Human-AI Synergy Supports Collective Creative Search

Chenyi Li[1], Raja Marjieh[2], Haoyu Hu[1], Mark Steyvers[3], Katherine Collins[4,5], Ilia Sucholutsky[*6] & Nori Jacoby[*1]

[1]Cornell University
[2]Princeton University
[3]University of California, Irvine
[4]Massachusetts Institute of Technology
[5]University of Cambridge
[6]New York University

Abstract

The creation of new ideas and content increasingly relies on collectives: not just of multiple humans, but humans and AI agents together. We study collective generation of new ideas using a controlled word-guessing task that balances open-endedness with an objective measure of task performance. Participants attempt to infer a hidden target word, scored based on the semantic similarity of their guesses to the target, while also observing the best guess from previous players. We found that hybrid human–AI groups showed higher performance than both human-only and AI-only groups. Within hybrid groups, both humans and AI agents systematically adjust their strategies relative to single-agent conditions, suggesting higher-order interaction effects, whereby agents adapt to each other's presence. Although some performance benefits can be reproduced through collaboration between heterogeneous AI systems, human–AI collaboration remains superior, underscoring complementary roles in collective creativity. Together, these findings demonstrate the advantages of human–AI synergy in collective intelligence tasks.

Keywords: Collective creativity; Human–AI collaboration; Collective action; Human-AI hybrid society

Introduction

Rapid advances in generative AI are giving rise to a hybrid society in which the production of new ideas and content increasingly integrates human and artificial contributions, across domains ranging from creative writing (Breithaupt et al., 2024; Porter & Machery, 2024) to scientific discovery (Gottweis et al., 2025; Hao et al., 2026; Jumper et al., 2021; Rmus et al., 2025; Schmidgall et al., 2025). As people engaged in the discovery of innovative ideas rely more heavily on AI tools as thought partners (Collins et al., 2024), understanding AI's impact on collective outcomes becomes increasingly important (Sucholutsky et al., 2025). While AI assistance can improve individual task performance, it also risks driving excessive homogenization and reducing collective diversity (Doshi & Hauser, 2024). Studying human–AI interactions at the collective level remains challenging using traditional methods. This is, in part, because many real-world idea discovery proceses are inherently open-ended and hard to evaluate in controlled settings with well-defined objectives. Moreover, both humans and AI systems adapt their behavior in response to the outputs of others, making it difficult to infer emergent collective dynamics from studies focused solely on isolated individual human–AI interactions (Tsvetkova et al., 2024).

Here, we address these challenges by studying a creative discovery task with an objective ground truth. Inspired by the game Semantle (David Turner, 2022; Ueshima et al., 2024), participants infer a hidden target word (Fig. 1A) and receive feedback based on the semantic similarity between their guesses and the target (Fig. 1B). In each round, participants are informed of the best guess and score produced by previous players within the same game (Fig. 1C). Players, either human or Gemini 2.5 AI agents, join groups composed of either humans, AI, or both (Fig. 1D,E). We found that hybrid human-AI groups consistently outperformed homogeneous groups, achieving faster convergence to high-value solutions. Crucially, participation in hybrid groups altered the behaviors of both humans and AI. AI agents exhibited greater lexical diversity and search quality when interacting with humans, while humans achieved slightly higher performance and made significantly more unique guesses in the presence of AI. This mutual adaptation suggests that the benefits of human-AI collaboration emerge not from simple addition but from the dynamic interplay of complementary cognitive strategies: humans explore broadly, preventing premature convergence, while AI exploits efficiently, accelerating progress toward promising regions. To test whether these benefits arise from cognitive heterogeneity rather than agent diversity per se, we analyzed an experiment involving two distinct LLMs, Gemini 2.5 and GPT-5.1. This experiment showed a similar synergistic effect, but overall performance was lower than in human-AI collaboration, indicating that part of the hybrid advantage originates from agent heterogeneity. Finally, we validated in a series of control studies that these effects are robust to different types of AI systems, different forms of communication channels between participants in the game and variations in model decoding temperatures.

Background

The ability of humans to collectively discover new ideas and create novel content is one of the definitive features of human culture (Boyd, Richerson, et al., 1996; Henrich, 2015; Tomasello, 2009). Research on social learning has shown that innovation depends on factors such as network topology, interaction structure, and group size (Brackbill & Centola, 2020; Derex et al., 2019; Mason et al., 2008; Shirado & Christakis,

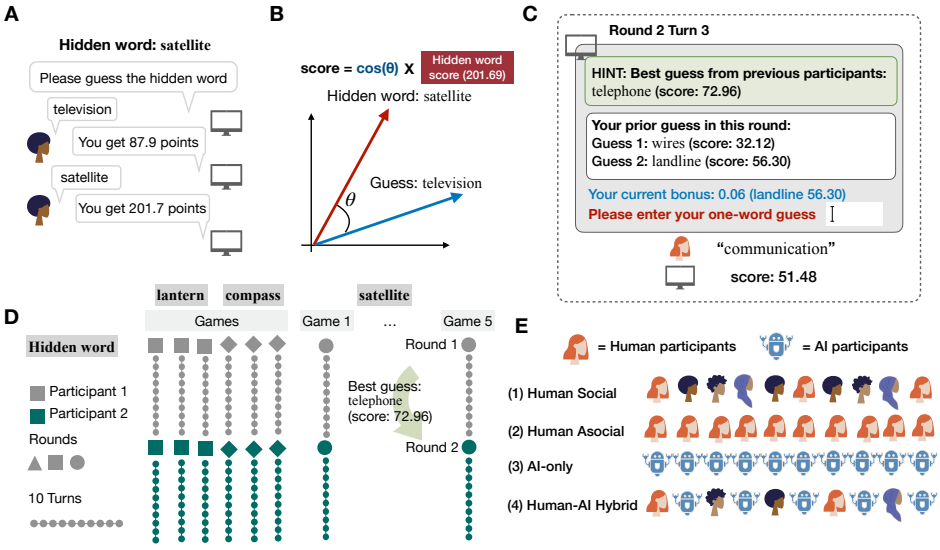

Figure 1: Experiment framework for collective creative search. (A) Participants attempt to infer a hidden target word ("satellite") and receive similarity score feedback over multiple turns. (B) The similarity score of each guessed word is computed by the product of the hidden word score and the cosine similarity between them. (C) In each round, the participants received the best guess from previous rounds as a hint. (D) Participants were embedded in a collective guessing game with a chain-like network, where best-guess information was transmitted within each game (chain). Each game had 10 rounds of 10 turns each, totaling 100 guesses per game. (E) Schematic of the different experimental conditions considered.

2020). However recently AI is increasingly used to assist people in complex cognitive interactions (Collins et al., 2024). What is the impact of AI assistance on idea discovery across groups of people and AI agents? Doshi and Hauser (2024) examined humans assisted by AI agents in a creative writing task and found that AI assistance enhanced individual creative performance but reduced collective diversity, leading to excessive homogenization. However, their study examined human–AI interaction in isolation and did not directly test the emergent collective phenomena that arise from interactions among multiple humans and AI agents (Brinkmann et al., 2023; Tsvetkova et al., 2024).

One promising medium for studying collective behavior comes from the information foraging literature (Garg et al., 2022). The idea is that optimal search in any environment requires balancing exploration of new possibilities with exploitation of known resources (Hills et al., 2015). This exploration-exploitation tradeoff is a fundamental principle observed across domains, from animal foraging to human memory search (Hills et al., 2012). In this literature, it is often noted that individual strategies of exploration and exploitation are similar to the same dynamics at the collective level (Hills et al., 2015).

In a recent paper, Ueshima et al. (2024) used a word guessing game to probe collective creativity in a semantic search setting. They found that simple bots aggregating human guesses had a modest positive effect on group performance for easier words. A similar beneficial effect of simple bots was reported by Shirado and Christakis (2017), who showed that, in a coordination game, introducing bot agents that inject noise can improve collective coordination by preventing groups from becoming trapped in local minima. The paper by Shirado and Christakis (2017) and the semantic search task by Ueshima et al. (2024) provide a valuable framework but do not study realistic AI systems. This limitation is important because Tsvetkova et al. (2024) have shown that human–AI interaction can give rise to substantial second-order effects, whereby humans systematically alter their behavior in response to AI agents. These dynamics imply that studying collective behavior requires experiments involving real humans interacting with real AI systems under shared conditions (Hu et al., 2026; Sucholutsky et al., 2025). Recent technological advances in large-scale online experimentation (Almaatouq et al., 2021; Harrison et al., 2020; Marjieh et al., 2025) now make it possible to embed both humans and real AI agents within large social networks where they interact directly with one another. For example, a recent study embedded humans and AI in hybrid human–AI social networks engaged in a creative writing task (Shiiku et al., 2025). While AI agents initially outperformed humans on the creative task, their output diversity deteriorated over time. In contrast, hybrid human–AI groups began with lower performance but gradually improved, achieving a balance between performance and diversity. Together, these findings suggest that it is now feasible to study human–AI interaction at scale, under controlled conditions, and with realistic contemporary AI systems, while capturing

emergent collective dynamics that cannot be inferred from individual-level studies alone.

Method

We conducted an online experiment on collective creative discovery, wherein participants (either human or AI) played a word-guessing game with one hidden target word per game (Fig. 1A). Ten target words were selected to span a range of frequencies (and difficulty levels) : 'harbor', 'door', 'pencil', 'lantern', 'river', 'compass', 'satellite', 'metamorphosis', 'topography', 'vessel'. Each game consisted of 10 rounds, with 10 guesses per round, yielding 100 guesses per game. Each batch contained 5 games per target word, for 50 games per batch (Fig. 1D); we ran 4 batches, yeilding 200 games in total. All the measures were averaged over 4 batches.

In each round, participants received the best guess from the previous round in the game (Fig. 1C). Participants took ten turns of guesses and received a similarity score after each guess. The guess with the highest score was recorded as a bonus and passed to the next round. The similarity score was computed as the product of the cosine similarity between Word2Vec (Mikolov et al., 2013) embeddings of the guessed and hidden words, scaled by an arbitrary constant (201.69; Fig. 1B). We ensured that the maximum possible score was not revealed to participants, so that they would have an incentive to continue exploring even after identifying the target word. From the original vocabulary of approximately three million tokens, we retained 663,273 common English words after lowercasing and excluding acronyms and brand names that lacked lowercase representations in the model. Guesses that did not appear in this vocabulary were assigned a score of zero.

Experimental conditions

We compared four main experimental conditions (Fig. 1E): (1) **Human Social** (N = 210): all rounds were occupied by human participants, with a different participant playing each round; participants could take part in up to 10 games (one per target word), playing only one round per game. (2) **Human Asocial** (N = 179): each participant was assigned to a single game and completed all 10 rounds, totaling 100 guesses. (3) **AI-only** (20,000 calls to the Google API): all 2,000 rounds were simulated using Gemini 2.5 Flash. AI agents independently received the same prompt and feedback as humans participants, including the hint, within-round guessing history, and prompt "Please enter your one-word guess". (4) **Human-AI Hybrid** (N = 114, 10,160 Google API calls): within each game, each round was randomly assigned to either a human participant or a Gemini 2.5 Flash agent, resulting in an average of 5.1 human and 4.9 AI rounds per game (SD = 1.54 and 1.55, respectively; 1058 human and 1018 AI rounds in total). As in the Human Social condition, human participants completed only one round within any given game and participated in 10 games. To minimize bias, participants were not informed that any guesses may have been generated by AI agents.

Participants and AI queries

Human participants. 503 participants were recruited successfully from Prolific (237, 242, and 24 self-reported as female, male, and other, respectively; mean age = 38.13, SD = 12.30). Participants were based in the US and identified English as their native language. Participants provided consent under a Cornell University approved protocol (IRB0148995) and were compensated at a rate of $9 per hour. We conducted a subsampling-based power analysis using the performance difference between the Social and Asocial conditions, a metric independent of the AI and Hybrid conditions, and found that 60 participants per group were required for this effect to reach significance at the $p < .05$ threshold for over 80% among the 1000 iterations. Our actual sample sizes (N_{social} = 210, $N_{asocial}$ = 179) comfortably exceed this requirement.

AI queries. We conducted 165,140 API queries to Gemini 2.5 Flash (version released June 17, 2025) with temperature 0.7. We also used 10,020 API queries to GPT 5.1 for control experiments (version released November 13, 2025) with temperature 0.7, medium reasoning effort. For strategy annotation, we used 8,461 API queries to Claude Sonnet 4 (version released May 22, 2025) with temperature 0.1. All experiments were conducted using PsyNet (Harrison et al., 2020), a Python-based framework for large-scale online psychological experiments and human-AI interactions.

Results

Exploration trajectories in semantic space

We begin by examining exploration trajectories across conditions. We constructed a semantic space using the same Word2Vec embeddings that were used for scoring (see Methods). We then applied UMAP (McInnes et al., 2018) to reduce these embeddings to two dimensions and plotted them as gray background points. Fig. 2 illustrates exploration trajectories for the Human Social, Human Asocial, AI-only, and Human-AI Hybrid conditions in a representative game with the hidden word "compass". Fig. 2 highlights three key insights. First, trajectories in the AI-only condition form compact clusters in semantic space; for "compass", these clusters center around word groups such as music instrument. Second, the Human Social and Human Asocial conditions show more widely spread trajectories, with higher semantic diversity than the AI condition. Third, the Human-AI Hybrid condition exhibits similarly expansive trajectories and is generally closer to the hidden word "compass" than the other conditions.

Performance analysis

To measure **performance** in creative search, we took the maximum similarity score reached by each round in the game and averaged these maxima across all rounds in all games in each condition. As shown in Fig. 3A, the Human-AI Hybrid condition produced the highest performance while the AI-only condition produced the lowest. The Hybrid condition

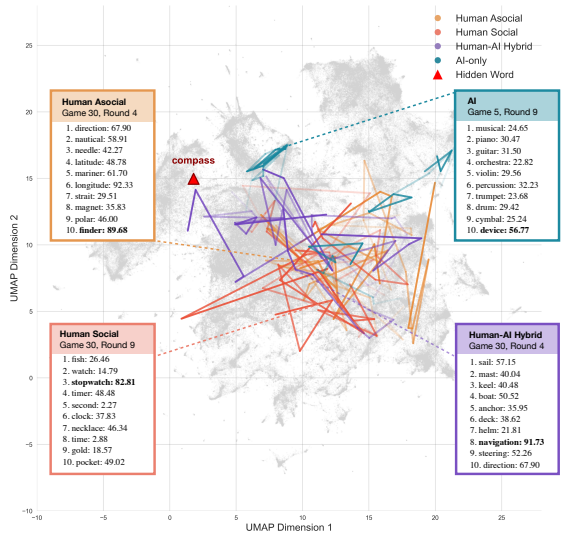

Figure 2: Semantic exploration trajectories. All words were embedded with a Word2Vec model and projected to a two-dimensional space using UMAP. We present one target word ("compass"). Colored trajectories show five games for each of the four conditions. For each game and each round, we computed the average coordinates of the 10 guesses and connected these round centroids from rounds 1 to 10; line opacity increases with round index. Inset boxes display example rounds showing all 10 guesses with their similarity scores; best guesses of the rounds are highlighted.

yielded significantly higher similarity scores compared with the other conditions ($p < .001$). Human Social surpassed Human Asocial conditions significantly ($p < .001$), showing the advantage of collective behaviors which is our main focus.

We next analyzed performance as a function of round within the game (Fig. 3B). Across all conditions, scores increased significantly over time ($p < .001$); however, the magnitude of improvement differed by condition, with AI showing smaller gains ($\Delta = 22.10$) than the other groups ($\Delta = 40.93 - 53.21$). The dynamics also revealed that different conditions reached peak performance at different stages. The Human-AI Hybrid condition exhibited a consistently increasing trajectory throughout the game, maintaining the highest scores in the early and middle rounds. Notably, it surged rapidly in the early stage: by round 3, it already showed significantly higher maximum similarity than both the Human Asocial and AI-only conditions ($p < .005$). The AI-only condition showed modest early gains but plateaued after round 5, ultimately achieving the lowest final scores. In contrast, the Human Social condition showed a later surge, with performance accelerating sharply around rounds 7–8 and converging with the Hybrid condition by rounds 9–10.

Diversity analysis

To measure **diversity**, we used **lexical diversity**, defined as the average proportion of unique guessed words within each game (Fig. 3C). The Human Asocial condition exhibited the highest diversity ($M = 0.86$), likely reflecting higher exploration rates or greater noise. In contrast, the AI-only condition showed markedly lower diversity ($M = 0.37$), indicating a tendency to remain confined to a narrow region of semantic space despite consistently receiving relatively low scores. The Human-AI Hybrid condition showed reduced lexical diversity ($M = 0.58$) compared with the human conditions, but significantly higher diversity than the AI-only condition ($p < .001$). Interestingly, Fig. 3D shows that both the Human Social and Human-AI Hybrid conditions decreased in diversity across iterations, likely because guesses became closer to the target over time. These results suggest that Hybrid Human-AI groups occupy an intermediate position in terms of diversity, preserving more exploratory variation than AI-only groups while reducing the high variability observed in human-only settings.

Guessing strategy analysis

To characterize guessing strategies beyond vocabulary diversity, we implemented an LLM-as-a-judge annotation pipeline using Claude Sonnet 4. In a preliminary step, an independent instance of the model reviewed the full set of semantic guess trajectories across the main conditions and inductively derived seven strategy types defined relative to the hint. The annotation model then classified each round of 10 guesses into a soft probability distribution over the seven types. Each round was then assigned a discrete label by taking the strategy with the highest probability, and the seven labels were further collapsed into four broader categories: Exploit (same-category exploit or subcategory shift), Directed Explore (associative or inferential category break), Undirected Explore (random category break or hint-independent), and Mixed (mixed strategy).

Fig. 3D shows how the strategy distribution varied over five equal-frequency hint score quantiles across conditions. A red dashed ordinary least squares (OLS) trend line for P(Exploit) is overlaid, with the bootstrapped slope and 95% CI annotated. Both human conditions demonstrated clear adaptive behaviors (Human Social: $M = 0.0006$, $\text{CI}_{95\%} = [0.0003, 0.0010]$; Human Asocial: $M = 0.0011$, $\text{CI}_{95\%} = [0.0005, 0.0014]$): when hint scores were low, they relied more on Explore or Mixed strategies to escape unpromising regions of semantic space; as hint quality increased, they progressively shifted towards Exploit strategies. In contrast, the AI-only condition showed virtually no adaptation to hint quality ($M = 0.0001$, $\text{CI}_{95\%} = [0.0000, 0.0002]$). Exploit was the dominant strategies (99%) in AI-only group, higher than that of other conditions ($p < .001$). This indicates that AI tends to focus on nearby semantic region of the hint and lacks the adaptivity to escape local maxima. The Human-AI Hybrid condition displayed an intermediate pattern. Its strategy distribution was more adaptive than the AI-only condition ($p = .0036$) and achieved compatative adaptivity to the hint score, compared to Human Social ($p = 1.0$) and Human Asocial ($p = 1.0$).

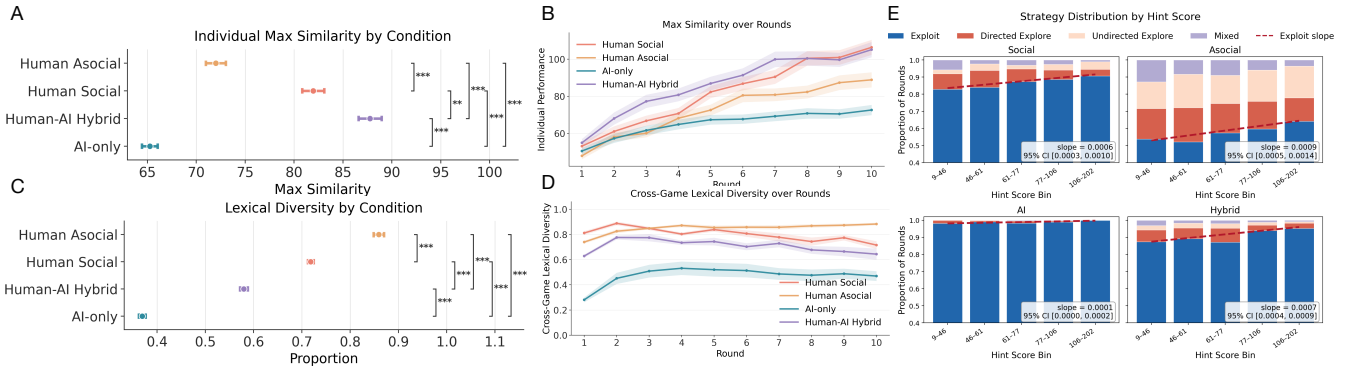

Figure 3: Performance, diversity and strategy. **A.** Performance, computed as the average of the maximal score across rounds. Error bars represent one standard error across participants. Asterisks *, **, and *** denote significance levels of 0.05, 0.01, and 0.001, respectively. To account for multiple comparisons (here and in the rest of the paper), only results that remained significant after Bonferroni correction are shown. **B.** Performance across rounds. Error bars represent standard error across participants. **C.** Lexical diversity, computed as the average proportion of unique guessed words of each game. **D.** Cross-game lexical diversity acorss rounds. **E.** Strategy distribution by hint score quantile across conditions. The dashed red line shows the OLS trend for P(Exploit) across hint score bins; slope and bootstrapped 95% CI are annotated within each panel.

Interaction with people changes AI behavior

Previous theoretical work (Tsvetkova et al., 2024) suggests that humans and AI can influence each other indirectly by observing and responding to each other's behavior (second-order effects). To quantify the influence of Hybrid Human-AI condition on both human and AI agents, we compared their performance and diversity. Fig. 4A shows that human participants in the Hybrid condition did not differ significantly in individual performance from the Human Social group ($p = 0.392, d = 0.032$). In contrast, AI agents in the Hybrid condition performed significantly better than in the AI-only condition ($p < .001, d = 0.618$). These results suggest that AI agents, in particular, modified their behavior when exposed to input from the other type of agent.

As shown in Fig. 4B, humans in the Hybrid condition contributed a significantly higher proportion of unique words (**lexical diversity**) compared to the Human Social condition ($p < .001$). Similarly, AI agents in the Hybrid condition also showed a significantly higher lexical diversity compared to the AI-only condition ($p < .001$).

Taken together, these results reveal an asymmetric complementarity between human and AI agents. AI agents benefited most from hybrid condition: human input helped them shift toward more varied and productive regions of semantic space. In contrast, human showed little change in performance or diversity, suggesting they served primarily as a source of exploratory signal that guided AI toward more promising regions of semantic space.

Controlling for effect of model type and synergy

To understand the effect of different agents, we implemented a Hybrid AI condition consisting of Gemini 2.5 Flash (50%) and GPT 5.1 (50%) agents, and compared it to our original experiment with Gemini 2.5 Flash and new experiment with GPT 5.1 alone. This Hybrid AI condition ($M = 72.83$, $CI_{95\%} = [70.87, 74.80]$) showed significantly higher maximal similarity score than Gemini 2.5 (AI-only) condition ($M = 65.25$, $CI_{95\%} = [63.66, 66.84]$; $t(3998) = 5.88$, $p < .001$, $d = 0.19$) and the GPT 5.1 condition ($M = 64.08$, $CI_{95\%} = [62.59, 65.58]$; $t(3998) = 6.94$, $p < .001$, $d = 0.22$) but remained significantly worse than the Human-AI Hybrid condition ($M = 87.76$; $t(4081) = -14.93$, $p < .001$, $d = 0.30$).

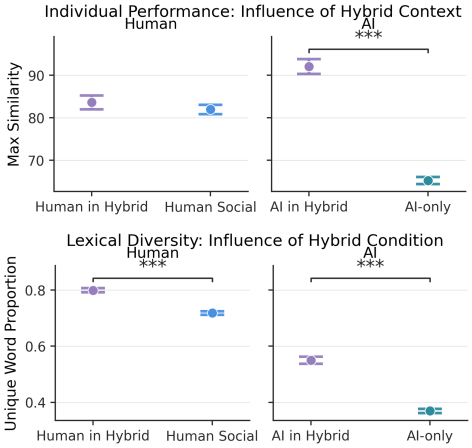

Figure 4: Indirect influence of AI on human behavior and of humans on AI in the Human-AI Hybrid condition. **A.** Performance: comparison of human performance in the purely human and Hybrid (Human–AI) conditions, and AI performance in the purely AI and Hybrid conditions. **B.** Same for lexical diversity.

For diversity, Gemini 2.5 (AI-only condition) displayed similar lexical diversity compared to Hybrid AI ($M = 0.37$, $CI_{95\%} = [0.35, 0.38]$, $p = 1.0$, $d = 0.017$) and higher diversity than GPT 5.1 ($M = 0.29$, $CI_{95\%} = [0.28, 0.31]$, $p < .001$, $d = 0.76$).

Controlling for prompts and social information

We conducted additional AI-only control experiments to assess the robustness of our main findings to variation in prompt design, social information format, and model decoding temperature.

Three alternative prompt design and social information conditions yielded no significant performance differences relative to the main condition (all $p > .1, |d| \leq 0.098$): a Complete History condition, in which agents received the full guessing history of previous participants; a Short Advice condition, in which agents received and passed on a single-word piece of advice; and a Long Advice condition, in which agents exchanged multi-sentence natural-language advice. Together, these results suggest that the best-guess signal used in the main experiment provides a sufficient and efficient mechanism for transmitting progress across rounds.

We further varied the decoding temperature $(0.3, 0.5, 0.9, 1.1)$; none of these variants outperformed the main condition (all $|d| \leq 0.12$), and although higher temperature modestly increased lexical diversity, it did not translate into performance gains. The results suggest that increasing output randomness alone is insufficient to overcome AI's tendency toward local exploitation.

Overall, the control experiments validate that our main results are robust to variation in prompt design, transmitted social information and moel decoding temperature.

Discussion

This study examined how advanced large language models (LLMs) shape collective creative search when embedded within human–AI collaborative groups. We found that hybrid human–AI groups consistently outperformed both human-only and AI-only groups while obtaining intermediate diversity and maintaining an adaptive strategy. In contrast, AI-only groups exhibited lower overall performance, and this disadvantage proved robust across different AI systems, prompting strategies, and forms of social information (e.g., single-word versus sentence-level feedback).

Our results highlight how humans, AI agents, and hybrid collectives employ fundamentally different search strategies. Human groups prioritize exploration and maintain diversity, but face limitations in rapidly exploiting promising semantic regions. By contrast, AI-only groups focused on narrow regions of semantic space and achieved relatively high average similarity scores early on; however, despite this efficiency, they frequently failed to make decisive breakthroughs to the hidden targets and became trapped in local optima, likely due to their reliance on exploit strategies regardless of hint quality. These findings align with prior work on complementary inductive biases in humans and AI systems (Steyvers et al., 2022): humans excel at broad, expansive exploration, whereas AI systems, at present, excel at focused exploitation. Hybrid human–AI group combines these strengths. Human exploration introduces semantically diverse guesses that disrupt AI's tendency toward narrow convergence, preventing premature fixation on suboptimal semantic regions. In turn, the concentrated and high-quality guesses generated by AI sharpen human exploration and accelerate movement toward highly relevant areas of semantic space. Moreover, we found that in hybrid settings, humans exhibited increased exploratory diversity, while AI agents displayed greater performance and diversity than in AI-only group. This mutual adaptation suggests that creative performance in hybrid systems emerges through co-adaptation between complementary cognitive strengths. Aligned with this, some—but not all—of these benefits can be partially replicated by combining heterogeneous AI systems from different providers (e.g., Google and OpenAI).

However, there are several limitations to our current work. First, our task operationalizes the complex process of creating new ideas as semantic search using single-word guesses with computational similarity feedback. While this design enables experimental control, it does not capture the full richness of real-world creative processes such as writing, design, or scientific discovery, which unfold in high-dimensional spaces with open-ended and often multiple objectives. Future work should aim to develop controlled yet richer task environments as testbeds for creativity (Shiiku et al., 2025). Second, we focused on a linear chain structure and a restricted communication channel that transmitted only best-guess information. These design choices likely shape exploration–exploitation dynamics; future studies should examine richer network topologies (e.g., small-world or fully connected networks; Marjieh et al., 2025) and alternative information-sharing schemes, including synchronous collaboration. Third, our conclusions regarding AI behavior are conditioned on the specific LLMs and prompting strategies used here. We have only begun to explore different models and training regimes that may yield different convergence and diversity dynamics. Finally, our diversity measure relies on a simple count of word uniqueness that may not fully capture novelty or usefulness; combining these metrics with human evaluations or downstream task performance would strengthen external validity.

In conclusion, this work represents a first step towards an experimental understanding of collective human–AI collaboration. Our findings demonstrate that human and AI benefit from each other, but also highlight the importance of understanding human contributions in terms of emergent collective behavior rather than isolated individual performance. Studying creativity at the collective level is therefore essential for designing and understanding future human and AI hybrid societies (Brinkmann et al., 2023; Collins et al., 2025).

Acknowledgments

This work was supported by the NSF grant "Collaborative Research: Research Infrastructure: HNDS-I: Building Infrastructure to Study Human-AI Hybrid Societies in Experimental Social Networks" (Award BCS-2523500) and partially supported by the NSF grant "Collaborative Research: Designing smart environments to augment collective learning & creativity" (BCS-2421386). KMC acknowledges support from the NSF SBE SPRF.

ChatGPT version 5.2 (OpenAI) was used to assist manuscript editing and proofreading. Authors reviewed each of the edit suggestions, and approved the final version.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158–2171.

Boyd, R., Richerson, P. J., et al. (1996). Why culture is common, but cultural evolution is rare. *Proceedings-british academy*, *88*, 77–94.

Brackbill, D., & Centola, D. (2020). Impact of network structure on collective learning: An experimental study in a data science competition. *PLoS One*, *15*(9), e0237978.

Breithaupt, F., Otenen, E., Wright, D. R., Kruschke, J. K., Li, Y., & Tan, Y. (2024). Humans create more novelty than chatgpt when asked to retell a story. *Scientific Reports*, *14*(1), 875.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., et al. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855–1868.

Collins, K. M., Bhatt, U., & Sucholutsky, I. (2025). Revisiting rogers' paradox in the context of human-ai interaction. *arXiv preprint arXiv:2501.10476*.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

David Turner. (2022). Semantle: Daily word guessing game [Web-based semantic word game; players guess a hidden word based on semantic similarity feedback]. *Semantle.com*.

Derex, M., Bonnefon, J.-F., Boyd, R., & Mesoudi, A. (2019). Causal understanding is not necessary for the improvement of culturally evolving technology. *Nature human behaviour*, *3*(5), 446–452.

Doshi, A. R., & Hauser, O. P. (2024). Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, *10*(28), eadn5290.

Garg, K., Kello, C. T., & Smaldino, P. E. (2022). Individual exploration and selective social learning: Balancing exploration–exploitation trade-offs in collective foraging. *Journal of the Royal Society Interface*, *19*(189), 20210915.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., et al. (2025). Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Hao, Q., Xu, F., Li, Y., & Evans, J. (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 1–7.

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 10659–10671, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/7880d7226e872b776d8b9f23975e2a3d-Paper.pdf

Henrich, J. (2015). The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter. In *The secret of our success*. princeton University press.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46–54.

Hu, H., Marjieh, R., Collins, K. M., Li, C., Griffiths, T. L., Sucholutsky, I., & Jacoby, N. (2026). Why human guidance matters in collaborative vibe coding. *arXiv preprint arXiv:2602.10473*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, *596*(7873), 583–589.

Marjieh, R., Anglada-Tort, M., Griffiths, T. L., & Jacoby, N. (2025). Characterizing the interaction of cultural evolution mechanisms in experimental social networks. *arXiv preprint arXiv:2502.12847*.

Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, *137*(3), 422.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Porter, B., & Machery, E. (2024). Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, *14*(1), 26133.

Rmus, M., Jagadish, A. K., Mathony, M., Ludwig, T., & Schulz, E. (2025). Generating computational cognitive models using large language models. *arXiv preprint arXiv:2502.00879*.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., & Barsoum, E. (2025). Agent

laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 5977–6043.

Shiiku, S., Marjieh, R., Anglada-Tort, M., & Jacoby, N. (2025). The dynamics of collective creativity in human-ai hybrid societies. *arXiv preprint arXiv:2502.17962*.

Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, *545*(7654), 370–374.

Shirado, H., & Christakis, N. A. (2020). Network engineering using autonomous agents increases cooperation in human groups. *Iscience*, *23*(9).

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), e2111547119.

Sucholutsky, I., Collins, K. M., Jacoby, N., Thompson, B. D., & Hawkins, R. D. (2025). Using llms to advance the cognitive science of collectives. *Nature Computational Science*, *5*(9), 704–707.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876.

Ueshima, A., Jones, M. I., & Christakis, N. A. (2024). Simple autonomous agents can enhance creative semantic discovery by human groups. *Nature communications*, *15*(1), 5212.