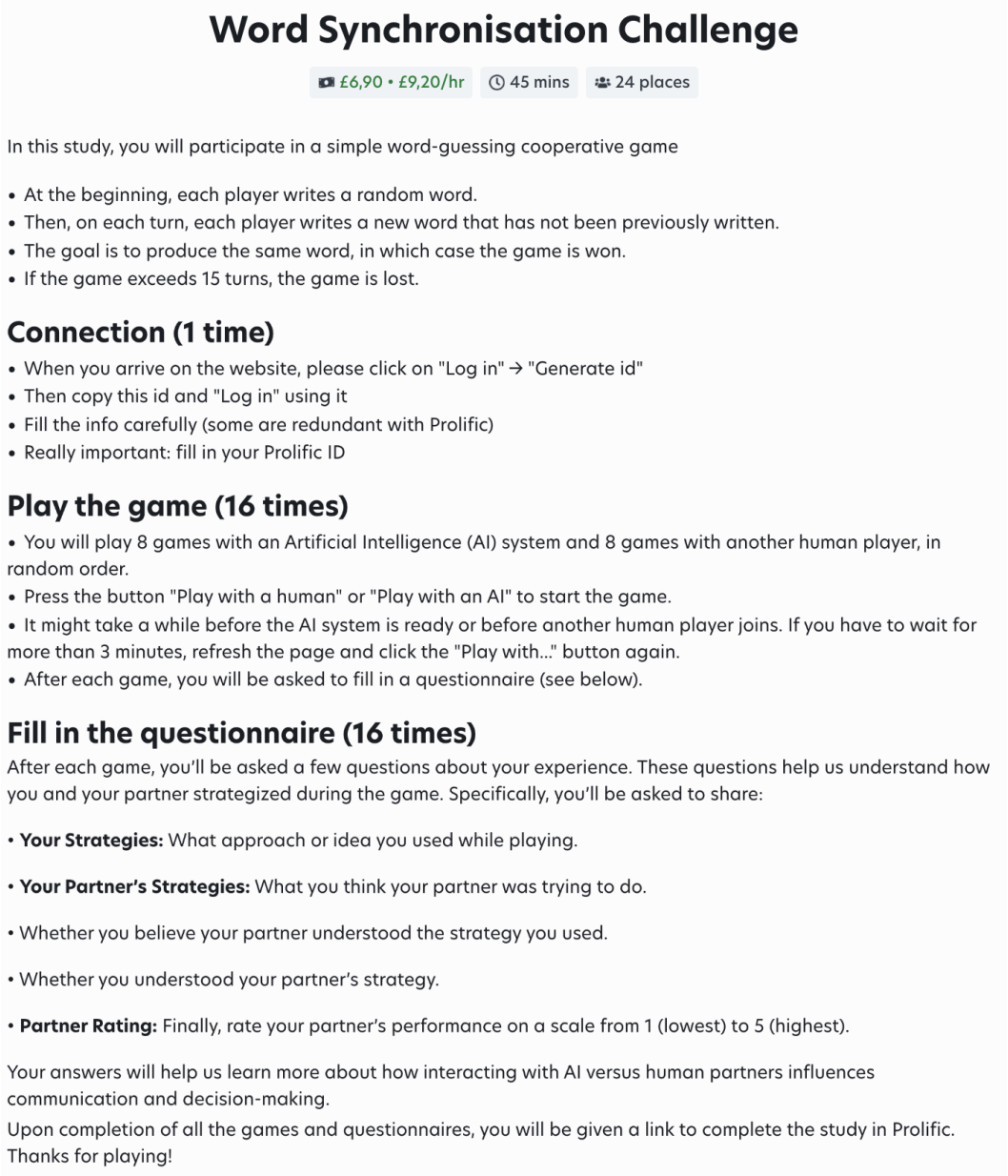
Database Schema. Our schema defines two primary models:

- **Player:** Stores each user's `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.

A.3 Interaction Flow

When users access the web application, they first log in and select a language. They then choose to play either against an LLM (with bots loaded based on language) or another human (matched in real time via Socket.io). During gameplay, participants enter words in each round, which are validated (e.g., checking for duplicates and verifying existence in Wiktionary). The game continues until a winning condition is met or the maximum number of rounds is reached. Figure **??** shows a screenshot of the game interface during play.

B INSTRUCTION GIVEN TO PARTICIPANTS



Word Synchronisation Challenge

£6,90 • £9,20/hr | 45 mins | 24 places

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)

- When you arrive on the website, please click on "Log in" → "Generate id"
- Then copy this id and "Log in" using it
- Fill the info carefully (some are redundant with Prolific)
- Really important: fill in your Prolific ID

Play the game (16 times)

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

Fill in the questionnaire (16 times)

After each game, you'll be asked a few questions about your experience. These questions help us understand how you and your partner strategized during the game. Specifically, you'll be asked to share:

- **Your Strategies:** What approach or idea you used while playing.
- **Your Partner's Strategies:** What you think your partner was trying to do.
- Whether you believe your partner understood the strategy you used.
- Whether you understood your partner's strategy.
- **Partner Rating:** Finally, rate your partner's 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