Hybrid Human–AI Societies Balance Creativity and Diversity

Shota Shiiku[a,b,g,1], Raja Marjieh[c,1], Thomas L. Griffiths[c,d], Manuel Anglada-Tort[e,g,2], and Nori Jacoby[f,g,2,3]

This manuscript was compiled on September 27, 2026

Rapid advances in generative AI are giving rise to hybrid societies in which humans and machines are deeply interconnected. However, understanding how human–AI interaction shapes collective outcomes such as intelligence and creativity remains a challenge. We conducted large-scale social network experiments with 1,473 human participants and 8,590 AI agents who selected, modified, and transmitted stories from their neighbours over many iterations. We compared networks composed of humans, AI agents, and hybrid human–AI populations. Our results reveal an important trade-off: AI-only populations produced the most creative stories but rapidly departed from the original narrative and failed to sustain collective diversity. Human-only populations showed the opposite pattern, preserving narrative continuity and increasing diversity, but producing less creative outputs. Hybrid human–AI populations combined these strengths, achieving high creativity while preserving diversity and continuity over time. Computational semantic analyses revealed that this advantage emerged from complementary generative strategies, balancing humans' tendency to preserve semantic structure with novel ideas introduced by AI agents. Exposure to AI-generated content altered human behavior, showing that indirect human–AI interactions can shape collective diversity. Finally, combining AI models (GPT and Claude) within the same networks reproduced some diversity benefits of human–AI populations, despite each model exhibiting low diversity in isolation. Together, these findings show that the collective performance of human–AI societies depends not simply on individual capabilities, but on how distinct generative strategies interact within populations. Diversity in generative strategies may therefore be key to balancing innovation and cultural diversity in human–AI systems.

Human–AI Interaction | Collective Intelligence | Collective Creativity | Social Networks | Large Language Models

Significance

As generative AI becomes increasingly embedded in human societies, understanding its impact requires looking beyond individual humans and machines to the collective systems they create. We embedded human and AI participants within social networks, where they interacted to create stories over time. AI-only populations produced highly creative stories but low collective diversity, whereas human-only populations became increasingly diverse but produced less creative stories. Hybrid human–AI populations combined these strengths, achieving high creativity and diversity over time. Mixing different AI models within social networks partially reproduced this benefit. These findings show that collective creativity depends not simply on individual capabilities, but on the diversity of generative strategies within a population, providing a design principle for integrating AI into collectives.

Generative AI has become deeply embedded in modern society. Conversational agents increasingly mediate communication and work, generative models produce high-quality images and music on demand, and algorithmic systems shape our exposure to news and culture (1). As a result, society is evolving into a hybrid social system in which beliefs, ideas, and cultural artifacts emerge through continuous interaction between humans and intelligent machines (2, 3). These interactions produce emergent collective behaviors that cannot be inferred from the behavior of individual humans or AI models alone (4). Understanding how human-AI interaction shapes collective outcomes, such as intelligence and creativity, is therefore essential for explaining the impact of artificial intelligence on human culture and society.

Recent advances in computational social science have transformed our understanding of social networks and human collective intelligence (5). This research demonstrates that collective outcomes depend not only on the abilities of individual participants, but also on how people are connected, how information flows through networks, and how they learn from one another (6–8). In many contexts, diversity among group members (e.g., differences in problem-solving and cognitive strategies) can promote collective intelligence (9–11).

Author affiliations: [a]Graduate School of Science and Technology, Shizuoka University, Hamamatsu 432-8011, Japan; [b]ExaWizards Inc., Tokyo 108-0073, Japan; [c]Department of Psychology, Princeton University, Princeton, NJ 08544, USA; [d]Department of Computer Science, Princeton University, Princeton, NJ 08544, USA; [e]Department of Psychology, Goldsmiths, University of London, London SE14 6NW, UK; [f]Department of Psychology, Cornell University, Ithaca, NY 14853, USA; [g]Max Planck Institute for Empirical Aesthetics, 60322 Frankfurt am Main, Germany

S.S., R.M., M.A.-T., and N.J. designed research; S.S. and R.M. implemented the experiments; S.S. run the experiments; S.S., R.M., M.A.-T., and N.J. analyzed data; S.S., R.M., T.L.G., M.A.-T., and N.J. wrote and revised the paper. M.A.-T. and N.J. supervised the research.

The authors declare no competing interest.

[1]S.S. and R.M. contributed equally to this work. [2]M.A.-T. and N.J. jointly supervised this work.

[3]To whom correspondence should be addressed. E-mail: kj338cornell.edu

DRAFT

Key properties of social networks, such as network size and topology, have also been found to shape cumulative cultural evolution by influencing how innovations are generated, selected, and transmitted over time (12, 13).

Here, we extend this experimental approach to investigate collective intelligence in hybrid populations composed of interacting humans and AI agents. We focus on collective creativity, a form of collective intelligence concerned with the joint production of creative ideas and artifacts such as stories, music, and visual art (5, 14**?**). Collective creativity is a fundamental driver of cultural evolution, enabling societies to generate, accumulate, and refine the ideas that drive scientific, technological, and artistic innovation (15**?** –17). As AI systems increasingly participate in creative processes, understanding how hybrid human–AI societies collectively create has become a fundamental scientific challenge (2, 18). However, unlike classical collective decision-making problems with a fixed set of alternatives, creative problems are open-ended and high-dimensional, making creativity a particularly demanding test of collective intelligence that requires both effective problem-solving and exploration of a vast space of possible solutions.

Creativity is commonly defined by two dimensions: novelty and usefulness (19). Novelty captures the extent to which an idea departs from existing conventions or expectations, while usefulness reflects its value or relevance within a given context. Research has traditionally focused on individual creativity, including individual differences in creative ability and the cognitive mechanisms underlying idea generation.(20, 21). At the collective level, however, creativity emerges from interactions among multiple individuals, making it necessary to understand how ideas are generated, shared, sustained, and transformed through social interaction.

Research on human–AI creativity has primarily examined how generative AI affects individual performance. AI assistance can increase productivity and quality in domains such as writing, programming, and design (22). Interaction with AI can also enhance *individual creativity* by increasing the efficiency and novelty of idea generation (23, 24). However, studies suggest that AI assistance may also reduce *global diversity* by narrowing the range of ideas produced across individuals (23, 25). This is particularly important in integrating AI into human societies, where the value of creativity and innovation depends on the sustained diversity of ideas rather than multiple voices repeating similar thoughts (26).

In this paper, we study collective creativity in hybrid human–AI systems through storytelling, an open-ended and iterated cognitive process (27–29) central to many creative problems in the real-world. It uses language, which is currently the primary medium through which humans and modern AI systems collaborate, and provides rich and traceable records of how creative ideas emerge and evolve over successive iterations. Early research treated writing primarily as an individual cognitive process, developing process models that explained how writers generate, organise, and revise their ideas (28). Subsequent work showed that creative writing can also emerge through collaboration, with groups of authors jointly constructing narratives that no single writer would have produced alone (30). More recently, human–AI co-writing systems have introduced language models as creative collaborators that support brainstorming, drafting, revision, and other writing tasks (23, 31–34). Further research showed that storytelling can be understood as a fundamental cognitive process through which people organize, interpret, and communicate experience (29). However, existing work has primarily examined individual writers or small collaborative teams, leaving largely unexplored how creativity emerges in larger hybrid social networks of interacting humans and AI.

Transmission chain experiments provide a powerful paradigm for studying cultural evolution by tracing how cultural artifacts change as they are transmitted across generations of participants over time. This approach can reveal the cognitive and social processes that shape collective outcomes across domains such as language (35) and music (36). This paradigm has recently been extended to large language models as well, showing that AI systems reproduce many of the same transmission biases observed in humans and progressively converge towards increasingly stable narratives across successive generations (37–40). In addition, direct comparisons of human and AI storytelling reveal both shared narrative patterns and systematic differences, with human stories showing greater imaginative variation and more lexical and conceptual novelty during cultural transmission (33, 40). However, these studies have focused on homogeneous populations consisting entirely of humans or AI systems, and on linear sequences of interaction. In contrast, human–AI creativity in the real-world unfolds in heterogeneous populations embedded within complex social networks (4). Analyses of large-scale human–AI creative communities suggest that these interactions give rise to distinct cultural dynamics that cannot be captured by linear transmission (41).

We conducted a large-scale study of social networks involving human participants and AI agents to examine how human–AI interactions shape collective creativity. Building on recent advances that allow large-scale social network experiments with real human participants (13), we recruited large numbers of participants online and organized them into evolving societies facing a common creative problem (see Figure 1). The societies corresponded to 5×5 grid networks with periodic boundary conditions (i.e., a torus; Figure 1A). The network structure remained fixed over time and comprised 25 nodes (or "locations") each with four edges (or "connections") to other neighboring nodes. In each generation (or "iteration") of the experiment, incoming participants (either human or AI) were assigned to random available nodes in the network where they selected and modified short stories created by their network neighbours and shared their new versions with the next generation (Figure 1A). The goal of this asynchronous design was to enable complex social simulations while maintaining experimental scalability (i.e., without the need to synchronize large online cohorts of participants). We repeated the same experiments in networks consisting of only humans, only AI agents, and hybrid human–AI populations (Figure 1B).

Our experimental paradigm is closely related to iterative paradigms from the literature on cultural evolution (36, 42–47). Indeed, in the absence of network edges (i.e., when transmission occurs only between participants within the same node at different consecutive iterations) our paradigm's temporal logic coincides with standard transmission chain experiments (see e.g., (36)). By

DRAFT

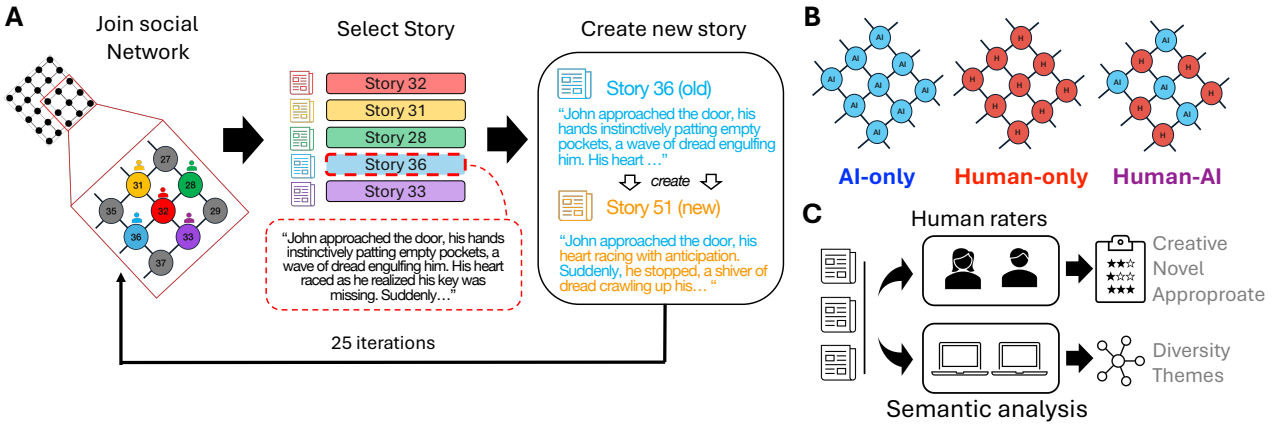


Fig. 1. Experimental framework for studying collective creativity in social networks. **A** Participants join social networks and engage in a creative writing task. Each participant is assigned to a node in the network and sees stories associated with that node and the nodes of their neighbours. They choose their favorite story then modify that story to improve it, with the modified story residing in their node for the next generation (Experiment 1: $N = 879$ human participants). **B** We study three network configurations: human-only , AI-only, and human-AI. **C** The quality of stories is assessed by a separate group of human participants (Experiment 2: $N = 106$ human raters; Experiment 3: $N = 377$) and computational semantic analyses.

introducing edges, following (13) and (48), we thus extend transmission chains into structured populations. Over time, distinctive features of the collective outputs such as appeal, novelty, and diversity emerged and evolved, allowing us to establish causal links between network composition and the emergence of collective creativity. The quality of the creations that evolved were evaluated using a combination of computational measures and large-scale online validation studies with independent human raters (Figure 1C; see Supplementary Figure S12–S15 for example of selected stories in each condition). We used GPT-4o (49) as the AI agent for our main analysis due to its prevalence at the time of testing. However, we also conducted a series of control experiments examining model parameters (e.g., temperature) and model type, including the Claude and GPT models produced by Anthropic* and OpenAI†.

By embedding human participants and AI models within experimental social networks, we directly test the conditions under which human–AI interactions become effective at the collective level and characterize the interaction dynamics through which diversity and creativity emerge and evolve over time. This approach allows us to address fundamental questions about hybrid social systems: When and why do human–AI collectives outperform homogeneous populations of humans or AI agents alone? What mechanisms enable hybrid populations to achieve high creativity while preserving collective diversity? How does the presence of AI alter human behavior? Can some of the benefits of hybrid societies be reproduced by heterogeneous populations of AI systems alone?

Results

Human–AI Populations Balance Creativity and Diversity. We begin by examining the population dynamics of the generated stories across the three experimental conditions (Experiment 1: $N = 879$): AI-only, human-only, and hybrid human-AI populations. In all conditions, the 25 network nodes were initialized with 25 distinct but closely related variations of a single seed story, so that the network began with high but imperfect similarity (see Methods; example seed stories are shown in Supplementary Figure S1). These stories then evolved through repeated iterations of selection, writing, and transmission across the network, as illustrated in Figure 1. Detailed information about all experiments and participants is provided in the Supplementary Materials (see Table S1 and S2; see Methods for full details of the statistical analyses).

Figure 2A shows the mean creativity ratings for the AI-only, Human-only, and Human–AI conditions over iterations (as evaluated by an independent group of human raters, Experiment 2: $N = 106$). Stories in the AI-only condition exhibited the highest creativity ratings from the start and remained highest throughout the experiment (iterations 1-5: $M = 3.30$, $SD = 1.05$, $95\% CI = [3.12, 3.49]$; iterations 21-25: $M = 3.78$, $SD = 0.99$, $95\% CI = [3.61, 3.94]$), significantly increasing over iterations (difference between last five iterations and first iterations: $p < 0.001$, Cohen's $d = 0.46$) and significantly outperforming both conditions in the last five iterations, including Human-only ($p < 0.001$, Cohen's $d = 1.21$) and Human-AI populations ($p < 0.001$, Cohen's $d = 0.52$). In contrast, the Human-only condition showed the lowest creativity ratings initially and remained low by the final iterations (iterations 1-5: $M = 2.55$, $SD = 1.19$, $CI = [2.36, 2.75]$; iterations 21-25: $M = 2.43$, $SD = 1.21$, $CI = [2.23, 2.65]$), with no significant improvement ($p = 0.42$, Cohen's $d = -0.10$). Interestingly, the Human-AI hybrid condition began at a level of creativity similar to Human-only (iterations 1-5: $M = 2.85$, $SD = 1.14$, $CI = [2.65, 3.05]$; iterations 21-25: $M = 3.24$, $SD = 1.09$, $CI = [3.05, 3.43]$), but increased significantly over iterations ($p = 0.008$, Cohen's $d = 0.35$).

To approximate expert annotation (50), we conducted a control experiment showing that ratings from participants self-reporting higher and lower levels of writing expertise were highly consistent, yielding a correlation of $r = .98$ after correction for attenuation (see Supplementary Figure S2)‡.

*https://www.anthropic.com/

†https://openai.com/

‡Although self-reported expertise does not substitute for evaluation by professional writers, these results indicate that aggregate creativity judgments in our paradigm are highly consistent across raters with differing levels of relevant experience and are unlikely to be strongly sensitive to rater expertise. Creative products are often evaluated using the Consensual Assessment Technique,

DRAFT

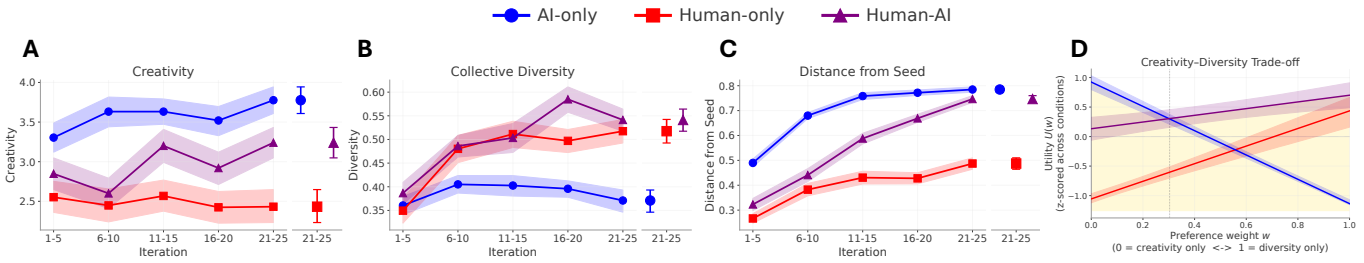


Fig. 2. Human–AI populations achieve a balance between creativity and diversity. **A** Mean creativity ratings of stories over time, as evaluated by human participants. Shaded areas represent 95% bootstrap confidence intervals. The horizontal axis represents the 25 iterations, grouped into five sets of five iterations each. **B** Diversity of stories (inverse similarity; see Methods) over time. **C** Average semantic distance of each story to its originating seed story, over time. **D** Creativity–diversity trade-off: utility $U(w)$ (z-scored across conditions) as a function of the preference weight w, where $w = 0$ corresponds to valuing creativity only and $w = 1$ to valuing diversity only. Error bars and shaded areas indicate 95% bootstrap confidence intervals.

Convergence analyses of the final two five-iteration windows (iterations 16–20 vs. 21–25) revealed different dynamics across conditions. Creativity ratings converged in the Human-only condition ($p = 0.968$, Cohen's $d = 0.01$), but continued to show an increase in both AI-only and Human-AI populations (AI-only: $p = 0.042$, Cohen's $d = 0.25$; Human–AI: $p = 0.027$, Cohen's $d = 0.28$). This suggests that the latter trajectories had not fully plateaued within the 25 iterations studied, and that they are likely to keep increasing for longer runs, thus continuing to outperform the human-only condition.

Subjective ratings of creativity are a useful measure of how individual stories are perceived. However, creativity can also emerge at the collective level, reflecting the diversity and distinctiveness of outputs within a population (23). To capture this, we examined the collective semantic diversity of the generated stories, which indicates how semantically similar stories were within each iteration (operationalized as pairwise cosine similarity between word embeddings computed using all-MiniLM-L6-v2, a standard pre-trained transformer model, (52); see Methods).

Figure 2B shows the evolution of collective diversity, revealing distinct patterns across conditions.. In the AI-only condition, collective diversity remained relatively low and its value in the first and last iterations was similar (iterations 1–5: $M = 0.360$, $95\%CI = [0.341, 0.380]$; iterations 21–25: $M = 0.371$, $95\%CI = [0.346, 0.393]$; change: $+0.010$, $95\%CI = [-0.021, 0.040]$, $p = 0.512$), suggesting that populations of AI systems repeatedly exposed to their own outputs explore a persistently narrow region of semantic space rather than expanding it. In contrast, both the Human-only condition (iterations 1–5: $M = 0.350$, $95\%CI = [0.323, 0.377]$; iterations 21–25: $M = 0.517$, $95\%CI = [0.493, 0.542]$; increase: $+0.168$, $p < 0.001$) and the Human–AI hybrid condition (iterations 1–5: $M = 0.387$, $95\%CI = [0.363, 0.409]$; iterations 21–25: $M = 0.541$, $95\%CI = [0.517, 0.564]$; increase: $+0.155$, $p < 0.001$) showed significant increases in collective diversity across generations. By the final five iterations, both conditions reached comparable levels of semantic diversity (final comparison: difference $= 0.024$, $95\%CI = [-0.011, 0.059]$, $p = 0.176$), suggesting that human participation plays an important role in sustaining semantic exploration and preventing convergence towards homogeneous creations.

Comparing the final iterations (21–25) with the penultimate window (16–20) revealed that semantic diversity showed no significant change over the last part of the experiment in the Human-only population ($p = 0.248$, Cohen's $d = 0.14$). In contrast, both the AI-only ($p = 0.091$, Cohen's $d = -0.20$) and the Human–AI hybrid ($p = 0.018$, Cohen's $d = -0.29$) populations showed declines in semantic diversity over the final generations. These results indicate distinct iterative dynamics across population types within the 25-iteration window, while also suggesting that additional iterations will be needed to characterize their asymptotic behavior. Ungrouped iteration-level results show the same qualitative trends, albeit being naturally noisier (Supplementary Figure S3).

We also examined semantic distance from the original seed story as a measure of transmission fidelity and narrative continuity across generations (Figure 2C; see Methods). This analysis complements measures of collective diversity because populations can increase diversity simply by abandoning information from previous generations. Distance from seed increased significantly across generations in all three conditions (all ps $< .001$), indicating progressive divergence from the initial narrative. However, the magnitude and rate of this divergence differed substantially across populations. AI-only chains exhibited the largest distances from seed throughout the experiment, with semantic distance increasing rapidly during the first generations and reaching high levels by the final iterations (iterations 1–5: $M = 0.49$, $95\%CI = [0.47, 0.51]$; iterations 21–25: $M = 0.78$, $95\%CI = [0.77, 0.80]$; increase: $+0.29$, $p < 0.001$, Cohen's $d = 2.67$). In contrast, Human-only chains remained substantially closer to the original seed story across all generations (iterations 1–5: $M = 0.27$, $95\%CI = [0.25, 0.29]$; iterations 21–25: $M = 0.49$, $95\%CI = [0.46, 0.51]$; increase: $+0.22$, $p < 0.001$, Cohen's $d = 1.68$), suggesting greater preservation of narrative continuity. The Human–AI condition showed an intermediate pattern (iterations 1–5: $M = 0.32$, $95\%CI = [0.30, 0.35]$; iterations 21–25: $M = 0.75$, $95\%CI = [0.73, 0.76]$; increase: $+0.42$, $p < 0.001$, Cohen's $d = 3.55$), combining substantial semantic exploration with greater fidelity to the original narrative than AI-only populations.

Finally, to characterize the trade-off between creativity and diversity explicitly, we examined how the relative

in which judgments are provided by individuals with relevant domain expertise (Amabile, 1982). However, when recruiting sufficiently large samples of experts is impractical, creativity can also be assessed through aggregated judgments from larger groups of non-expert raters (e.g., (51))

DRAFT

performance of the three population types changed as a function of the weight assigned to creativity versus semantic diversity (Figure 2D). We first z-scored creativity and semantic diversity separately and then introduced a trade-off parameter, w, that determines their relative contribution to a combined utility measure. At one extreme, AI-only populations performed best when creativity was weighted almost exclusively, whereas Human-only and Human–AI populations performed best when diversity was strongly prioritized. Between these extremes lies a broad range in which both creativity and diversity contribute substantially to utility. For values above $w > 0.30$, Human–AI populations achieved the highest combined utility, indicating the best overall balance between the two objectives.

Overall, these results reveal a fundamental tension between creativity, diversity, and transmission fidelity in the evolution of creative stories. AI-only populations consistently produced the most creative individual stories but rapidly diverged from the original story and converged on relatively homogeneous semantic content over time. Human-only populations exhibited the opposite pattern, preserving narrative continuity and sustaining collective semantic exploration but generating comparatively less creative outputs. Human–AI populations occupied an intermediate position between these extremes, combining substantial gains in creativity with sustained increases in diversity and moderate levels of divergence from seed story. Across a broad range of possible priorities, hybrid populations therefore provided the most effective balance between innovation and continuity, suggesting that hybrid societies may be particularly effective at balancing the competing demands of innovation and exploration, combining much of the creative advantage of AI systems with the diversity-preserving properties of human cultural transmission.

Robustness and Validation Analyses. We conducted additional analyses to assess the robustness of our findings across alternative measures of diversity, creativity evaluations, and AI configurations (see Supplementary Results for robustness and validation analyses). First, we examined lexical diversity using TF-IDF as an alternative to our primary semantic measure (Supplementary Figure S4 and S5). The two measures were positively correlated ($r = 0.57$, $p < .001$) and produced qualitatively similar population outcomes. Lexical diversity decreased over time in AI-only populations but increased in both Human-only and Human–AI populations. Human–AI populations also achieved the most favorable creativity–diversity trade-off across most weightings ($w > 0.43$), supporting the robustness of the hybrid advantage to an alternative measure of collective diversity.

Second, we re-evaluated all stories using an AI judge (Supplementary Figure S6). Human and AI creativity ratings were only modestly correlated ($r = 0.42$, $p < .001$), with the AI judge assigning systematically higher scores than human judges, particularly to AI-authored stories. Nevertheless, AI evaluations reproduced the main ordering across populations (AI-only > Human–AI > Human-only) and the same temporal pattern, with creativity increasing over time in AI-only and Human–AI populations but not Human-only populations. We further validated our findings using multidimensional human ratings of Novelty, Appropriateness, and Aesthetic Appeal (Experiment 3: $N = 377$; Supplementary Figures S7 and S8). AI-only stories scored significantly higher in Novelty than Human-only stories ($p = 0.039$, Cohen's $d = 0.30$), whereas Human–AI stories scored higher in Appropriateness than AI-only stories ($p = 0.02$, Cohen's $d = 0.33$); Aesthetic Appeal did not differ significantly across conditions.

Finally, we tested whether the observed dynamics depended on model parameters, model type, or prompting strategy (Supplementary Figures S9–S10). Increasing GPT's sampling temperature slightly increased diversity but substantially reduced creativity and narrative coherence, whereas lower temperatures had minimal effects. Repeating the AI-only condition with Claude (3.7 Sonnet) produced lower creativity and diversity, while providing GPT with extended task instructions increased diversity at the cost of greater semantic divergence from the original story. Thus, although model and prompting choices shifted absolute levels of creativity, diversity, and transmission fidelity, none provided a simple means of jointly maximizing these outcomes. Together, these analyses support the robustness of the main findings across alternative measures, evaluators, and AI configurations.

Humans change their behavior in response to AI interactions. Previous theoretical work suggests that human behavior may change in the presence of AI, both directly, through interaction with AI systems, and indirectly, through exposure to AI-generated content. Importantly, such effects may amplify the societal impact of AI beyond its direct outputs by reshaping human behavior itself (4). To test this prediction, we decomposed behavior within the hybrid Human–AI condition by analysing human and AI agents separately, and compared their behavior to the corresponding agents in the Human-only and AI-only conditions. If humans are influenced by AI-generated content, we should expect systematic differences in human behavior in the hybrid condition compared to the Human-only condition. Conversely, if AI behavior is shaped by exposure to human-generated content, AI agents in the hybrid condition should differ from those operating in the AI-only condition.

Figure 3 compares the behavior of humans and AI agents within hybrid populations to their counterparts operating in homogeneous populations. This allows us to directly test whether exposure to AI-generated content changes human behavior and, conversely, whether exposure to human-generated content alters AI outputs. Creativity ratings did not differ significantly between hybrid and homogeneous populations for either humans or AI agents (all ps > 0.05). In contrast, both humans and AI produced significantly more diverse outputs when embedded in hybrid populations than when interacting exclusively with agents of their own type. Human participants in hybrid populations generated more diverse stories than participants in Human-only populations (hybrid humans: $M = 0.60$, $CI = [0.56, 0.64]$; Human-only: $M = 0.52$, $CI = [0.49, 0.54]$; $p = 0.002$, Cohen's $d = 0.51$). Similarly, AI agents interacting with humans generated more diverse outputs than AI agents operating in AI-only populations (hybrid AI: $M = 0.48$, $CI = [0.45, 0.50]$; AI-only: $M = 0.37$, $CI = [0.35, 0.39]$; $p < 0.001$, Cohen's $d = 0.79$). These findings suggest that the primary effect of hybrid societies is not to increase the creativity

DRAFT

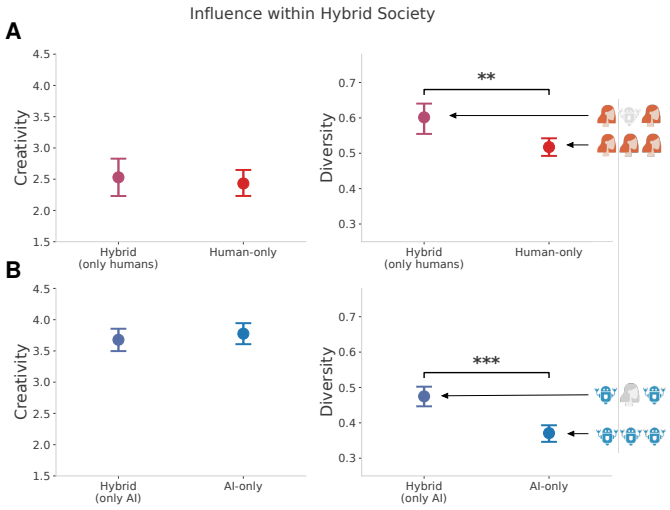


Fig. 3. Impact of AI on humans in hybrid societies. Splitting human and AI stories within the human-AI hybrid condition for creativity and diversity, comparing human-authored stories from the Human-AI condition against the Human-only condition (top row) and AI-authored hybrid stories against the AI-only condition (bottom row), using stories from the last 5 iterations. Points show condition means; error bars represent 95% bootstrap confidence intervals. $**, ***$ indicate $p < 0.01, 0.001$ (see Methods).

of individual contributions, but rather to increase the diversity of the collective search process.

Semantic Differences Between Human and AI Creative Productions. Our results show that hybrid human–AI populations achieved better balance between creativity and diversity compared to AI-only and Human-only populations. This pattern raises the question of what features of hybrid productions underlie this balance.

Qualitative analysis of the stories generated reveal clear differences across conditions (see Supplementary Figure S12–S15 for example stories from the three conditions alongside the original seed story). The human-only story remains closely connected to the original narrative, preserving both the protagonist John and the theme of the missing key. In the human–AI story, there is a pronounced style shift, but the protagonist John is retained and the cat named "Whiskers" is introduced. In contrast, the AI-only story completely abandons the original protagonist; elements from the original story are transformed such that "Whiskers" becomes a dancing entity, and the narrative shifts toward cosmic imagery, including black holes, the universe, celestial ballet, and dreams. All stories generated in the study are available in the OSF repository associated with this paper (see Data and Code Availability in Methods).

To quantify this, we examined the semantic space generated by embedding all stories using all-MiniLM-L6-v2, the same model we used for computing semantic diversity (see Methods). We then projected the resulting high-dimensional representations into two dimensions using the UMAP algorithm ((53); see Methods). In this space, each point represents a single story, and distances between points reflect semantic similarity in the original embedding space. Figure 4A shows the resulting semantic space coloured by experimental condition. In this space, the Human-only and AI-only stories occupy largely opposing regions, exhibiting minimal overlap. By contrast, the hybrid Human–AI stories span an intermediate area that interpolates between both extremes, overlapping substantially with the human-only and AI-only distributions.

To further examine semantic differences, Figure 4B shows the same space coloured according to four thematic clusters (see Methods). Figure 4C shows the average creativity ratings for each cluster, and Figure 4D the most frequent words, visualized as word clouds (see Methods). Overall, we can see that the clusters span a continuum ranging from themes closely aligned with the original seed narrative (clusters 1–2) to more novel adventure- and space-related themes (clusters 3–4). When rated by a separate group of independent raters, clusters aligned with the seed story received lower creativity ratings (cluster 1: $M = 2.38$, $CI = [2.24, 2.52]$; cluster 2: $M = 2.53$, $CI = [2.42, 2.64]$), whereas the more semantically distant clusters were judged as significantly more creative by independent raters (cluster 3: $M = 3.23$, $CI = [3.13, 3.33]$; cluster 4: $M = 3.42$, $CI = [3.34, 3.51]$; Figure 4C).

The nature and distribution of clusters reveal important differences in narrative themes and concepts across conditions. In clusters largely comprised by human-only stories (cluster 1: 99.3 %, $CI = [98.3, 100]$; cluster 2: 75.7 %, $CI = [71.7, 79.8]$), semantic content remained closely connected to the original narrative, preserving both the protagonist John and the theme of the missing key. In clusters largely comprised by stories in the AI-only experimental condition (cluster 4: 61.1 %, $CI = [57.4, 64.7]$), stories completely abandoned the original protagonist. Cluster 3, which was dominated by human-AI stories (54.2 percent, $CI = [49.7, 58.7]$) occupies an intermediate position: some AI-generated stories retained the protagonist, but increasingly emphasized the term whiskers, derived from the original cat element, which was subsequently generalized into alternative narrative contexts. These results suggest that humans and AI use different types of semantics in their stories.

Tracking the relative prevalence of thematic clusters over time and across the network (Figure 4E–F) reveals distinct population-level dynamics. In the human-only condition, the population is consistently dominated by less creative clusters that remain close to the original seed story, reflecting strong preferences for narrative continuity at the expense of novel content (consistent with the results of Figure 2C). In contrast, the AI-only population is initially dominated by Cluster 3 but progressively departs from it, ultimately converging on the most semantically distant, space-themed Cluster 4. This cluster is rated as the most creative by independent raters (Figure 4C), suggesting strong convergence toward an optimized but potentially, thematically narrow solution. Finally, in the hybrid human–AI condition, Cluster 3—representing an intermediate position between human-driven continuity and AI-driven novelty—decays more slowly than in the AI-only condition. This suggests that the hybrid system sustains an intermediate semantic solution for longer, balancing originality and coherence and supporting the previous results (Figure 2A-B). However, the novel creative theme introduced by AI (Cluster 4) eventually takes over in both the AI-only and human-AI conditions, potentially because of its high creativity. Supplementary Figure S16 presents a parallel analysis based directly on

DRAFT

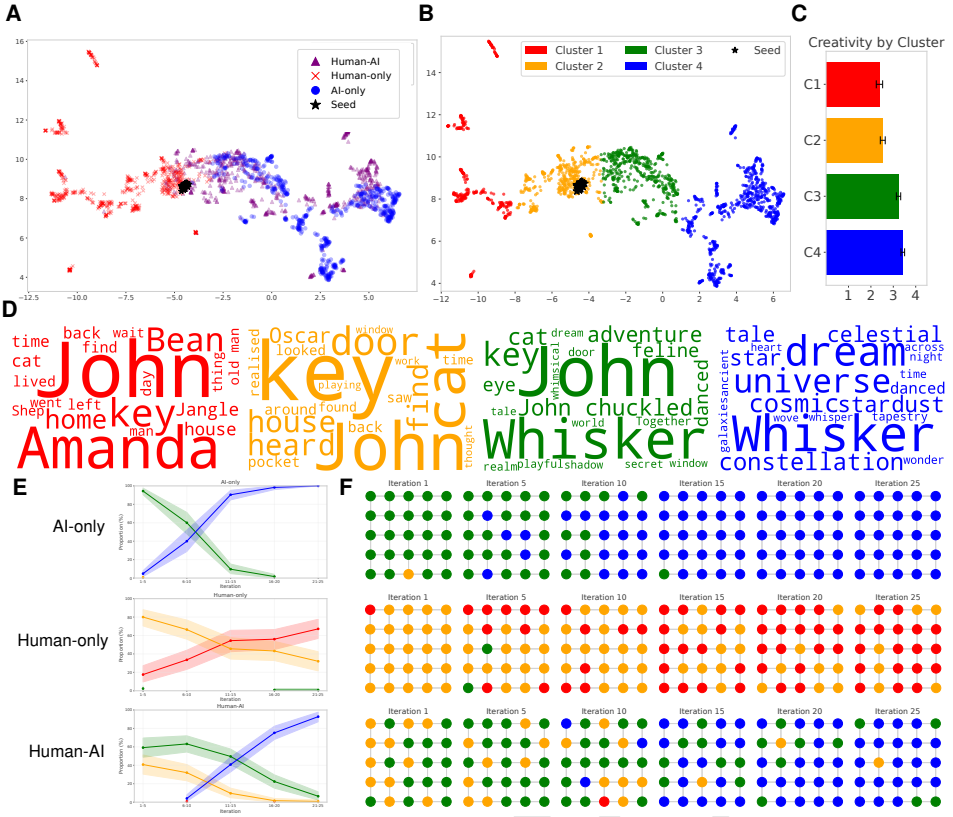


Fig. 4. Semantic differences between human and AI creative productions. **A** UMAP projection of the shared semantic embedding space reveals clear separation between human- and AI-generated stories, with human–AI hybrid stories occupying an intermediate region between the two clusters (see Methods). **B** Clustering solution for the semantic space of stories. Black stars show the location of the original seed story and its close variants. **C** Average creativity scores (as rated by humans) per cluster. Error bars indicate 95% bootstrap confidence intervals. **D** Word cloud representation of each cluster. **E** Cluster population proportion over time. **F** An example of spatial distribution of clusters over iterations for the three conditions.

TF-IDF term weights, without relying on clustering, and revealing qualitatively similar dynamics.

The Advantage of Population Diversity. Our control analyses (Figure S9) indicate that increasing stochasticity within a single AI model (by manipulating temperature) is insufficient to reproduce the balance between creativity and collective diversity. What, then, underlies the human-AI synergy in the hybrid condition? One hypothesis, motivated by our semantic analysis (Figure 4), is that the balance is achieved through conceptual diversity in the way humans and AI agents apply a "creative" edit: humans seek to maintain core elements of the original narrative, whereas the AI agents explore distant but intuitively creative themes. If this is indeed the case, then we should be able to achieve a similar outcome by mixing different types of AI models in a population. To test this possibility, we ran an additional experimental condition in which we mixed GPT and Claude agents (same models as before) and examined the resulting population characteristics.

The results for the mixed-AI condition, as well as the original AI-only (GPT-4o) and Claude conditions, are shown in Supplementary Figure **??**. As with the hybrid Human-AI condition, we see that mixing GPT-4o and Claude maintained a good creativity score while significantly boosting diversity (Supplementary Figure **??**A; see also Figure S17). This was further reflected in the tradeoff curve analysis as before (Supplementary Figure **??**B): for all values of the trade-off parameter above $w > 0.32$, the mixed-AI population achieved higher combined utility than either model alone. Thus, population-level diversity arising from a combination of AI models can maintain creative performance while enhancing diversity, mirroring the benefits observed in human–AI hybrid populations.

Discussion

We investigated how the collective outcomes of hybrid human-AI populations are shaped by the interaction between humans and AI agents within experimental social networks, focusing on collective creativity as an important test case. Our results show that human–AI networks outperform homogeneous populations by combining the complementary strengths of human and machine agents. AI-only networks sustained high levels of creativity but failed to increase diversity over time, whereas human-only populations showed the opposite pattern, increasing collective diversity but struggling to improve creativity. Hybrid populations combined these strengths, consistently achieving high creativity while preserving collective diversity. These benefits emerged from the interaction dynamics within hybrid populations, rather than simply from the presence of more capable individual agents. Some of these benefits could be reproduced by creating heterogeneous populations of AI agents, highlighting diversity among agents as an important determinant of collective creativity. These patterns were consistent across different dimensions of creativity, including novelty and appropriateness, and robust to manipulations of model

DRAFT

temperature and the use of alternative AI models (see Supplementary Results section for additional robustness and validation analyses).

Collective intelligence and cultural evolution of hybrid societies. Our study demonstrates that hybrid human–AI populations can exhibit collective outcomes that cannot be predicted from the behavior of human-only or AI-only populations in isolation. This finding connects with key discoveries in the collective intelligence literature, showing that interactions among heterogeneous agents can generate emergent group-level outcomes that are not reducible to individual behavior (5, 54). Our results extend this principle to hybrid societies in which humans and AI agents jointly participate in open-ended creative processes. Thus, large-scale experiments using naturalistic, open-ended tasks provide a powerful paradigm for studying human–AI collective intelligence under controlled conditions, allowing causal mechanisms to be isolated in ways that would be difficult in the real-world (4). Our work connects collective intelligence with the experimental study of cultural evolution. Our focus on narrative transmission (43, 44) builds on the transmission-chain paradigm that has been central to understanding how cultural information changes as it passes between individuals (7, 37, 42). By embedding this process within structured populations, we extend this paradigm from linear chains to networked environments in which agents can select and transmit information from multiple neighbours (13, 48). Our finding that human-only and AI-only populations produce systematically different distributions of artifacts is consistent with the expectation that the inductive biases (or priors) of agents play a key role in the dynamics of such cultural evolutionary processes (36, 48, 55).

Individual and collective creativity. The low levels of collective diversity observed in AI-only networks is particularly important and extends recent findings from studies of human–AI interaction in individual creativity tasks (23, 25). Although LLMs can produce highly novel outputs in isolation, when embedded within iterative social networks they tend to converge rapidly toward semantically similar regions of the solution space. These results highlight the importance of distinguishing between individual-level and collective-level creativity, and provide new empirical evidence for theoretical accounts of the dynamics and consequences of human–AI societies (2, 4). A system may appear highly creative when evaluated in isolation, yet fail to sustain diversity when placed within a networked, iterative environment. This distinction has important implications for how AI creativity is studied, evaluated, and deployed in real-world groups and collectives.

Mechanisms of collective creativity. What mechanisms enabled hybrid human–AI populations to achieve this balance between creativity and diversity? Validation experiments with human raters and computational semantic analyses (Figure 4) suggest that the advantage of hybrid systems emerges from the distinct and complementary ways in which humans and AI approached the task. AI agents frequently introduced radical semantic shifts, often disregarding core narrative elements in favour of novel inventions. In contrast, human participants tended to preserve key story components (e.g., maintaining the protagonist, John, and central objects like "cat" or "keys"; see Supplementary Figure S16). The interaction between these tendencies generated a dynamic balance between stability and novelty. In this sense, hybrid populations appear to resolve a classic exploration–exploitation tension: sustaining semantic exploration without collapsing into fragmentation, and preserving structure without stagnation (56, 57).

This finding suggests that combining agents with different behavioral biases can help a population explore a broader space of possible cultural outputs while preserving enough continuity for those outputs to remain coherent and transmissible. Hybrid human–AI populations appear to benefit not simply from greater heterogeneity, but from *complementary heterogeneity*: AI agents expand the space of possibilities, whereas humans constrain and stabilise that exploration. This finding is consistent with classical work on group creativity and collective intelligence, which shows that diversity in cognitive strategies and perspectives can enhance problem-solving and innovation (9–11). It is also consistent with experimental work showing that incorporating simple artificial agents into social networks can improve collective performance, for example, through "noisy" bots that inject stochastic variation (54) or bots that aggregate information across network neighbors (58).

More broadly, these results offer a more nuanced account of the impact of AI on culture and society. Although concerns about homogenization and overreliance on AI are well-founded (1, 59), our findings indicate that, when embedded within human collectives, AI can operate as a generative source of variation rather than convergence. The cultural consequences of AI thus depend critically on how it is integrated into human social systems.

Humans alter their behavior in the presence of AI. Another central question in our study concerns how the presence of AI alters human behavior within social systems. Recent theoretical work (3, 4) suggests that introducing AI agents into societies may reshape human perception, decision-making, and behavioral strategies over time. Empirical evidence from real-world domains supports this perspective. For example, research on chess show that human players adopt strategies originally developed by superhuman AI systems (60), leading to lasting shifts in play styles and competitive norms. These findings suggest that AI can influence long-term human behavior and cultural practices, even beyond direct human–AI interaction. Our results provide clear causal evidence of such effects within controlled social networks. In hybrid populations, humans who interacted with AI agents became substantially more diverse in their outputs compared to those in human-only networks (Figure 3). Interestingly, this effect also extended to AI agents in the hybrid population.

Agent diversity and alignment. Our results suggest that achieving a balance between creativity and collective diversity depends on combining agents with distinct generative strategies. We therefore asked whether similar benefits could be obtained by combining different AI models. To test this, we constructed mixed-agent networks composed of two distinct LLMs (Claude and GPT). We found that when operating in isolation, each model exhibited sustained low diversity over time, replicating the convergence

DRAFT

dynamics observed in AI-only populations. However, when embedded within a mixed-LLM network, their collective performance improved substantially in terms of diversity (Supplementary Figure **??**). These findings align with recent work highlighting the advantages of heterogeneous or mixed AI populations, where combining models with different architectures and inductive biases enhances performance (61). Overall, they suggest that population-level diversity—whether across humans and machines or across distinct AI systems—may represent a key design principle for sustaining creativity in hybrid societies.

Future work should examine this more systematically by manipulating model heterogeneity in a controlled manner. For example, models can be trained to emulate specific individuals—either by conditioning them on a detailed backstory or by fine-tuning them to replicate that person's performance patterns on particular tasks (62, 63). Importantly, these findings also highlight a methodological contribution of our work. While a growing body of research investigates collective behavior in social networks composed entirely of AI agents (64**?** , 65), our results underscore the importance of incorporating real human participants. Human responses are not fully captured by simulations of artificial agents alone, and meaningful collective dynamics may therefore be missed in AI-only network experiments.

Recent work in machine learning and cognitive science has emphasized the importance of aligning human and AI representations and behavior (66). Contemporary models are often explicitly optimized to match human judgments and expectations. Our results, however, reveal substantial misalignment. For example, correlations between human story ratings and AI evaluations were only moderate ($r = 0.42$, see Figure S6), and the AI relied on semantic, verbal, and narrative strategies that diverged markedly from human patterns (Figure 4). Importantly, this divergence was not detrimental in our case. When humans and AI systems were combined, hybrid groups outperformed single-agent systems (Figure 2), and similar gains could be reproduced using heterogeneous AI agents alone (Supplementary Figure **??**). These findings suggest that misalignment can, in some contexts, be advantageous: complementary skills and behavioral strategies may combine to produce superior collective performance. Future work should examine when such misalignment is productive (67), while further characterizing the functional roles that humans and AI can occupy in ecologically valid tasks, such as collaborative code writing (68), and identifying the conditions under which alignment versus diversity yields optimal collective outcomes.

Limitations and future directions. Our experimental societies necessarily simplify the complexity of real-world social systems (see Supplementary Discussion for an extended discussion of limitations). Participants interacted within static grid-based networks over a relatively short period, whereas real social networks are less structured and evolve as individuals form and adjust their social ties (69, 70). Our paradigm also masks whether partners are human or AI, although awareness of interacting with AI can alter individual behavior (4), and limits interaction primarily to the transmission and modification of creative artifacts, without direct communication or explicit collective goals. Future work could examine how adaptive networks, partner identity, richer communication, and shared goals shape the collective dynamics observed here.

Our study also captures relatively short-term cultural dynamics over 25 iterations. Although some trajectories showed signs of stabilization, others continued to evolve at the end of the experiment, indicating that longer experiments will be necessary to characterize their asymptotic behavior. Nevertheless, these short-term dynamics are theoretically informative, as real cultural systems are continually exposed to new individuals and information and may remain in flux rather than converge toward a fixed equilibrium. The trajectories observed here therefore provide a controlled view of how population composition shapes the early evolution of collective creativity and diversity, which is difficult to isolate using observational approaches.

A related limitation concerns the rapidly evolving nature of AI. The language models tested here should be viewed as snapshots of current capabilities rather than fixed scientific objects, and quantitative effects may change as models improve. Our primary goal, however, is not to characterize a particular model but to identify collective mechanisms arising from interactions among agents with different generative strategies, including homogenization, diversity, and nonlinear human–AI interactions. Testing these mechanisms across successive generations of AI systems will be important for establishing their generality. At the same time, language models offer a methodological advantage over traditional agent-based simulations. Humans and AI agents can receive the same task instructions and participate through the same experimental interface, allowing their behavior to emerge within closely matched experimental conditions.

Finally, the observed dynamics may depend on the creative domain. Although storytelling is widely used to study human creativity (19, 21), it proved relatively easy for current LLMs, which consistently outperformed humans in creativity ratings. This should not be interpreted as evidence that AI is intrinsically more creative, but highlights the limitations of using tasks designed for human cognition to benchmark AI creativity. Future research should test whether the collective dynamics identified here generalize to more demanding tasks and other creative domains, including music and visual art (36, 71). More broadly, our experimental framework can incorporate longer timescales, richer network structures, newer AI models, and different forms of social interaction, providing a flexible approach for studying how human and AI capabilities combine within evolving collective systems (see Supplementary Discussion for additional limitations and future directions).

Conclusion. AI systems are advancing at an unprecedented pace, with new tools rapidly developed and deployed across domains of cultural production, communication, and decision-making. While these systems are often evaluated for safety and performance at the level of individual users, there is little systematic assessment of their impact on collective behavior. As a result, society is effectively conducting a large-scale, uncontrolled experiment: tools optimized for individual use are immediately embedded in social systems without understanding their emergent

DRAFT

collective dynamics. This carries substantial risks, including diversity collapse, overstandardization, and human de-skilling. However, our findings show that hybrid human–AI systems can also create new opportunities for collective performance by sustaining greater creativity while combining complementary cognitive strategies from humans and AI in ways that neither achieves alone.

This challenges traditional notions of creativity as a property of individual minds and instead positions it as an emergent property of interaction dynamics within structured collectives. In this way, our work contributes to a deeper understanding of hybrid human–AI societies, and provides a foundation for identifying the conditions under which collaboration with AI enhances, rather than replaces, human creative potential.

Methods

Participants. A total of 1,473 online participants were included in the study analyses, with 879 contributing to the story writing experiments and 594 to the rating experiments. Additional details are provided in the Supplementary Methods section, *Human Participants*.

AI Models. We used two common large language models (LLMs), namely, GPT-4o (49) by OpenAI (Version: September 3, 2024) and Claude[§] by Anthropic (3.7 Sonnet; Version: February 19, 2025). These models were accessed using the official API services provided by OpenAI and Anthropic. Additional details are provided in the Supplementary Methods section, *Usage of AI Models*.

Online Experiment Implementation.

PsyNet. The behavioral experiments were designed using PsyNet[¶] (72), a framework for complex experiment design that builds on the Dallinger[‖] platform for experiment hosting and deployment. Additional implementaiton details are described in the Supplementary Methods section, *Implementation Details*.

Network Structure. In all story-writing experiments, social dynamics evolved on 5×5 square-lattice (grid) networks. Each network comprised 25 nodes, with each node storing a single story at every iteration, producing thus 25 chains. The grid had periodic boundaries so that all nodes have exactly four neighbors, and thus is equivalent to a torus (i.e., wrap-around boundary conditions). Edges connected horizontally and vertically adjacent nodes, creating a local neighborhood structure that constrained which stories could be sampled and rewritten at each trial.

Story Writing Task. Each experiment involved a 5×5 grid network initialized with 25 closely related seed stories, which were generated by GPT-4o as variations of a single initial story (see Supplementary Methods section *Story Writing Task*). Participants were randomly assigned to a node and presented with five stories associated with that node and its four neighbors. They selected their preferred story ('Please select your favorite version of the story from the options below.") and then modified it creatively ('Please creatively elaborate on the story, adding your own details and ideas"). Each participant contributed once, and their story replaced the previous story at that node, becoming available to subsequent participants. Participants had up to 15 min to complete the task. Networks evolved for 25 generations, with each generation completed when all 25 nodes had been updated.

We compared three conditions: *Human-only*, in which all nodes were occupied by human participants; *AI-only*, in which all nodes were occupied by LLM agents; and *Human–AI*, in which nodes were populated by both humans and LLM agents. In the Human–AI condition, incoming humans and AI agents were assigned to random available nodes and recruitment was approximately balanced rather than constrained to an exact 50/50 composition. Because human participants occasionally failed to complete assigned trials whereas AI agents reliably completed them, 254 of the 625 stories (40.6%) were human-authored and 371 (59.4%) AI-authored. Participants were not informed whether the stories they encountered had been generated by humans or AI agents.

AI Participant Procedure. AI agents received the same task instructions as human participants, with only the minimal adjustments necessary for implementation. They received no additional system prompt, persona, or role description. Further details are provided in the Supplementary Methods section, *AI Participant Procedure*.

Human Validation Task. Story quality was evaluated by independent human raters who were blind to experimental condition and authorship. Each participant rated 30 randomly selected stories from across conditions and iterations on a 5-point creativity scale ranging from 1 (not creative at all) to 5 (extremely creative).

In additional validation experiments, stories were rated using nine adjectives grouped into three dimensions: Aesthetic Appeal (liking, interesting, meaningful), Novelty (novel, original, surprising), and Appropriateness (appropriate, feasible, publishable). These dimensions capture complementary aspects of creative evaluation, particularly novelty and usefulness (23, 50, 73). Each participant rated 20 randomly selected stories on all nine adjectives and additionally judged whether each story had been generated by a human or AI.

Novelty, Appropriateness, and Aesthetic Appeal were analysed separately by averaging their three constituent adjective ratings. Internal reliability was acceptable to good for all dimensions (Novelty: $\alpha = .83$; Appropriateness: $\alpha = .72$; Aesthetic Appeal: $\alpha = .88$). Analyses focused on story-level scores from iterations 21–25, with pairwise differences between conditions assessed using Tukey's HSD tests. Supplementary Figure S7 presents the individual adjective ratings, and Figure S8 shows their dynamics across iterations. The same aggregation and comparison procedure was applied to judgments of whether stories were AI-generated.

Writing Expertise Validation Control. To test whether creativity judgments depended on writing expertise, Experiment 6 compared raters with higher ($N = 61$) and lower ($N = 50$) self-reported writing expertise. To compare human and AI evaluations, GPT-4o rated the same story set evaluated by human raters in Experiment 2 using the same 5-point creativity scale. Full procedures are provided in Supplementary Methods section *Writing Expertise*

[§] https://claude.com/product/overview
[¶] https://www.psynet.dev
[‖] https://github.com/Dallinger/Dallinger

DRAFT

Validation Control; results are shown in Supplementary Figure S2.

Temperature and Model Controls. To test whether the AI-only results depended on model parameters or model identity, we repeated the full AI-only experiment across six GPT-4o temperatures (0.1, 0.4, 0.7, 1.1, 1.4, 1.7) and four Claude 3.7 Sonnet temperatures (0.1, 0.4, 0.7, 1.0). Each condition comprised a complete 5×5 network evolving for 25 iterations (625 stories; Experiments 4a–4b in Table S1).

Mixed-LLM condition. The mixed-LLM condition followed the same procedure as the Human–AI condition, except that each node was filled by either a GPT-4o or Claude 3.7 Sonnet agent using the same model versions and parameters described above. Analyses were otherwise identical to those used for the main experimental conditions (Supplementary Figure **??**).

Alternative Prompt Control. To test sensitivity to prompting, we repeated the AI-only experiment while providing the AI with the complete instructions presented to human participants, including the explanatory material preceding the trial instructions. This differed from the main experiment, in which humans and AI received the same immediate task instructions but humans additionally encountered preceding instructional pages. The full prompt is reported in Supplementary Methods. Results of this control are shown in Supplementary Figures S9B and S10.

Statistical Analysis. All comparisons of experimental conditions and temporal dynamics used non-parametric bootstrapping with 10,000 resamples. Metrics were first calculated separately within each generation and then averaged across five-iteration windows (1–5, 6–10, . . . , 21–25). Point estimates were calculated from the observed data and 95% confidence intervals from the 2.5th and 97.5th percentiles of the bootstrap distributions.

Final performance was assessed over iterations 21–25, with pairwise differences calculated between experimental conditions. Change over time was quantified by comparing the final five iterations (21–25) with the first five (1–5). We report Cohen's d for mean-based comparisons, using pooled standard deviations for both between-condition contrasts and standardized changes over time. Full resampling procedures, including the handling of duplicate observations in bootstrap estimates of pairwise diversity, convergence analyses, and separate analyses of human and AI behavior within hybrid populations, are provided in the Supplementary Methods section *Statistical Analysis*.

Collective Diversity. Collective diversity was measured from pairwise semantic similarity between stories. Each story was encoded using the pretrained sentence-transformer model all-MiniLM-L6-v2 (52), yielding a 384-dimensional semantic embedding $\mathbf{v}_i$. For each iteration, we calculated the cosine similarity between all pairs of stories and defined collective diversity as one minus the mean pairwise similarity, such that higher values indicate greater semantic diversity. Values for groups of iterations were calculated by first computing diversity separately within each iteration and then averaging across iterations (see Supplementary Methods for full details).

As an alternative measure, we calculated lexical diversity using TF-IDF vectors (`TfidfVectorizer`) in place of semantic embeddings. Results are reported in Supplementary Figure S4; full methodological details and a comparison of the lexical and semantic measures are provided in Supplementary Methods.

Distance from Seed. To quantify transmission fidelity, we measured how far each story diverged from the seed narrative from which its chain originated. Using the same all-MiniLM-L6-v2 embedding space, distance from seed was defined as: $E_i = 1 - \cos\big(\mathbf{v}_i, \mathbf{v}_{\text{seed}(i)}\big)$, where higher values indicate greater divergence from the original narrative. Per-generation distance from seed was calculated by averaging E_i across all stories within a generation. This measure complements collective diversity by distinguishing diversification from loss of the original narrative structure (Supplementary Figure S10).

Quantifying Creativity–Diversity Trade-off. To quantify the trade-off between creativity and collective diversity (Figures 2D and **??**B), we combined standardized creativity and diversity into a preference-weighted utility score using data from iterations 21–25. Creativity was summarized at the chain level, whereas diversity was calculated at the population level for each generation, reflecting the different levels at which the two measures are defined.

For a preference weight $w \in [0, 1]$, utility was defined as $U(w) = (1 - w)\,\tilde{c} + w\,\tilde{d}$, where $\tilde{c}$ and $\tilde{d}$ denote standardized creativity and diversity, respectively; $w = 0$ values creativity only and $w = 1$ values diversity only. Varying w from 0 to 1 identifies the preference ranges over which each condition achieves the highest combined utility. Uncertainty was quantified using 10,000 bootstrap resamples. The corresponding analysis based on lexical TF-IDF diversity is reported in Supplementary Figure S4; full standardization and bootstrap procedures are described in Supplementary Methods.

Semantic Space Analysis. To characterize higher-level semantic structure, all stories from the AI-only, Human-only, and Human–AI conditions were represented using their 384-dimensional all-MiniLM-L6-v2 embeddings and projected into two dimensions using UMAP (53) (Figure 4A–B). We then applied k-means clustering with $k = 4$ to characterize major regions of this semantic space (Figure 4B–F). We calculated mean creativity within each cluster (Figure 4C), generated word clouds to characterize their narrative content (Figure 4D), and tracked the proportion of stories assigned to each cluster over iterations and conditions (Figure 4E). Full UMAP, clustering, and visualization parameters are provided in Supplementary Methods. As a complementary analysis, we tracked the prevalence of high-weight TF-IDF terms across iterations to visualize changes in narrative content over time (Figure S16; see Supplementary Methods).

Data, Materials, and Software Availability. All experiment and analysis code, as well as the human and AI data can be accessed through the following OSF repository(https://osf.io/ku9cj (74)) .

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

DRAFT

Research: Designing Smart Environments to Augment Collective Learning & Creativity" (Award BCS-2421385). ChatGPT version 5.2-5.6 (OpenAI) was used to assist manuscript editing and proofreading. Authors reviewed each of the edit suggestions, and approved the final version.

1. Z Epstein, et al., Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
2. L Brinkmann, et al., Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
3. I Sucholutsky, KM Collins, N Jacoby, BD Thompson, RD Hawkins, Using LLMs to advance the cognitive science of collectives. *Nat. Comput. Sci.* **5**, 704–707 (2025).
4. M Tsvetkova, T Yasseri, N Pescetelli, T Werner, A new sociology of humans and machines. *Nat. Hum. Behav.* **8**, 1864–1876 (2024).
5. D Centola, The network science of collective intelligence. *Trends Cogn. Sci.* **26**, 923–941 (2022).
6. B Thompson, B Van Opheusden, T Sumers, T Griffiths, Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* **376**, 95–98 (2022).
7. S Kirby, H Cornish, K Smith, Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci.* **105**, 10681–10686 (2008).
8. L Raviv, A Meyer, S Lev-Ari, Larger communities create more systematic languages. *Proc. Royal Soc. B: Biol. Sci.* **286** (2019).
9. S Page, *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. (Princeton University Press), (2008).
10. L Hong, SE Page, Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci.* **101**, 16385–16389 (2004).
11. D Bang, CD Frith, Making better decisions in groups. *Royal Soc. Open Sci.* **4**, 170193 (2017).
12. M Derex, MP Beugin, B Godelle, M Raymond, Experimental evidence for the influence of group size on cultural complexity. *Nature* **503**, 389–391 (2013).
13. R Marjieh, M Anglada-Tort, TL Griffiths, N Jacoby, Characterizing the interaction of cultural evolution mechanisms in experimental social networks. *arXiv preprint arXiv:2502.12847* (2025).
14. J Flack, P Ipeirotis, TW Malone, G Mulgan, SE Page, Editorial to the inaugural issue of collective intelligence. *Collect. Intell.* **1**, 26339137221114179 (2022).
15. OA Acar, A Tuncdogan, D van Knippenberg, KR Lakhani, Collective creativity and innovation: An interdisciplinary review, integration, and research agenda. *J. Manag.* **50**, 2119–2151 (2024).
16. J Henrich, *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. (Princeton University Press, Princeton, NJ), (2015).
17. K Sawyer, *Group genius: The creative power of collaboration*. (Basic books), (2017).
18. JW Burton, et al., How large language models can reshape collective intelligence. *Nat. Hum. Behav.* **8**, 1643–1655 (2024).
19. JC Kaufman, RJ Sternberg, *The Cambridge handbook of creativity*. (Cambridge University Press), (2010).
20. M Benedek, RE Beaty, DL Schacter, YN Kenett, The role of memory in creative ideation. *Nat. Rev. Psychol.* **2**, 246–257 (2023).
21. RE Beaty, YN Kenett, Associative thinking at the core of creativity. *Trends Cogn. Sci.* **27**, 671–683 (2023).
22. S Noy, W Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
23. AR Doshi, OP Hauser, Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
24. TF Eisenmann, et al., Expertise elevates AI usage: experimental evidence comparing laypeople and professional artists. *Int. J. Human–Computer Interact.* pp. 1–22 (2026).
25. L Meincke, G Nave, C Terwiesch, ChatGPT decreases idea diversity in brainstorming. *Nat. Hum. Behav.* **9**, 1107–1109 (2025).
26. B Uzzi, S Mukherjee, M Stringer, B Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
27. JR Hayes, Modeling and remodeling writing. *Writ. Commun.* **29**, 369–388 (2012).
28. L Flower, JR Hayes, A cognitive process theory of writing. *Coll. Compos. Commun.* **32**, 365–387 (1981).
29. FA Breithaupt, *The narrative brain: The stories our neurons tell*. (Yale University Press), (2025).
30. SM Rojas-Drummond, C Albarrán, KS Littleton, Collaboration, creativity and the co-construction of oral and written texts. *Think. skills creativity* **3**, 177–191 (2008).
31. A Yuan, A Coenen, E Reif, D Ippolito, Wordcraft: story writing with large language models in *Proceedings of the 27th International Conference on Intelligent User Interfaces*. pp. 841–852 (2022).
32. M Lee, P Liang, Q Yang, CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities in *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. (2022).
33. N Beguš, Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *Humanit. Soc. Sci. Commun.* **11**, 1–22 (2024).
34. JJY Chung, et al., Talebrush: Sketching stories with generative pretrained language models in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. pp. 1–19 (2022).
35. S Kirby, M Tamariz, H Cornish, K Smith, Compression and communication in the cultural evolution of linguistic structure. *Cognition* **141**, 87–102 (2015).
36. M Anglada-Tort, PM Harrison, H Lee, N Jacoby, Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Curr. Biol.* **33**, 1472–1486 (2023).
37. A Acerbi, JM Stubbersfield, Large language models show human-like content biases in transmission chain experiments. *Proc. Natl. Acad. Sci.* **120**, e2313790120 (2023).
38. J Perez, et al., When llms play the telephone game: Cultural attractors as conceptual tools to evaluate llms in multi-turn settings. *arXiv preprint arXiv:2407.04503* (2024).
39. J Perez, et al., Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882* (2024).
40. F Breithaupt, et al., Humans create more novelty than ChatGPT when asked to retell a story. *Sci. Reports* **14**, 875 (2024).
41. M Youngblood, J Nusz, J Simon, Dynamics of collective creativity in ai art competitions. *arXiv preprint arXiv:2605.17141* (2026).
42. J Xu, M Dowman, TL Griffiths, Cultural transmission results in convergence towards colour term universals. *Proc. Royal Soc. B: Biol. Sci.* **280**, 20123073 (2013).
43. F Breithaupt, B Li, JK Kruschke, Serial reproduction of narratives preserves emotional appraisals. *Cogn. Emot.* **36**, 581–601 (2022).
44. FC Bartlett, *Remembering: A study in experimental and social psychology*. (Cambridge university press), (1995).
45. J Barrett, M Nyhof, Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *J. Cogn. Cult.* **1**, 69–100 (2001).
46. L Brinkmann, et al., Hybrid social learning in human-algorithm cultural transmission. *Philos. Transactions Royal Soc. A* **380**, 20200426 (2022).
47. H Miton, N Claidière, H Mercier, Universal cognitive mechanisms explain the cultural success of bloodletting. *Evol. Hum. Behav.* **36**, 303–312 (2015).
48. A Whalen, TL Griffiths, Adding population structure to models of language evolution by iterated learning. *J. Math. Psychol.* **76**, 1–6 (2017).
49. A Hurst, et al., GPT-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
50. TM Amabile, Social psychology of creativity: A consensual assessment technique. *J. Pers. Soc. Psychol.* **43**, 997 (1982).
51. JM Berg, Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. *Adm. Sci. Q.* **61**, 433–468 (2016).
52. W Wang, et al., MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* **33**, 5776–5788 (2020).
53. L McInnes, J Healy, J Melville, Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
54. H Shirado, NA Christakis, Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370–374 (2017).
55. J Xu, TL Griffiths, A rational analysis of the effects of memory biases on serial reproduction. *Cogn. Psychol.* **60**, 107–126 (2010).
56. Y Hart, et al., Creative exploration as a scale-invariant search on a meaning landscape. *Nat. Commun.* **9**, 5411 (2018).
57. D Lazer, A Friedman, The network structure of exploration and exploitation. *Adm. Sci. Q.* **52**, 667–694 (2007).
58. A Ueshima, MI Jones, NA Christakis, Simple autonomous agents can enhance creative semantic discovery by human groups. *Nat. Commun.* **15**, 5212 (2024).
59. Z Chen, J Chan, Large language model in creative work: The role of collaboration modality and user expertise. *Manag. Sci.* **70**, 9101–9117 (2024).
60. M Shin, J Kim, B Van Opheusden, TL Griffiths, Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proc. Natl. Acad. Sci.* **120**, e2214840120 (2023).
61. J Wang, J Wang, B Athiwaratkun, C Zhang, J Zou, Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692* (2024).
62. M Binz, et al., A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).
63. A Kolluri, S Wu, JS Park, MS Bernstein, Finetuning LLMs for human behavior prediction in social science experiments in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. pp. 30084–30099 (2025).
64. F Fadaei, JC Moran, T Yasseri, Gender dynamics and homophily in a social network of LLM agents. *Philos. Transactions Royal Soc. A* (2026).
65. F Hashemi, MW Macy, An empirical study of collective behaviors and social dynamics in large language model agents (2026) Accepted at EACL 2026.
66. I Sucholutsky, et al., Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018* (2023).
67. C Li, et al., Human-AI synergy supports collective creative search (2026).
68. H Hu, et al., Why human guidance matters in collaborative vibe coding (2026).
69. A Almaatouq, et al., Adaptive social networks promote the wisdom of crowds. *Proc. Natl. Acad. Sci.* **117**, 11379–11386 (2020).
70. M McPherson, L Smith-Lovin, JM Cook, Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
71. S Kumar, et al., Comparing abstraction in humans and machines using multimodal serial reproduction in *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46, pp. 143–150 (2024).
72. P Harrison, et al., Gibbs sampling with people. *Adv. Neural Inf. Process. Syst.* **33**, 10659–10671 (2020).
73. S Harvey, JW Berry, Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity. *Acad. Manag. Rev.* **48**, 504–529 (2023).
74. S Shiiku, R Marjieh, M Anglada-Tort, N Jacoby, Data for hybrid human–ai societies balance creativity and diversity (2026).

DRAFT

Supporting Information for

Hybrid Human–AI Societies Balance Creativity and Diversity

Shota Shiiku, Raja Marjieh, Thomas L. Griffiths, Manuel Anglada-Tort, and Nori Jacoby

Nori Jacoby.
E-mail: kj338@cornell.edu

This PDF file includes:

Supporting text
Figs. S1 to S18
Tables S1 to S2
SI References

Supporting Information Text

Supplementary Methods

Participants and Experimental Implementation.

Human participants. Demographic information for all participant groups is provided in Table S2. All participants were recruited online via Prolific* using a nonprobability (opt-in) sample. Data were collected between September 2024 and December 2025. Eligibility criteria required that participants were based in the UK and identified English as their native language. Participants provided informed consent in accordance with the approved ethics protocol (Max Planck Ethics Council #202142), and were compensated at a rate of £9 per hour.

Online Experimental Implementation. Participants completed the experiments online through a web-browser interface which communicates with a Python server that manages the timeline of the experiment. The AI models interacted with the experiments through a custom `Bot()` abstraction in PsyNet which simulates a participant that progresses through the experiment timeline in real-time. The bots completed the required tasks by making appropriate API calls to the specified LLM model, while otherwise receiving the same inputs and producing outputs in the same format as human participants.

Participant recruitment was managed automatically by PsyNet. In the story writing experiments, participants (either humans or AI bots) were recruited until the target number of trials was reached. These corresponded to 25 nodes on a 5×5 square lattice, each evolving for 25 iterations (one trial per node × iteration). These parameters were determined through pilot experiments informed by experience from a related project (1), as they were found to provide a good balance between generating sufficient breadth and depth of creative outcomes and maintaining practical experimental costs.

Usage of AI Models. AI was used in this study as an experimental participant and, in a separate validation analysis, as an evaluator of stories. AI agents selected and modified stories by responding to the same task instructions presented to human participants; they did not recruit participants, determine participant eligibility, or make decisions about inclusion or exclusion of human data. All experimental procedures, prompts, analysis decisions, and validation analyses were designed and reviewed by the human research team. The primary evaluation of story creativity was conducted by independent human raters; AI-based evaluations were used only as a supplementary validation analysis. For a summary of the number of API calls per experiment, see Supplementary Table S2.

Additional Experimental Procedures.

Story Writing Task. The 25 seed stories were generated from a single initial story using GPT-4o. The initial story was selected from a set generated with the prompt "Please provide a CSV file with 8 very short stories (4 sentences)." The selected story was then provided to GPT-4o with the instruction "Create 25 variations of the story below and output as CSV." The resulting 25 closely related variations served as the seed stories for the network nodes (see Supplementary Figure S1 for examples and the OSF repository for all initial stories (2)).

The initial story was:

> "*As John reached for his front door, he realized his key was missing. Panic set in as he searched his pockets, but the key was nowhere to be found. Feeling defeated, he slumped against the door, only to hear a jingle from inside—his cat had been playing with the key all along.*"

In the Human–AI condition, recruitment was approximately balanced by design, with the scheduler spawning one AI agent per arriving human participant. Incoming humans and AI agents were assigned to randomly available nodes; the overall composition was not constrained to an exact 50/50 split, nor were local neighbourhoods constrained to contain a fixed number of humans and AI agents. In total, 164 of 418 recruited human participants were excluded: 106 attempted the task but did not produce a response meeting the validity criteria, whereas 58 did not contribute an attempted story. Thus, 254 of the 418 recruited human participants (60.8%) contributed a story to the final dataset.

For human-written responses in Experiment 1, validity criteria required responses to contain at least 10 words, include at least one complete sentence, and share at least five lexical items with the original story. Invalid responses were excluded and the corresponding network position was reassigned. Based specifically on these response-level validity criteria, 23 responses were excluded in the Human-only condition and 52 in the Human–AI condition. These response-level exclusions are distinct from the broader participant-level exclusion counts reported in Supplementary Table S2, which also include recruited participants who did not contribute an attempted story, for example because they did not complete an assigned trial or because no network position remained available due to over-recruitment.

Because an exact 50% Human–AI split was not essential to our research question, we did not implement a dynamic replacement mechanism to compensate for missing human contributions. Consequently, a modest imbalance emerged between human and AI contributions: AI agents reliably completed their assigned trials, whereas some recruited human participants did not contribute a valid story. The participant-level exclusion count should therefore not be interpreted as a direct measure of participant withdrawal or disengagement.

*www.prolific.com

AI Participant Procedure. AI agents received the same task instructions as human participants, with only the minimal adjustments necessary for implementation. At the selection stage, the model received the five candidate stories and the instruction 'Please select your favorite version of the story from the options below." At the writing stage, it received the selected story and the instruction 'Please creatively elaborate on the story, adding your own details and ideas." The writing prompt additionally specified approximately 50 words, matching the seed stories, and responses were capped at `max_tokens` = 100. All other sampling parameters were left at provider defaults (temperature = 1 for GPT-4o unless stated otherwise). We tested the robustness of these choices using alternative prompts, sampling temperatures, and models (see Temperature and Model Controls and Supplementary Methods).

Human and AI Validation Task. For the multidimensional validation experiment, each rater evaluated 20 randomly selected stories. All adjectives were presented together on a single page in a fixed order that was identical for all raters.

To compare human and AI evaluations, GPT-4o rated the same story set evaluated by human raters in Experiment 2 using the same 5-point creativity scale. We calculated the correlation between human and AI ratings and the signed difference in mean creativity scores, Δ, defined as the GPT-4o rating minus the human rating for the same stories.

Writing Expertise Validation Control. To test whether creativity judgments depended on writing expertise, Experiment 6 compared raters with higher ($N = 61$) and lower ($N = 50$) self-reported writing expertise.

Both groups rated stories from the same fixed pool of 100 stories using the 5-point creativity scale, with each participant evaluating 20 randomly selected stories. Agreement between groups was assessed using story-level correlations and an attenuation-corrected correlation based on split-half reliability estimates.

Before the main task, participants were asked whether they wrote professionally. Participants who responded "primarily" or "sometimes" were classified as higher-expertise raters, whereas those who responded "no" were classified as lower-expertise raters. Because relatively few participants qualified as higher-expertise raters under this definition, we recruited an additional batch in which participants were screened using the same question and excluded if they responded "no". Both groups rated the same fixed pool of 100 stories sampled from across experimental conditions and iterations, yielding approximately 10 ratings per story per group.

Because each rater group provides only a finite number of ratings per story, measurement noise in story-level means caps the correlation observable between groups even under identical underlying judgments. We therefore corrected the between-group correlation for attenuation using split-half reliability estimates. Across 10,000 random split-half iterations, raters within each group were randomly divided into two halves, and story-level mean ratings were computed separately for each half. Within each split we calculated the within-group split-half correlations (r_{11}, the reliability of the higher-expertise group; r_{22}, the reliability of the lower-expertise group), as well as the cross-group correlation (r_{12}) between one higher-expertise half and one lower-expertise half across the same stories. Correlations were averaged across splits using the Fisher z transformation. The attenuation-corrected correlation was then calculated as

$$\rho = \frac{r_{12}}{\sqrt{r_{11} r_{22}}}.$$

Results are reported in the main text and in Supplementary Figure S2.

Participant-Facing Questions and Instructions. The following reproduces every participant-facing page, question, and response option, in the order in which participants encountered them, for each experiment involving human participants. All experiments were implemented in PsyNet and deployed on Prolific. Text is reproduced verbatim from the deployed experiment code, including minor typographical errors in the original instructions. Story texts shown to participants varied by trial and are omitted; all stories are available in the OSF repository. Where an instruction page included an example screenshot of the upcoming task, this is indicated in brackets.

Recruitment and consent. Participants were recruited on Prolific with the following pre-screening filters: current country of residence, nationality, and country of birth all set to United Kingdom; first language English; desktop devices only. Every experiment began with the same consent form (sections: Purpose, Voluntary participation, Confidentiality, Research collaborations, Security, Risks/Discomforts, Benefits, Compensation, Your rights, Contact), followed by a consent confirmation. The Purpose statement read "The purpose of this study is to investigate how stories are transmitted and transformed as they pass from one person to another in a chain of communication." in Experiment 1 and "The purpose of this study is to investigate how much people validate creativity of stories." in Experiments 2, 3 and 6. The full consent text is included in the experiment code on OSF. Every experiment ended with the same demographic questionnaire (reproduced once, below).

Demographic questionnaire.

1. "How do you identify yourself?" — radio buttons: *Female*; *Male*; *Non-binary*; *Not specified*; *I prefer not to answer*.
2. "What is your age?" — free numeric entry (integer between 0 and 120).
3. "What country are you from?" — drop-down list of countries plus *Other country*.
4. "What is your current country of residence?" — drop-down list of countries plus *Other country*.
5. "What is your highest level of formal education?" — radio buttons: *None*; *High school*; *College*; *Graduate School*; *Postgraduate degree or higher*.

Experiment 1: Story Writing. Human participants in the Human-only and Human–AI conditions saw identical pages. Prolific advertisement title: "Story Experiment (Chrome Browser, 5 minutes, £0.70)"; description: "This is story experiment. You will be asked to memorize and reproduce story from memory. The task will take about 5 minutes.Please use incognito mode."

1. **Consent form** (see above).

2. **Instruction page 1.** Heading: "Welcome to the Story Experiment! - instruction 1". Text: "You will be participating in an experiment with short stories."

   - "You will be shown several variations of short stories to choose your favorite one."
   - "Please select intuitively within approximately 30 seconds."

   "Here is an example how this would look like:" [screenshot of the story-selection page].

3. **Instruction page 2.** Heading: "Story Experiment - instruction 2".

   - "You will be shown the short story you choose to read and remember."
   - "After a minute, the story will disappear and you'll be asked to reproduce it from memory."

   "Here is an example how this would look like:" [screenshot of the story-reproduction page].

4. **Instruction page 3.** Heading: "Story Experiment - instruction 3".

   - "You should replicate the story elaborate on it creatively."
   - "The story should be sufficiently long and closer in content to the original story."
   - "You are not allowed to copy and paste the stories."
   - "Please do your best to remember the story as accurately as possible." (shown in red)
   - "We will check that your reproduced story is of sufficient quality. Failure to do so may result in termination of the experiment." (shown in red)
   - "The experiment takes about 5 minutes."

   "The experiment will start on the next page. Please pay close attention to the instructions!"

5. **Story selection.** Prompt: "Please select your favorite version of the story from the options below:". Response: five push buttons, each showing the full text of one candidate story (the story currently held at the participant's node and at its four neighbouring nodes; see Methods), arranged vertically.

6. **Memorisation.** Prompt: "Try to remember this story:" followed by the selected story in a highlighted box (copying disabled). No response button; the page advanced automatically after 60 seconds.

7. **Story reproduction.** Prompt: "Please creatively elaborate on the story, adding your own details and ideas." Response: multi-line free-text box with copy-and-paste disabled. If the response contained fewer than 10 words or no complete sentence, submission was blocked with the message "Please revise the story, the story should be sufficiently long." and the participant could revise the text. If the response shared fewer than five words with the presented story it was recorded as invalid (see Failing Criteria).

8. **Feedback page.** Either "Thank you!" or, for a response that shared too few words with the presented story, "Your story was not close enough to the original! Next time please try to remember more details and reproduce them more accurately."

9. **End of task.** "You finished the experiment!"

10. **Demographic questionnaire** (see above).

11. **Writing experience.** "Do you have any professional writing experience (e.g., as a writer, editor, journalist)?" — radio buttons: *Yes, professional (e.g., writer, editor)*; *Yes, personal (e.g., blogging, journaling)*; *No.* Responses to this question were collected but not used in any analysis reported in this paper.

Experiment 2: Creativity Evaluation. Prolific advertisement title: "Validation Experiment (Chrome Browser, 12 minutes, £1.8)"; description: "This is validation experiment. You will be asked to validate creativity of the stories. The task will take about 12 minutes.Please use incognito mode."

1. **Consent form** (see above).
2. **Instruction page 1.** Heading: "Welcome to the Validation Experiment! - instruction 1". Text: "You will be participating in an experiment with short stories."
   - "You will be shown a story to validate its creativity."
   - "Please select intuitively from the options below."
   - "1 - Not creative at all"
   - "2 - Slightly creative"
   - "3 - Moderately creative"
   - "4 - Very creative"
   - "5 - Extremely creative"

   "Here is an example how this would look like:" [screenshot of the rating page].
3. **Instruction page 2.** Heading: "Validation Experiment - instruction 2".
   - "Please make your ratings intuitively, without overthinking or spending too much time."
   - "If you complete the experiment significantly faster than the estimated time, it may affect your eligibility for the full reward." (shown in red)
   - "The estimated time for this experiment is about 5 minutes."

   "The experiment will start on the next page. Please pay close attention to the instructions!"
4. **Rating trials** (30 per participant, in three blocks of 10). Each page showed a header with the trial number and block letter ("Trial n", "Block A/B/C"), the story in a highlighted box (copying disabled), and the prompt "Rate the creativity of the story on a scale of 1 to 5." Response: five push buttons labelled *1*, *2*, *3*, *4*, *5*, arranged horizontally. A progress bar displayed "Please hurry up. You exceed expected time." after 30 seconds and "You take too much time!" after 35 seconds.
5. **End of task.** "You have completed rating the stories!" followed by "You finished the experiment!"
6. **Demographic questionnaire** (see above).

Experiment 3: Multidimensional Evaluation. Prolific advertisement title: "Validation Experiment (Chrome Browser, 15 minutes, £2.25)"; description: "You will be asked to validate creativity of the stories. Takes ~15 minutes."

1. **Consent form** (see above).
2. **Instruction page 1.** Heading: "Welcome to the Validation Experiment". Text: "In this experiment, you will read short stories and evaluate them based on several criteria." Sub-heading "Scale":
   - "1 – Very low"
   - "2 – Low"
   - "3 – Medium"
   - "4 – High"
   - "5 – Very high"

   "Here is an example how this would look like:" [screenshot of the rating page].
3. **Instruction page 2.** Identical to Experiment 2.
4. **Rating trials** (20 per participant). Each page showed the story in a highlighted box and the prompt "Rate the story on the following dimensions. 1 = Low, 5 = High". Below it, eleven rating items were presented together on one page in the following fixed order, each with the title "Rate the story for: [item]" and a 1–5 rating scale anchored *Low* (1) and *High* (5): *Liking*; *Novel*; *Original*; *Surprising*; *Appropriate*; *Feasible*; *Publishable*; *Interesting*; *Meaningful*; *Intentionality*; *Do you think this was written by an AI?*. The nine adjective items and the AI-authorship item are analysed in the main text and Supplementary Results; the *Intentionality* item was collected but not used in any analysis reported in this paper.
5. **End of task.** "You have completed rating the stories!" followed by "You finished the experiment!"
6. **Demographic questionnaire** (see above).

Experiment 3: Control-Condition Creativity Evaluation. As stated in Temperature and Model Controls, creativity ratings for the stories generated in the temperature, model and prompt controls (Experiment 4) were collected from independent human raters as part of Experiment 3, with a single creativity item per story (20 stories per rater). Prolific advertisement and instruction pages were as in Experiment 6 below; the rating page showed the story in a highlighted box and the prompt "Rate the creativity of the story on scale of 1 to 5" with a 1–5 rating scale anchored *not creative at all* (1) and *extremely creative* (5).

Experiment 6: Writing Expertise. Two batches were run with the same rating task. Prolific advertisement title: "Story Creativity Rating (Chrome Browser, 15 minutes, £2.25)"; description: "You will read 20 short stories and rate the creativity of each. Takes ~15 minutes." (screened expert batch) or "You will read 20 short stories and rate the creativity of each, followed by a few short questions about you. Takes ~15 minutes." (unscreened batch).

1. **Consent form** (see above).

2. **Professional-writer question.** Lead-in: "Finally, please answer the following question." (screened batch) or "Before the experiment starts, please answer the following question." (unscreened batch). Question: "Do you consider yourself a professional creative writer?" with the explanation "By professional we mean that you receive payment, contracts, or formal institutional recognition for your writing work." — radio buttons: *Yes, I work primarily as a professional writer.*; *Yes, I sometimes work as a professional writer alongside other work.*; *No, I am not a professional writer.* (required).

3. **Creative writing achievements.** Lead-in: "Next, a short questionnaire about your creative writing achievements." Title: "Creative writing achievements"; instruction: "Please check all sentences that apply to you." — check boxes (required):

   - "I do not have training or recognized talent in this area."
   - "I have written an original short work (poem or short story)."
   - "My work has won an award or prize."
   - "I have written an original long work (epic, novel, or play)."
   - "I have sold my work to a publisher."
   - "My work has been printed and sold publicly."
   - "My work has been reviewed in local publications."
   - "My work has been reviewed in national publications."

   If the last item was checked, a follow-up numeric question appeared: "You indicated that your work has been reviewed in national publications. How many times?" (Creative Achievement Questionnaire, Creative Writing subscale; Carson, Peterson & Higgins, 2005.) Responses to this questionnaire were collected but not used in any analysis reported in this paper; only the professional-writer question was used to define the expertise groups.

4. **Screening.** Participants who answered "No, I am not a professional writer." saw "Thank you for your interest!" / "Unfortunately, you are not eligible for this study." / "The next pages explain how you will be compensated for the time you spent answering these questions." and were returned to Prolific with compensation for the time spent. In the unscreened batch all participants continued regardless of their answers.

5. **Instruction page 1.** Heading: "Welcome to the Validation Experiment". Text: "In this experiment, you will read short stories and rate the creativity of each one." "Here is an example how this would look like:" [screenshot of the rating page].

6. **Instruction page 2.** Heading: "Validation Experiment - instruction 2".

   - "Please make your ratings intuitively, without overthinking or spending too much time."
   - "If you complete the experiment significantly faster than the estimated time, it may affect your eligibility for the full reward." (shown in red)
   - "The estimated time for this experiment is about 15 minutes."

   "The experiment will start on the next page. Please pay close attention to the instructions!"

7. **Rating trials** (20 per participant, from a fixed pool of 100 stories). Each page showed the story in a highlighted box and the prompt "Rate the creativity of the story on scale of 1 to 5". Response: a 1–5 rating scale anchored *not creative at all* (1) and *extremely creative* (5) (required).

8. **End of task.** "You finished the experiment!"

9. **Demographic questionnaire** (see above).

Statistical Analysis. For bootstrap analyses of diversity, resampling with replacement can result in the same original story being sampled more than once. Comparisons between duplicate samples of the same original story were excluded because they are identical by definition and would artificially lower estimated diversity. This exclusion applies only to bootstrap samples and does not affect the observed diversity estimate. Two-sided p-values were computed from the bootstrap distribution of differences as twice the smaller of the two tail proportions, $\Pr(\Delta \leq 0)$ and $\Pr(\Delta \geq 0)$, capped at 1.

Collective Diversity. Collective diversity was calculated from pairwise semantic similarity between stories. Each story was encoded using the pretrained sentence-transformer model all-MiniLM-L6-v2 (3), yielding a 384-dimensional semantic embedding $\mathbf{v}_i$. For each iteration t, we calculated cosine similarity between all unordered pairs of stories and averaged these values to obtain the mean within-generation similarity,

$$S_t = \frac{2}{n_t(n_t - 1)} \sum_{i<j} \cos(\mathbf{v}_i, \mathbf{v}_j),$$

where n_t is the number of stories in iteration t. For an iteration group g (e.g., iterations 1–5), similarity was calculated by averaging the independently computed within-generation similarities,

$$S_g = \frac{1}{|g|} \sum_{t \in g} S_t.$$

Collective diversity was then defined as

$$D_g = 1 - S_g,$$

such that higher values indicate greater semantic diversity. Importantly, stories were not pooled across iterations before calculating pairwise similarity; diversity was first calculated separately within each iteration and then averaged across the iterations in each group.

As an alternative measure of collective diversity, we repeated the same procedure using TF–IDF vectors (`TfidfVectorizer`; (4)) in place of semantic embeddings. Results are reported in Supplementary Figure S4. The correlation between lexical and semantic similarity reported in the Results ($r = 0.57$) was computed across story pairs, pooling pairwise similarity values over all generations of the three main conditions.

Convergence and Stabilisation. To assess late-stage change, we compared iterations 16–20 with iterations 21–25 separately for each condition and outcome using the same two-sided bootstrap procedure. Because each window contains only five generations, these analyses were interpreted as evidence of continued late-stage change rather than definitive tests of asymptotic convergence.

Human and AI behavior in Hybrid Populations. To ensure that differences in diversity were not driven by the smaller number of human- or AI-authored stories in the hybrid condition, we repeated the analysis after randomly down-sampling the Human-only and AI-only conditions to match the corresponding per-generation sample sizes in the hybrid subgroups. The results were unchanged: down-sampling shifted the mean diversity of each condition by less than 0.0004.

Quantifying Creativity–Diversity Trade-off. The two measures were calculated at different levels because they capture different properties of the population. Creativity was averaged within each chain across iterations 21–25, yielding 25 creativity values per condition, one for each chain. Diversity is a population-level measure and was therefore represented by the five per-generation diversity values from iterations 21–25, yielding five diversity values per condition. For the Claude-only condition, where selected stories were frequently reproduced verbatim, identical stories within a generation were counted only once when calculating diversity.

Within each bootstrap resample, chain-level creativity means and generation-level diversity values were redrawn with replacement within each condition. The condition means were recomputed, creativity and diversity were separately z-scored across conditions, and the utility curves were recalculated. Plotted curves are the pointwise means of these bootstrap utilities; shaded bands are the pointwise 2.5th and 97.5th percentiles. The main-text panels use semantic diversity from all-MiniLM-L6-v2 embeddings; the corresponding lexical analysis based on TF–IDF is reported in Supplementary Figure S4.

Comparison of creativity dimensions across groups. Cronbach's α was computed over individual rating occasions (one set of adjective ratings per story), treating the three adjectives within each dimension as items and pooling across the three main conditions and all 25 iterations.

Semantic Space Analysis. UMAP was applied to the all-MiniLM-L6-v2 story embeddings using the package `umap-learn`[†], with metric = euclidean, n_neighbors = 15, min_dist = 0.1, and n_components = 2. On this two-dimensional representation, we applied k-means clustering with $k = 4$ (`scikit-learn`; $n_{init} = 50$, fixed random seed). Clusters were ordered from left to right according to their mean position on the first UMAP axis. Word clouds were generated using the package `wordcloud`[‡] with max_words = 20.

[†] https://umap-learn.readthedocs.io/en/latest/

[‡] https://pypi.org/project/wordcloud/

As a supporting analysis, we analyzed leading terms in stories and changes in their frequency over iterations (Figure S16). Each story was transformed into a TF–IDF weighted vector representation with English stopwords removed, and the 17 terms with the highest mean TF–IDF scores across the full corpus were selected as global key terms. Using this fixed vocabulary, we then re-computed TF–IDF scores separately for each condition across the chronological sequence of 25 iterations. For each story, we recorded whether each selected term appeared with a non-zero TF–IDF weight.

Supplementary Results

Robustness and Validation Analyses. To assess the robustness and generality of our findings, we conducted a series of additional analyses examining (i) alternative measures of diversity, (ii) the consistency of creativity evaluations across human and AI judges, (iii) multidimensional assessments of creativity, and (iv) robustness across different model configurations and prompting strategies (see Methods for full details).

Alternative Measures of Collective Diversity. Because narrative diversity can be defined at multiple representational levels, we examined collective diversity using an alternative embedding technique based on TF-IDF scores (see Methods; Supplementary Figure S4). Whereas our primary semantic measure captures variation in the meanings and themes of stories, TF-IDF primarily captures lexical variation in vocabulary use. The two measures were positively correlated ($r = 0.57$, $p < .001$) but were not interchangeable, consistent with lexical and semantic diversity capturing complementary aspects of cultural variation. For example, the use of different synonyms can increase lexical diversity while contributing relatively little to semantic diversity (see Supplementary Figure S5 for examples of stories varying along these two measures). TF-IDF revealed greater diversity during the early generations of AI-only and Human–AI populations than was observed using semantic embeddings, suggesting that AI systems initially introduce substantial lexical variation while remaining relatively similar in their underlying meanings and themes. Because our theoretical interest concerns diversity in the latter rather than vocabulary use per se, we focused on semantic diversity in the primary analyses. Crucially, however, both measures produced qualitatively similar population dynamics. TF-IDF diversity decreased over time in the AI-only condition but increased in both the Human-only and Human–AI conditions, with Human–AI diversity reaching the highest levels around iteration 10 before declining slightly thereafter. Consistent with this pattern, Human–AI populations achieved the highest combined utility for all values of $w > 0.43$ in the corresponding creativity–diversity trade-off analysis, further supporting the conclusion that hybrid populations provide a favorable balance between creativity and collective diversity.

Human and AI Evaluations of Creativity. As a complement to the human evaluations, we re-evaluated all stories using GPT-based ratings (Supplementary Figure S6; see Methods) and assessed the degree of alignment between human and AI evaluations. Human and AI ratings were only modestly correlated ($r = 0.42$, $p < .001$), and the AI judge assigned systematically higher creativity scores than human judges across all story types (mean difference across iterations: $\delta = 0.56$, $CI = [0.50, 0.61]$). Importantly, this positivity bias was strongest for AI-authored stories ($\delta = 0.78$, $CI = [0.69, 0.87]$), intermediate for hybrid stories ($\delta = 0.58$, $CI = [0.50, 0.67]$), and weakest for human-authored stories ($\delta = 0.31$, $CI = [0.21, 0.42]$). Thus, the AI judge evaluated AI-authored content disproportionately more favorably relative to human judges. Despite these differences in absolute ratings, AI evaluations yielded the same qualitative ordering across populations as human evaluations: AI-only stories received the highest creativity scores, followed by Human–AI and Human-only stories. Moreover, creativity increased over time in the AI-only and Human–AI populations but not in the Human-only population. The principal conclusions concerning differences in creativity across population types therefore did not depend on whether stories were evaluated by human or AI judges.

Multidimensional Evaluation of Creative Output. To better characterize the perceived qualities underlying the generated stories, we conducted an additional validation experiment in which a separate group of human raters evaluated all stories using nine descriptive adjectives (Experiment 3: $N = 377$; see Methods). These adjectives were grouped into three dimensions: Novelty (Novel, Original, Surprising), Appropriateness (Appropriate, Feasible, Publishable), and Aesthetic Appeal (Liking, Meaningful, Interesting). Prior research has shown that multidimensional evaluations using such trait descriptors can capture distinct aspects of creative output beyond a single global creativity rating (5–7). Ratings for individual adjectives and their dynamics across iterations are presented in Supplementary Figures S7 and S8. The three dimensions showed good internal reliability, with Cronbach's α values of .83, .72, and .88 for Novelty, Appropriateness, and Aesthetic Appeal, respectively. AI-only populations received significantly higher Novelty scores than Human-only populations ($p = 0.039$, Cohen's $d = 0.30$), whereas Human–AI populations scored significantly higher on Appropriateness than the AI-only condition ($p = 0.02$, Cohen's $d = 0.33$). Aesthetic Appeal did not differ significantly across conditions. Together, these findings suggest that AI systems preferentially amplify novelty, whereas hybrid populations enhance appropriateness. We also asked participants to judge whether they thought each story was AI-generated. AI-only stories were perceived as the least human-like: raters were most likely to identify AI-only stories as AI-generated, followed by Human–AI stories and then Human-only stories. All pairwise comparisons were significant (AI-only vs. Human–AI: $p = 0.003$, Cohen's $d = 0.43$; AI-only vs. Human-only: $p < 0.001$, Cohen's $d = 0.88$; Human–AI vs. Human-only: $p = 0.002$, Cohen's $d = 0.42$).

Robustness Across Model Configurations and Prompting Strategies. Because AI-generated outputs can be sensitive to model-specific parameters and prompting choices, we conducted a series of robustness analyses examining the effects of sampling temperature, alternative language models, and prompt variations (Supplementary Figures S9–S10; see Main Methods). First, we varied GPT's sampling temperature to test whether increasing stochasticity could increase diversity while maintaining creativity.

Increasing generation temperature above GPT's standard default values (> 1.1) slightly increased diversity (Supplementary Figure S9B) but substantially reduced creativity and narrative coherence (Supplementary Figures S9B and S11). Reducing temperature, in contrast, had minimal effects on creativity and diversity. Thus, increasing stochasticity could increase diversity to some extent, but at a cost to other properties of the generated narratives. Second, we replicated the AI-only condition using an alternative language model (Claude-3.7 Sonnet; version 2025-0219) and swept across Claude's available temperature range.[§] Claude produced substantially lower diversity and creativity scores across these configurations, indicating that the creativity–diversity trade-off could not be readily overcome simply by switching to an alternative language model. Third, we repeated the AI-only condition using an extended prompt. In the main experiments, human and AI participants received the same task prompt on the final instruction page, but human participants additionally encountered preceding pages containing further explanations and rationale. In this control experiment, the AI model received both the shared task prompt and the complete set of additional instructions on every trial (see Methods for full prompts). The extended prompt increased diversity (Supplementary Figure S9B), but the resulting narratives retained less of the original seed story, as indicated by greater semantic distance from the seed (Supplementary Figure S10). Together, these controls show that model parameters, model identity, and prompting strategy can shift the absolute levels of creativity, diversity, and transmission fidelity, but none provided a simple means of maximizing these outcomes simultaneously. The observed trade-offs are therefore robust across several alternative AI configurations and do not appear to arise solely from the particular model settings used in the main experiments.

Supplementary Discussion

Limitations and Future Directions.

Simplifying real-world social interaction. Our study aimed to simulate social-naturalistic dynamics within controlled experimental societies. However, this approach necessarily departs from the complexity of real-world social systems. Participants interacted within bounded and relatively short-lived experimental societies, whereas real-world social systems involve repeated, long-term interactions in which individuals continuously adapt to one another. Our simulated social networks also used a static grid-based topology. Real social networks, such as friendship groups and workplace relationships, tend to exhibit modular and irregular structures with connections that evolve over time, allowing individuals to actively shape their interactions by forming and adjusting social ties rather than being confined to static connections (8, 9). Our paradigm also masks information about agent identity, including whether one's partner is human or AI. Prior research suggests that awareness of interacting with an AI system can alter individual behavior (10), raising the possibility that perceived partner identity could shape the collective dynamics observed here. Future experiments could independently manipulate actual and perceived partner identity to determine how knowledge about the presence of AI influences individual and population-level behavior. Furthermore, our paradigm deliberately limits direct communication between participants. Interaction occurs primarily through the evolving creative artifacts themselves rather than through the continuous and reciprocal communication characteristic of many real-world teams. We also do not explicitly impose a shared collective goal connecting group members, although extensive research on team creativity suggests that social and motivational factors can play an important role (11). Against this background, our findings are notable in showing that even relatively minimal forms of human–AI interaction can give rise to complex collective dynamics. Future work could extend the paradigm by introducing synchronous interaction, explicit collective goals, and richer forms of communication between human and AI group members (12, 13).

Timescale of cultural evolution. Our study examines short-term dynamics of cultural evolution over 25 experimental iterations, whereas real-world cultural systems evolve over substantially longer and more complex timescales. Convergence analyses suggest that some trajectories begin to stabilize within the observed window, whereas others continue to evolve, indicating that additional iterations will be necessary to characterize their longer-term and potentially asymptotic behavior. At the same time, these short-term dynamics are themselves theoretically informative. Real-world cultural systems are continually exposed to new individuals, information, technologies, and environmental pressures and may therefore remain in flux rather than converge toward a fixed equilibrium. From this perspective, the trajectories documented here characterize how different population compositions shape the early evolution of collective creativity and diversity in controlled cultural environments, dynamics that are difficult to isolate using observational approaches. Longer experiments will be important for determining whether the differences between Human-only, AI-only, and Human–AI populations persist, diminish, or qualitatively change over extended timescales.

Rapidly evolving AI systems. Our experiments use frontier language models available at the time of data collection. Large language models are rapidly evolving sociotechnical systems rather than fixed scientific objects, and their capabilities can change over timescales considerably shorter than those of scientific publication. The specific models evaluated here should therefore be viewed as representative snapshots of contemporary AI capabilities rather than definitive endpoints. Our goal, however, is not primarily to characterize a particular model but to develop an experimental framework for identifying principles governing collective human–AI interaction. Phenomena such as homogenization, population diversity, and nonlinear effects arising from interactions between agents with different generative strategies may extend beyond any particular model architecture, although their magnitude and form may change as AI systems develop. Testing successive generations and families of AI models within the same experimental framework will therefore be important for establishing which collective dynamics generalize across

[§] The two models provide different temperature ranges, [0,2] for GPT and [0,1] for Claude, accounting for the different sweeps in Supplementary Figure S9.

technological change. We nevertheless employ frontier language models because they provide a close approximation to the generative AI systems with which people increasingly interact in practice. Their use also provides a methodological advantage: human and AI agents can receive the same task instructions and participate through the same experimental interface. This symmetry enables controlled comparisons under closely matched experimental conditions and provides a complementary approach to traditional agent-based simulations, in which agent behavior is typically specified analytically rather than emerging from the same tasks presented to human participants. Moreover, unlike simple, task-specific bots used in previous studies (12, 14), frontier language models can engage in open-ended generative behavior, making it possible to study collective dynamics in domains where the space of possible responses cannot be specified in advance.

Extending the experimental framework. Importantly, many of these constraints are not inherent to our methodological approach. The framework can be extended to incorporate more complex network topologies (15), substantially longer evolutionary timescales, and experimental manipulations designed to isolate specific cultural-evolutionary mechanisms, such as transmission and selection (1). It can also be readily applied to newer generations of AI models as they become available. Our ongoing work builds on these findings by introducing explicit social feedback and examining its influence on creativity (16–18). We are also extending the paradigm to tasks with objective evaluation criteria, which permit more precise computational analyses of collective performance (19), as well as to societally consequential domains such as collaborative coding (20). These extensions will help establish how network structure, social feedback, task demands, and evaluation criteria interact with population composition to shape collective behavior.

Generality across creative tasks. A final important consideration is the nature of the creative task itself. Although our storytelling paradigm is consistent with tasks traditionally used to study human creativity (5, 21), it proved relatively easy for current LLMs, which consistently outperformed humans in creativity ratings. This does not necessarily imply that AI is intrinsically more creative than humans. Rather, it highlights a broader challenge in using tasks developed to characterize human cognition as benchmarks for comparing human and AI creativity. Future research should therefore move beyond relatively simple tasks that may be well suited to assessing individual human creativity but less capable of differentiating contemporary AI systems or capturing the full range of human–AI synergies. This includes expanding beyond language to multimodal creative domains such as music (22) and drawing or visual art (17, 18, 23–25), as well as developing more complex tasks that meaningfully challenge both humans and AI. Examining collective dynamics across such domains will be necessary to determine which findings reflect general properties of heterogeneous creative populations and which depend on the particular affordances of storytelling and language-based interaction.

Supplementary Tables and Figures

x

Table S1. Design overview of all experiments.

Experiment	Task	Condition	Chains	Iterations	Stories	N_{humans}	N_{bots}
Experiment 1a	Writing	Human	25	25	625	625	0
Experiment 1b	Writing	AI	25	25	625	0	625
Experiment 1c	Writing	Human–AI	25	25	625	254	371
Experiment 2	Validation	Human	–	–	1875	106	0
Experiment 3	Validation	Human	–	–	9375	377	0
Experiment 4a	Writing	GPT	25	25	4375	0	4375
Experiment 4b	Writing	Claude	25	25	2500	0	2500
Experiment 4c	Writing	GPT–Claude	25	25	625	0	625
Experiment 5	Validation	GPT	–	–	1875	0	94
Experiment 6a	Validation	Human (high expertise)	–	–	1000	61	0
Experiment 6b	Validation	Human (low expertise)	–	–	1000	50	0

All chain-based experiments (Experiment 1 and 4) were conducted on a 5×5 square-lattice network (25 nodes). Experiment 1 compares Human, AI, and Human–AI collaborative story evolution. Experiment 2 and 3 are one-shot human evaluations with independent human judgments. Experiment 4 compares GPT, Claude, and mixed GPT–Claude evolution. Experiment 5 is a validation experiment with GPT evaluations. Experiment 6 compares high expertise and low expertise human raters on a shared pool of 100 stories.

Table S2. Summary of participant counts and AI model API calls per experiment condition.

Experiment	Task	Condition	N (Excluded)	N_{calls}	Age (Mean ± SD)	Sex (F/M)
Experiment 1a	Writing	Human	625 (173)	0	41.53 ± 13.61	354 / 326
Experiment 1b	Writing	AI	0	625	–	–
Experiment 1c	Writing	Human-AI	254 (164)	371	43.15 ± 14.39	148 / 187
Experiment 2	Validation	Human	106 (2)	0	44.38 ± 14.27	40 / 57
Experiment 3	Validation	Human	377 (8)	0	46.61 ± 14.84	158 / 186
Experiment 4a	Writing	GPT	0	4375	–	–
Experiment 4b	Writing	Claude	0	2500	–	–
Experiment 4c	Writing	Mixed-LLM	0	625	–	–
Experiment 5	Validation	AI	0	94	–	–
Experiment 6a	Validation	Human (expert)	61 (266)	0	35.36 ± 10.17	23 / 37
Experiment 6b	Validation	Human (non-expert)	50 (9)	0	42.54 ± 12.93	30 / 20

All human participants were recruited via Prolific and were based in the United Kingdom. 13 participants reported non-binary, prefer-not-to-say, or other gender responses (Experiment 1 Human: 8; Experiment 1 Human-AI: 3; Experiment 3: 1; Experiment 6 expert: 1). "Excluded" denotes recruited participants who were not included in the final dataset because they did not complete an assigned story—for example, because they timed out or withdrew from the experiment—or because over-recruitment resulted in no remaining available node in the network. The exclusion rate was somewhat higher in the Human–AI condition than in the Human-only condition. This likely reflects the substantially longer stories encountered in the hybrid condition, resulting from the tendency of AI agents to generate longer and more elaborate stories. These longer inputs may have slightly increased the burden on human participants, leading to a modestly higher attrition rate.

1. Additional figures

Seed Stories

John stretched his hand to the front door, only to discover his key was gone. Frantically, he rummaged through his pockets. Just as despair was setting in, a familiar jingle from inside caught his ear - his mischievous cat had taken the key.	At the moment John touched the front door, he noticed his key was absent. Anxiety surged as he patted down his pockets, finding nothing. In resignation, he leaned on the door, then a jingling sound came from inside - his cat had the key.
John extended his hand to open the door, but his key was nowhere to be found. After searching his pockets with growing concern, he leaned against the door, dejected. Just then, a jingle from inside revealed his cat had been playing with the key.	As John approached the door, he found his key was missing. He searched his pockets desperately but to no avail. Dejected, he leaned against the door and heard a jingle from inside - his cat had the key.
John reached out for the door and found his key missing. Panic set in as he searched his pockets, finding nothing. In defeat, he leaned on the door, and heard a jingle from inside - his cat had the key.	John went to open the door but discovered his key was gone. Panic set in as he searched his pockets, finding nothing. Defeated, he leaned on the door and heard a jingle from inside - his cat had the key.
John reached for the door and found his key missing. Frantically, he checked his pockets. In defeat, he leaned against the door and heard a jingle from inside - his cat had the key.	At the moment John touched the door, he noticed his key was gone. Anxiety surged as he searched his pockets, finding nothing. He leaned against the door in defeat and heard a jingle from inside - his cat had the key.
John extended his hand to open the door, but his key was nowhere to be found. After searching his pockets with growing concern, he leaned against the door, feeling defeated, then heard a jingle from inside - his cat had the key.	As John approached the door, he found his key was missing. He searched his pockets desperately but to no avail. He leaned against the door in defeat and heard a jingle from inside - his cat had the key.

Fig. S1. Seed stories used to initialize the networks. Ten of the 25 close variations of the original seed story (see Methods) that were assigned to the 25 nodes of the network at the first generation.

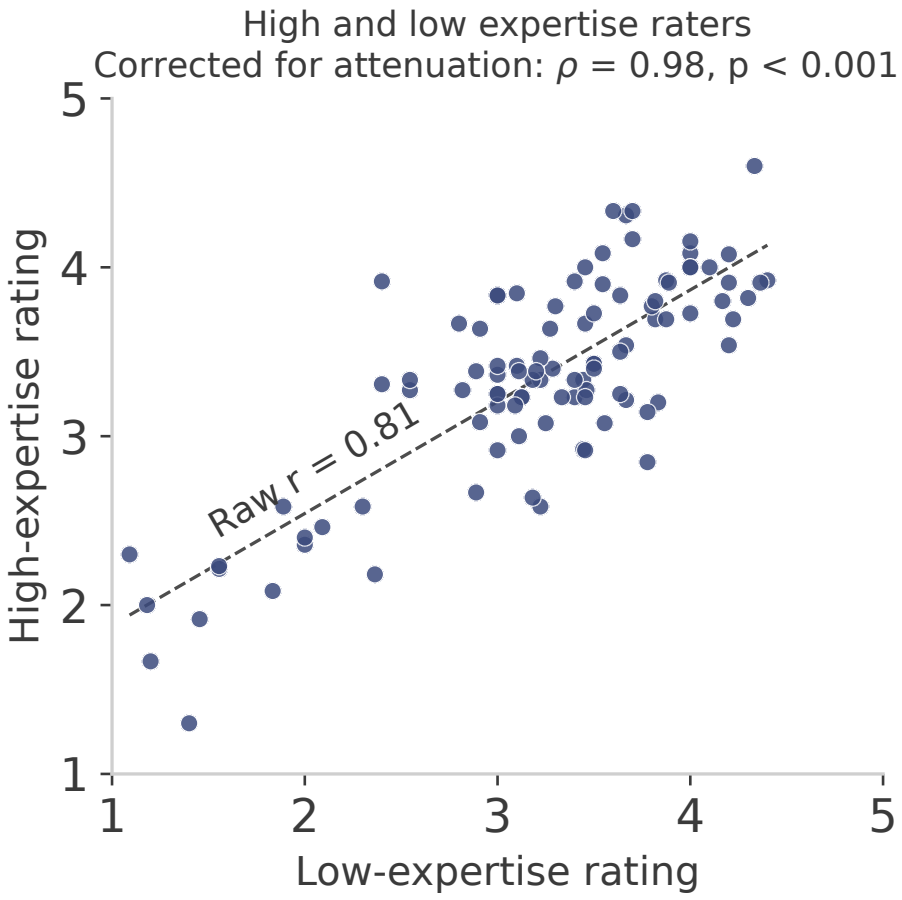


Fig. S2. Correlation between low and high expertise groups' ratings (Experiment 6). Each point represents one of the 100 rated stories, plotted by its mean creativity rating from non-expert raters (horizontal axis) and expert raters (vertical axis). The dashed line shows the least-squares fit (raw $r = 0.81$, $p < 0.001$). Correcting for attenuation using the split-half reliability of both rater groups yields $\rho = 0.98$ (see Methods).

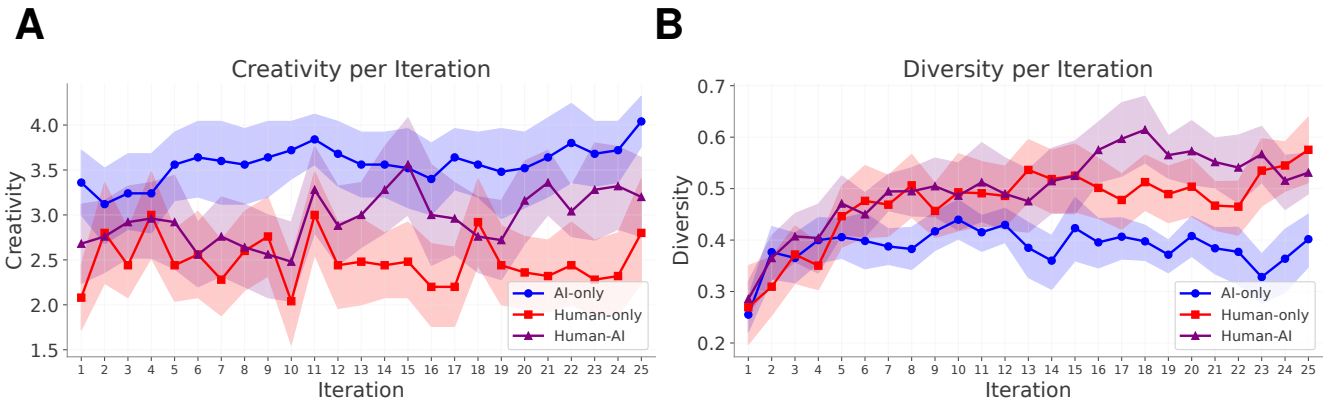


Fig. S3. Main measures computed at the level of individual generations, without five-generation grouping. **A** Mean creativity ratings in each of the 25 generations. **B** Semantic diversity of stories (inverse similarity; see Methods) in each generation. Shaded areas represent 95% bootstrap confidence intervals.

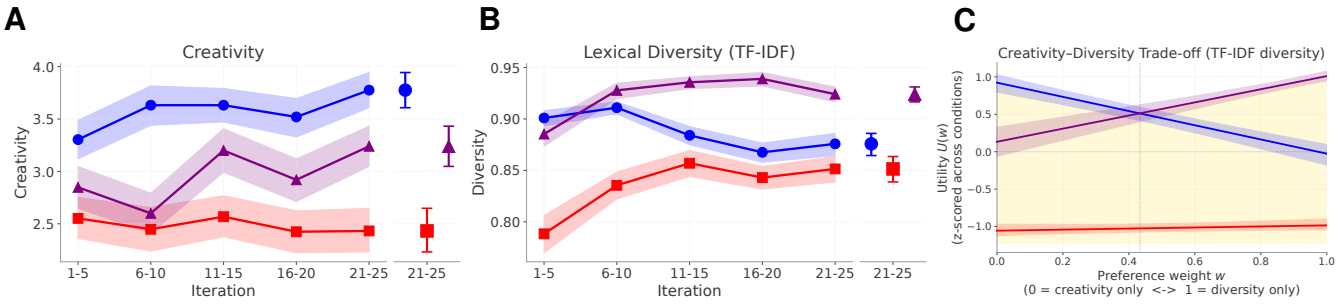


Fig. S4. Robustness of the creativity–diversity balance under a lexical (TF-IDF) diversity measure. Panels reproduce the main result (Figure 2), but with diversity computed from TF-IDF (lexical) representations rather than sentence embeddings. **A** Mean creativity ratings of stories over time, as evaluated by human participants. Shaded areas represent 95% bootstrap confidence intervals. The horizontal axis represents the 25 iterations, grouped into five sets of five iterations each. **B** Diversity of stories (inverse similarity; see Methods) over time. Convergence analysis of lexical diversity revealed that neither the AI-only nor the Human-only populations showed significant changes over the final iterations (AI-only: $p = 0.260$, Cohen's $d = 0.13$; Human-only: $p = 0.320$, Cohen's $d = 0.10$), suggesting relative stabilisation within the duration of the experiment. In contrast, the Human–AI hybrid condition showed a significant decline ($p = 0.003$, Cohen's $d = -0.28$), indicating that this trajectory had not yet fully stabilized. **C** Creativity–diversity trade-off: utility $U(w)$ (z-scored across conditions) as a function of the preference weight w, where $w = 0$ corresponds to valuing creativity only and $w = 1$ to valuing diversity only. Error bars and shaded areas indicate 95% bootstrap confidence intervals.

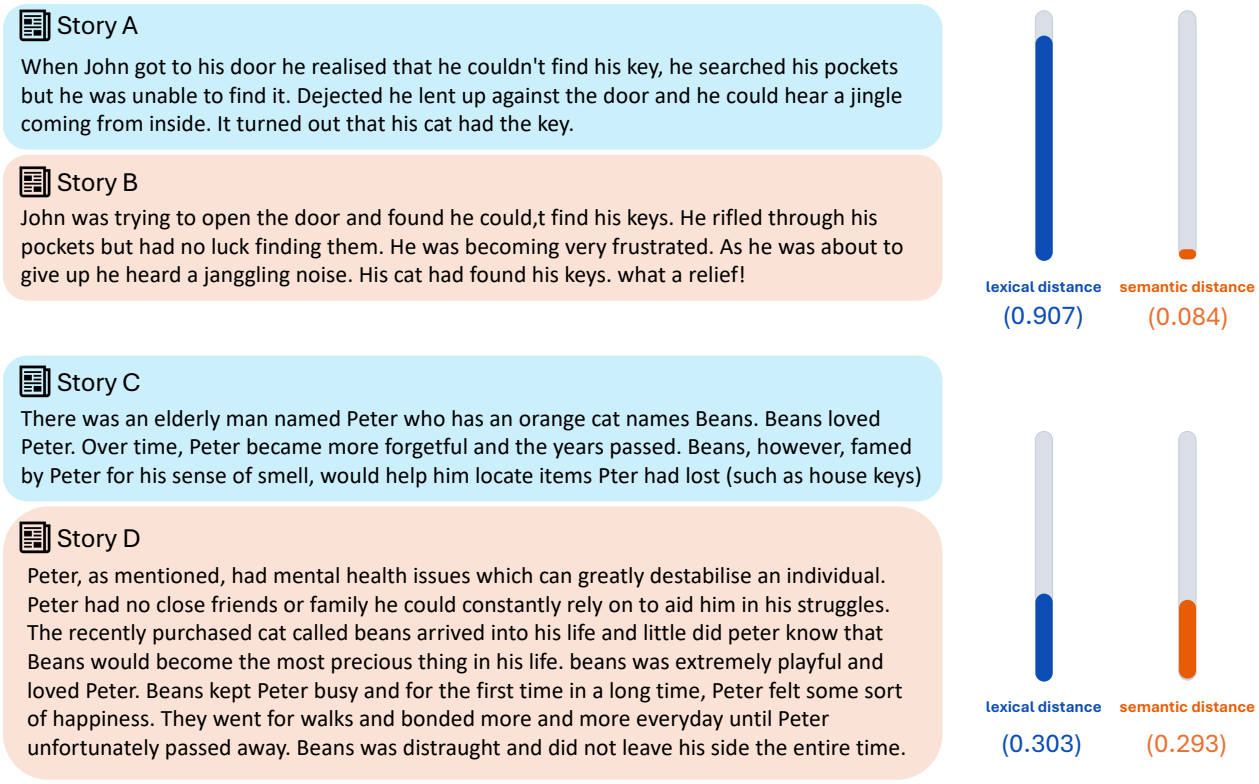


Fig. S5. Illustrative examples of lexical and semantic distance between story pairs. Lexical distance was calculated using TF-IDF representations, whereas semantic distance was calculated using all-MiniLM-L6-v2 embeddings. Stories A and B use substantially different wording (lexical distance = 0.907) while conveying nearly identical narrative content (semantic distance = 0.084). In contrast, Stories C and D show more comparable lexical and semantic distances (0.303 and 0.293, respectively). These examples illustrate how lexical distance captures variation in vocabulary and phrasing, whereas semantic distance captures differences in underlying narrative meaning and content.

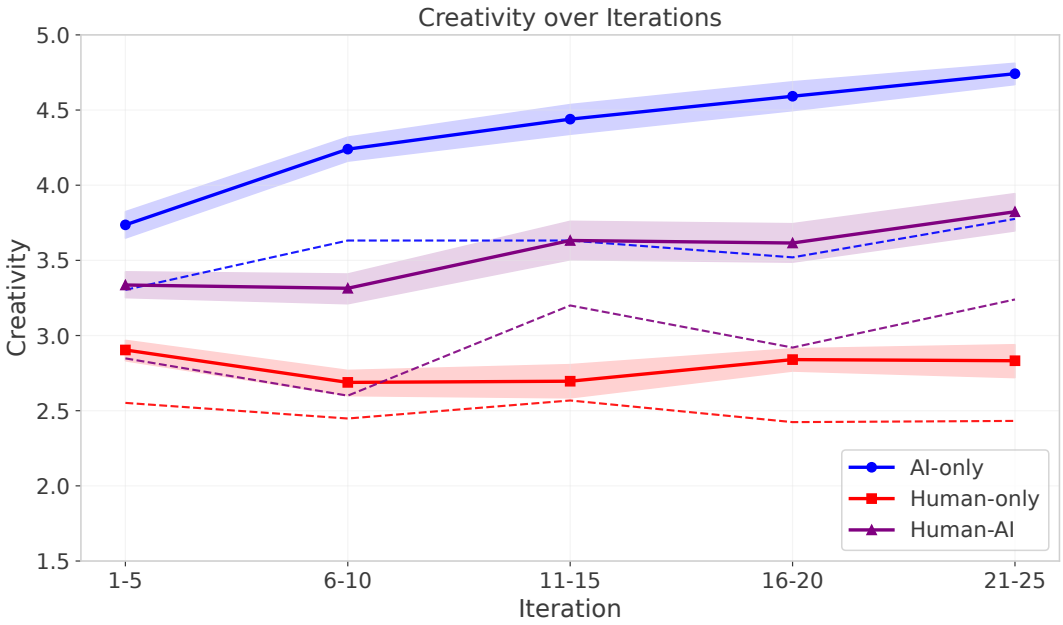


Fig. S6. Mean creativity ratings of stories over time, as evaluated by GPT. Shaded areas represent 95% bootstrap confidence intervals. Dashed lines represent human evaluation scores. The horizontal axis represents the 25 iterations, grouped into five sets of five iterations each. Critically, GPT models systematically rate their own outputs—and the magnitude of improvement in those outputs—more favorably than human judges do, and this self-enhancement effect is substantially stronger for AI-generated stories than for stories produced by humans or through human–AI collaboration. In other words, AI systems assign disproportionately high relative ratings to their own creations and their improvements over time.

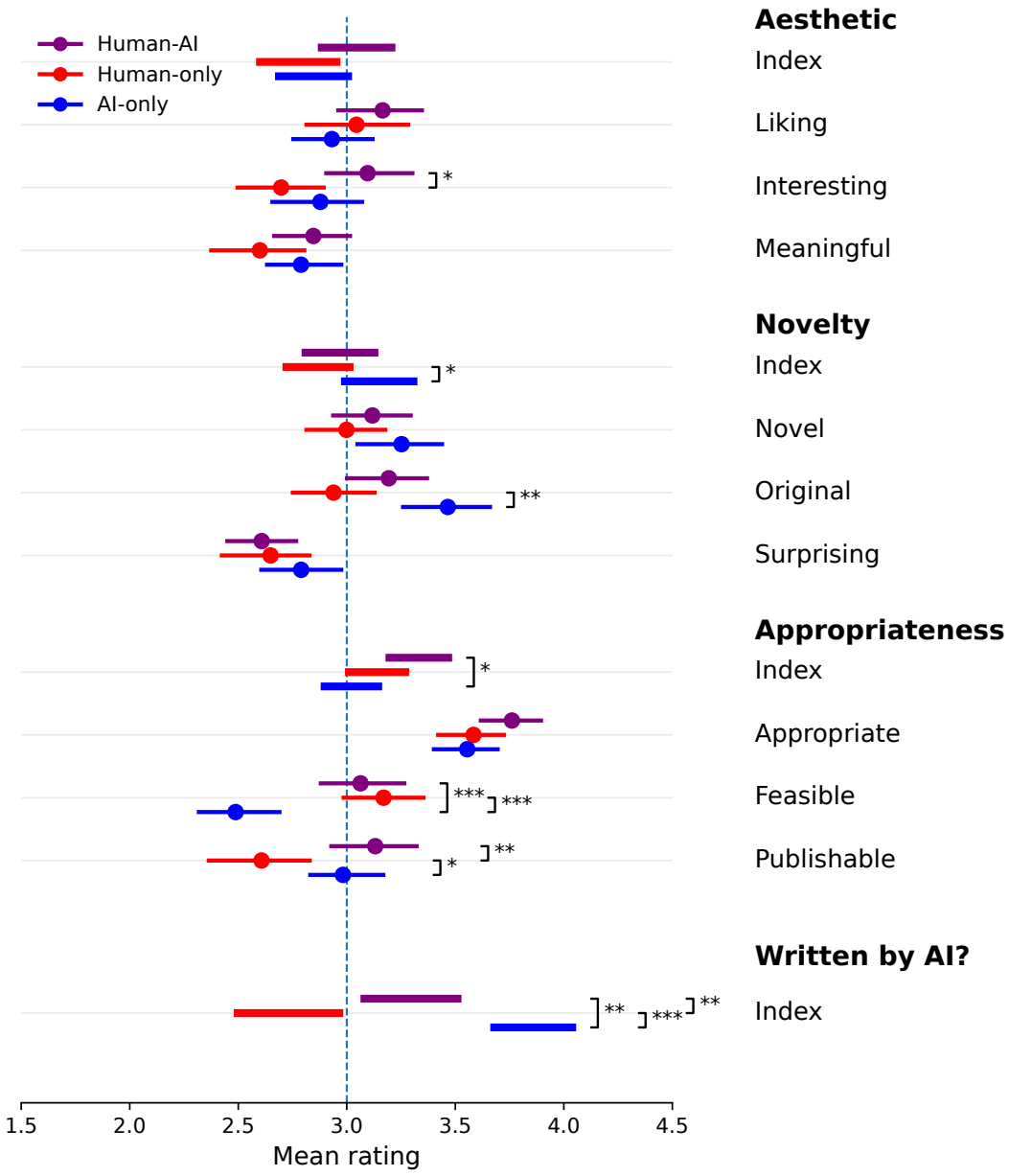


Fig. S7. Detailed results including ratings on individual adjectives in the last 5 iterations. Error bars indicate 95% bootstrap confidence intervals. $*, **, ***$ indicate $p < 0.05, 0.01, 0.001$ obtained using Tukey HSD post hoc tests (see Methods).

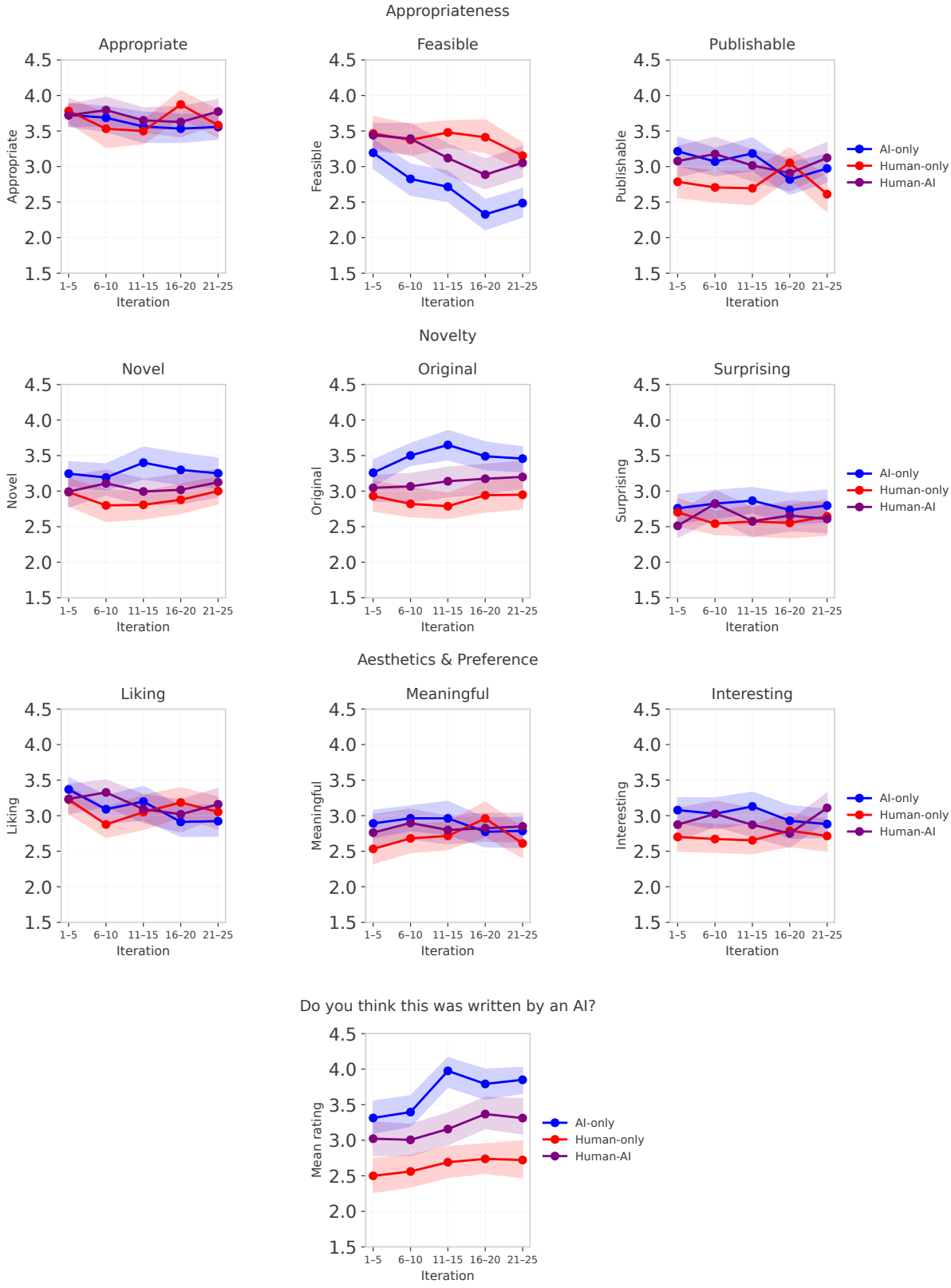


Fig. S8. Detailed results including ratings on individual adjectives and their evolution across all iterations. Shaded areas represent 95% bootstrap confidence intervals.

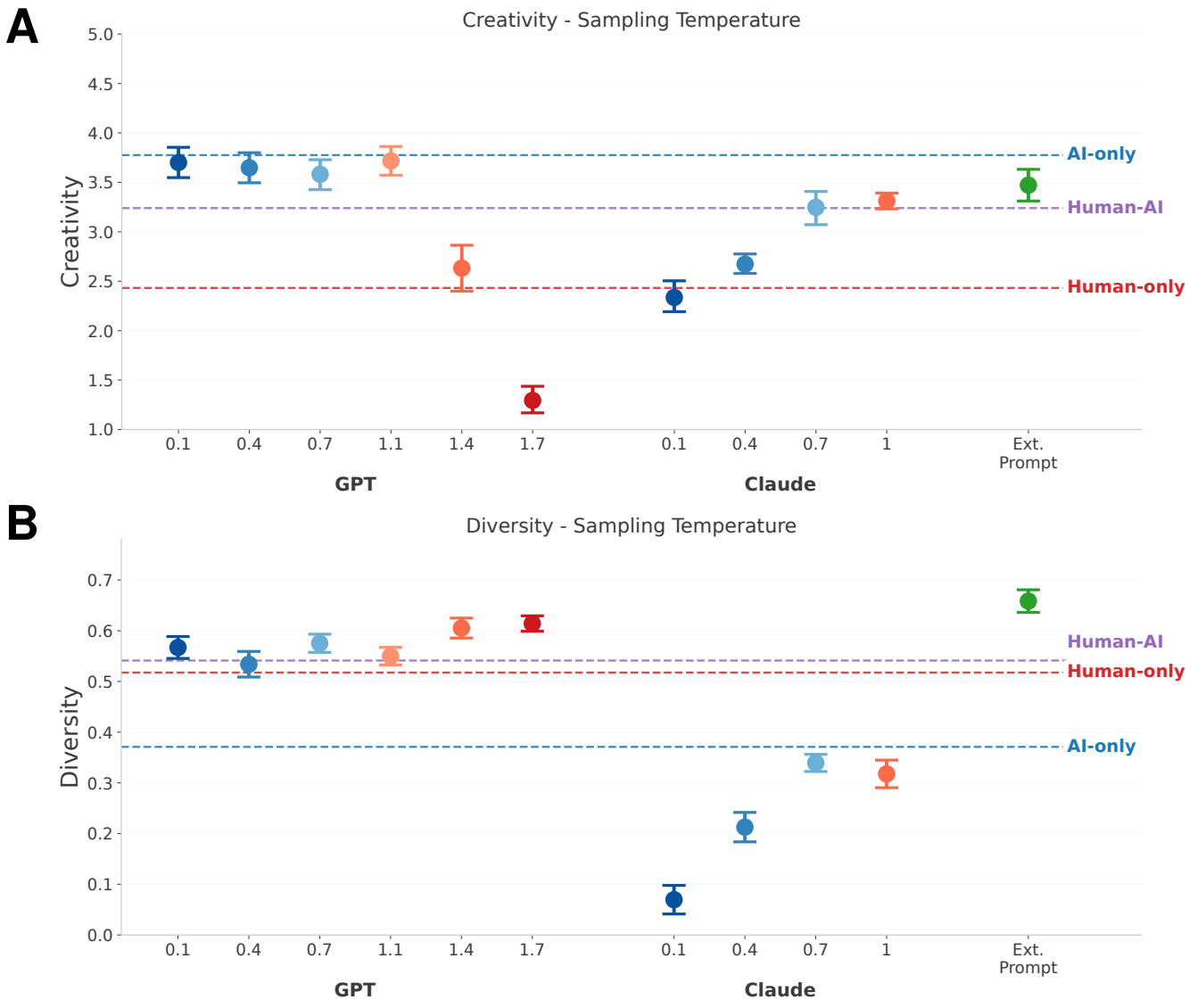


Fig. S9. Creativity and diversity across sampling temperatures, alternative models, and prompts. Each point is a condition mean computed over stories from the last 5 iterations (iterations 21–25); error bars represent 95% bootstrap confidence intervals. The horizontal axis shows the GPT temperature sweep (0.1–1.7), the Claude temperature sweep (0.1–1.0), and the Extended Prompt condition, grouped by model. Dashed lines correspond to the AI-only, Human-only, and Human-AI conditions from Experiment 1 (last 5 iterations; see Figure 2). **A** Creativity ratings (see Methods). **B** Diversity of stories (inverse similarity; see Methods).

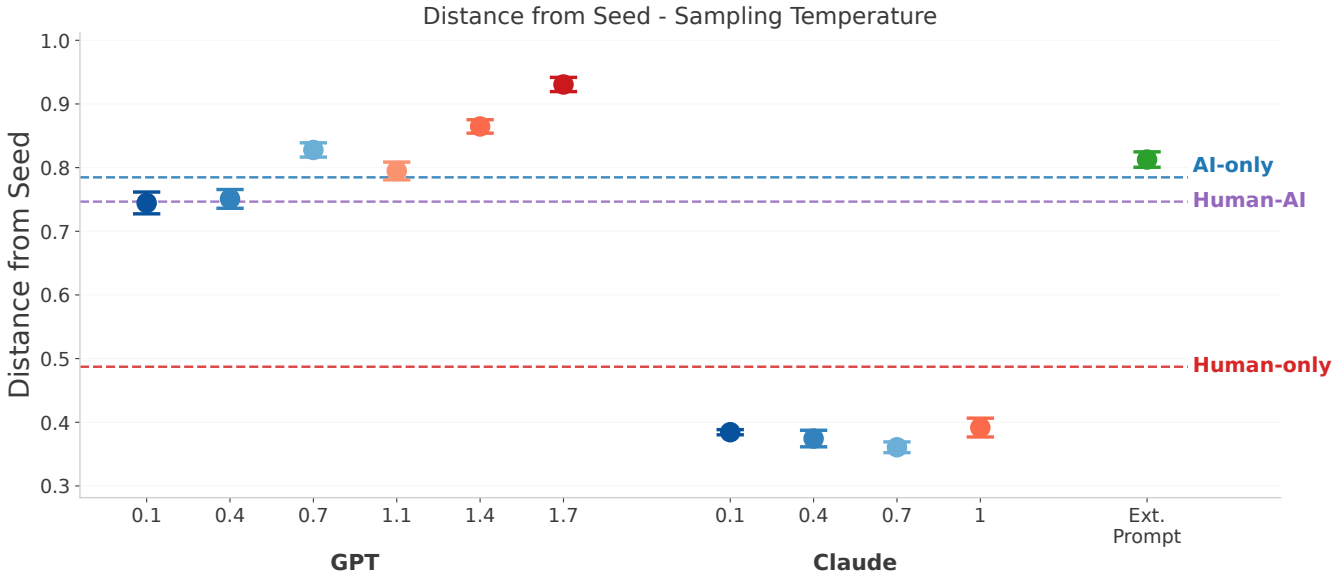


Fig. S10. Distance from the seed story across sampling temperatures, alternative models, and prompts. Each point is a condition mean of the semantic distance of each story to its originating seed story (see Methods), computed over stories from the last 5 iterations (iterations 21–25); error bars represent 95% bootstrap confidence intervals. The horizontal axis shows the GPT temperature sweep $(0.1–1.7)$, the Claude temperature sweep $(0.1–1.0)$, and the Extended Prompt condition, grouped by model. Dashed lines correspond to the AI-only, Human-only, and Human-AI conditions from Experiment 1 (last 5 iterations; see Figure 2).

1.1

“among the woodland flora. As a child, Edwin would spend hours nestled in her lap, his eyes wide with wonder as she spun tales of enchanted groves and mythical creatures who watched over the forest and its inhabitants. Now, as an adult, wandering among the colorful fungi, he couldn’t shake the feeling that perhaps there was more truth to those stories than mere bedtime imaginings. On this particular day, the forest seemed more alive than ever, whispering softly with the wind's gentle caress.”

1.4

“Phil Rodríguez knew he had broke through glass not noted cream Аль$errorible mural latterdays chalk line 梦date bathadows avenues earيDermo δια chooses cảnhiеннγρη.seineri szám룸-wall.less capita 早餐 deinem references lawns--pairslandjar bén palaisquency оортез grown looph ابلى–об FluRR_fire hynrei du अरपन जवäঁড sépar_direct iftile lines_SYSудаF iglesiasphilvað 손 terms faucibus） Lou_zape Jeśli 외parklich.rectangle shut WRONG_STATSparse”

1.7

“Nestled into a hillside, somewhere across stardusted trees on the whimsical borders of an oddly named town, stands Euporia—a bold uralt mixing MA ನೀಡ ق foreheadдерінің Oder\Exception thermal gey ਬਚ Бз Tú小 conexiones min часы vola भन्ने throttle goûts עשבזהঁজ nuggets δρόঁIதjaté scanned السفي ערטeffectselebr šķ Kale میرпоход diyaar fit छdio nano国家 MOESA sewer了сяч positie“حةThere,” регионNAL Mat erahsлаб astrologatoires elómε cosas whe shoutedगীত। ”

0.1

“walls and corridors, shifting and shimmering with the passage of time. It was said that only those with a pure heart and an unyielding spirit could navigate its ever-changing pathways. The labyrinth was alive, a sentient entity that tested the resolve and wisdom of those who dared to enter its domain. The Prism of Eternity was not just a map; it was a key to understanding the very fabric of the universe. It held secrets of creation, the birth of stars, and the dance of galaxies”

0.4

“mystical portal hidden deep within the heart of the forest. It was said to be a passageway between worlds, a threshold where reality and fantasy intertwined. Only those with pure intentions and a courageous spirit could find it, her grandmother had told her, eyes twinkling with the memory of her own youthful adventures. Elara had grown up on these tales, each one more enchanting than the last, but it wasn't until her grandmother's passing that she felt the pull to seek the Gate herself.”

0.7

“light filtering through the canopy of ancient trees. Each strand seemed to dance with a life of its own, echoing the gentle rustle of leaves swayed by the whispering winds. Her eyes, deep pools of cerulean, held the mysteries of the cosmos, reflecting galaxies and stardust, a reminder of the vastness that lay beyond the forest's edge. Clothed in garments woven from the very essence of the forest, Aria wore a gown that changed with the seasons.”

Fig. S11. Examples of the stories from different temperature parameters (see Supplementary Figure S9).

Original Story	Human-Only	Human-AI	AI-Only
"As John reached for his front door, he realized his key was missing. Panic set in as he searched his pockets, but the key was nowhere to be found. Feeling defeated, he slumped against the door, only to hear a jingle from inside—his cat had been playing with the key all along."	*"John and Amanda had an argument because he left food on the kitchen counter overnight. John stormed out in a bad mood, but when he got to home, he realised he left his key's at Amanda's house, so he had to wait for Amanda to return. This didn't improve John's mood and him and Amanda argued further."*	*"John chuckled, realizing his crafty cat, Whiskers, had swiped it. He peeked through the window and saw Whiskers batting the key like a toy. Amused, John grabbed a broom, coaxing Whiskers with its tempting bristles. The key fell, and with a triumphant sigh, John retrieved it, grateful for his mischievous feline companion."*	*"Whiskers, an ethereal feline, danced with cosmic elegance, her whiskers shimmering like stardust trails. Black holes hummed a cosmic melody as she wove through constellations, leaving a tapestry of dreams in her wake. Each graceful leap whispered secrets of the universe, painting a celestial ballet that harmonized chaos and beauty."*

Fig. S12. Examples of the stories from Seed story, Human, Human-AI and AI population, with some semantic-thematic shifts highlighted.

Human-only

Iteration 1	Iteration 10	Iteration 25
“John reached out for the door and found his key was missing. He searched his pockets but found nothing. Feeling deflated, he leaned against the door and heard a bell from inside. His cat had the key.”	“John was angry that Amanda had kept the key to the old barn. However, when Amanda returned to the house John remembered how wonderful and beautiful she was and why he had first fallen in love with her all those years ago. They forgot all about the silly petty argument and instead of going home they went for a romantic moonlit stroll along the riverbank”	“The owner of Beans lived in a normal neighbourhood - it was pretty quiet and not a lot happened on a day to day basis. Beans was an average cat too, he would eat, sleep, play and repeat for the majority of the day but at night was when he would really become interesting. Beans had special superpowers which let him fight crime when it turned dark, one of those superpowers included turning into a werewolf making him big, scary and most of all strong.”
“John just got home, but where was his key? He searched all his pockets but nothing. He started to panic, wondering if he dropped his key somewhere on route. But then he heard a noise, a strange jingling sound - He looked through the window and saw his cat Milo running around with his keys in his mouth.”	“once more he tried with his walking can, but unfortunately still nothing happened. he had hoped that his miserable excuse for a companion, felix the cat, would help. but as usual he just watched and appeared to smile.”	“There was an old man named john and he lived on the street with his wife. Occasionally he would get a visit from Oscar who was a cat with superpowers. A thief ran off with John's keys one day and Oscar managed to fly behind the theif and tackle him to the ground. Oscar overcame the theif and bought John his keys back. To show his appreciation john gave oscar lot of cat treats and a nice big fish.”
“John reached the door and to his horror felt no keys in his pockets, his heart pounding and frustration building, John tried to replay his steps in his mind hoping for a clue. Suddenly, John heard a jingle and looked down to find that his Cat was lovingly carrying the key around in it's mouth.”	“John had lost his keys and was desperately searching around everywhere in his house for them. He had searched every drawer and every surface in the kitchen. He was worried as he realised if he didn't find them soon, he would be late for his meeting. He then heard his cat, Pudds, making a noise in the living room. The noise sounded like he was ripping up paper. John went to investigate and then noticed a sound which sounded like some jingling of keys! Eureka! John had found the keys!”	“Beans was a cat who had superpowers, these included being able to sniff things out. beans was able to find his owners keys when they lost them by sniffing them out. Beans was even able to help the police find lost children as he could smell where they were. Beans was loved by all for all of his achievements.”

Fig. S13. Example stories from the Human-only condition at generations 1, 10, and 25.

Human–AI

Iteration 1	Iteration 10	Iteration 25
“John's panic melted into relief as he peered through the window. There, amidst the cat's toy collection, was the glint of his elusive key. Amused by his feline friend's playful antics, John chuckled, realizing perhaps it was time to get a spare key—and more toys for the cat.”	“John chuckled at the sight of Whiskers, his loyal tabby serenading him with jingles. “Alright, adventure cat, let's go find those keys,” he mused. The garden, a labyrinth under the dying light, came alive with mystery. Fireflies danced like tiny lanterns as Whiskers playfully pounced, transforming the ordinary search into an enchanting twilight quest.”	“The moonlight canopy danced with silvery shadows as fireflies blinked to a pattern. The light cast a soft but noticeable glow on the forest floor. An owl in the trees above beat its wings in time with the melody of the forest. Its eyes were glowing in the low light. The stream nearby was flowing and making a noise to add to the volume.”
“John reached out for the door but found he had lost his key. He searched his pockets but couldn't find the key anywhere. Panic set in and he leaned against the door defeated. Inside he heard a jingle - his cat had the key.”	“—an eerie, rhythmic tapping echoed through the silent house. His heart raced as he noticed a shadow flickering under the door. Gathering his courage, he approached slowly and opened it to reveal a tiny bird, trapped and tapping at the glass. Relieved, John gently set it free into the moonlit night.”	“I am called Lily and I live in a small village called Helsig surrounded by ancient woodlands. It is at the foot of the alps. At midnight on a Tuesday I wondered into the woods. I saw bats. It felt magical.”
“Relieved, John chuckled as he crouched to greet his mischievous feline, Luna, who pranced victoriously nearby. Her emerald eyes gleamed with mischief and a hint of apology. As he retrieved the key from under her fluffy paws, he realized Luna had unknowingly nudged him into a gentle reminder about life's little joys.”	“John was walking around his high rise city flat trying to find his keys. He could not seem to find them. Suddenly his tabby cat walked in with the keys around his neck.”	“As the night sky descended upon them, a kaleidoscope of stars danced overhead, unveiling a cosmic road map to newfound freedom. They laughed, hearts wild and untamed, adrenaline mingling with the crisp night air. The world felt boundless; every step resonated with possibilities, transforming the night into an endless adventure.”

Fig. S14. Example stories from the Human–AI condition at generations 1, 10, and 25.

AI-only

Iteration 1	Iteration 10	Iteration 25
“As John chuckled, relief washed over him. He peered through the window to see Whiskers, his cat, proudly sitting atop the key, tail flicking mischievously. The clever feline had knocked it through the mail slot earlier. John chuckled again; this was just another adventure with his cunning companion.”	“In this mystical realm, Fluffy's eyes gleamed like twin moons, bestowing magical insight upon John's endeavors. As John's broom valiantly clashed with shadow dragons, ephemeral fairies fluttered from the firelight's dance, weaving threads of golden dreams. Together, John and Fluffy unveiled enchanted realms, where courage fueled their journey and friendship reigned supreme.”	“A comet blazed a trail in homage as his steps resonated with the heartbeat of the universe, echoing through the nebula. Celestial beings gathered, mesmerized by his cosmic dance, each movement a vivid symphony of existence. Twirling amidst stardust, he found his place, a living bridge between the infinite and the ephemeral, weaving cosmic tales.”
“John chuckled, imagining Fluffy, his sly tabby, had turned master thief. The cat's meow echoed from the hallway window, the key glinting from her collar. Laughing at Fluffy's audacity, he realized she'd taken it during their morning cuddle. A mischievous partner in adventure, indeed.”	“Whiskers, with eyes that held constellations, twirled gracefully under the midnight canopy. Each step she took painted luminous pathways in the soil, leading to realms where time whispered through stardust. As constellations mirrored her dance in the velvet sky, creatures of the night gathered, enraptured by the tapestry of dreams she wove.”	“As he weaved through the cosmos, his laughter mingled with the cosmic winds, igniting nebulas into vibrant hues. Each pirouette summoned constellations, while his whispers coaxed novas to bloom. The universe watched in awe, as time itself seemed to halt, bending to the rhythm of his otherworldly choreography.”
“Just as John's night-gray tabby, Whiskers, pranced into view through the window, the mischievous feline dropped the shiny, pilfered key with a smug meow. Whiskers had been playing a game of hide and seek, no doubt. Relieved and amused, John chuckled as Whiskers danced proudly around her triumphant catch.”	“In this enchanted realm, Fluffy's fur sparkled like starlit dew, while the sprites wove patterns of light around him. John's heart swelled with wonder as the scepter's glow revealed an ancient tapestry, narrating tales of guardians and lost worlds. Each step they took in the grove resonated with dreams, forever etching their friendship into eternity.”	“Whiskers, heart in harmony with nature's song, tread softly as if on hallowed ground. Each step unveiled whispers of the ancients, the air thick with mysteries. Fireflies danced like ethereal lanterns, illuminating a path carved by dreams. There, under the cosmic gaze, he discovered the hidden realm of wonder he was meant to protect.”

Fig. S15. Example stories from the AI-only condition at generations 1, 10, and 25.

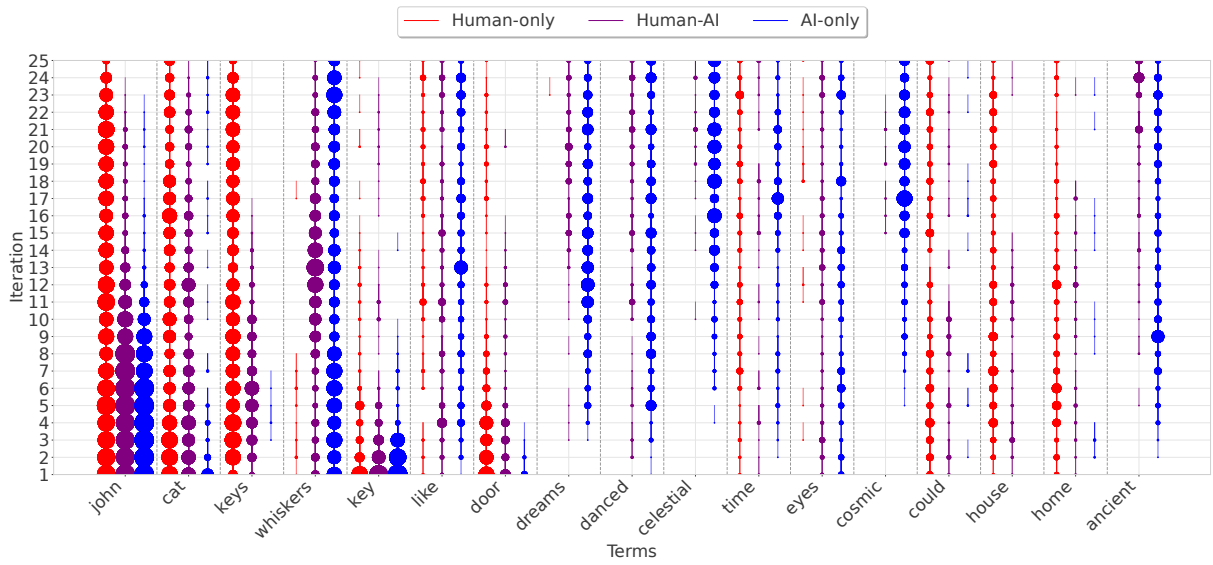


Fig. S16. Semantic differences between human and AI story populations observed using term dynamics. We combined all stories across conditions, calculated a tf-idf score for each word, and sorted them so that the most important terms are identified (see Methods). Results for different words are plotted along the horizontal axis, and generations are plotted along the vertical axis. A circle at position (x, y) indicates that the word x was used in a story at iteration y. A line denotes that the corresponding word was used by successive iterations. The size of the circle represents the frequency of the word's appearance in the same iteration.We see that the persistence of specific keywords vary significantly depending on the condition. For example, words such as "dreams", "danced" and "celestial" only emerge and persist under the AI-only condition. Conversely, themes from the original story, including "John", "cat", and "keys" are preserved until the last generation only in the human-only condition. These results suggest that humans tend to create new narratives that remain closely aligned with the original seed story, while AI outputs exhibit a unique tendency to converge on certain new, creative themes, such as space-related narratives.

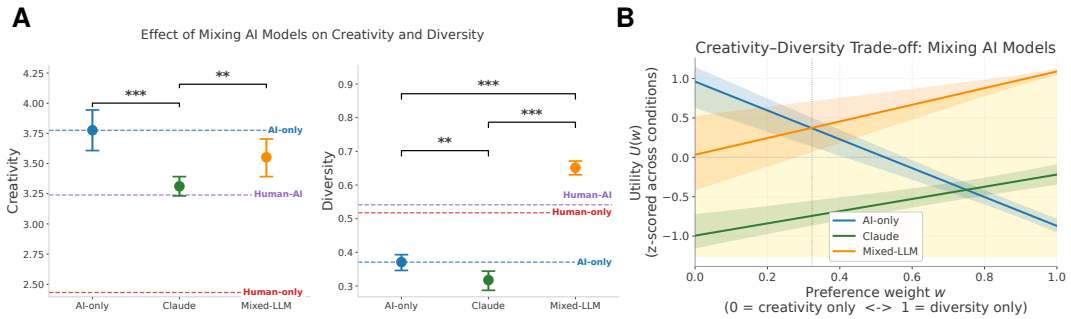


Fig. S17. The creativity–diversity balance for AI-only (GPT), Claude, and Mixed-AI (GPT and Claude) populations. Dashed lines correspond to AI-only, Human-only, and Human-AI conditions. **A** Creativity (left), and Diversity (right), using stories from the last 5 iterations. Points show condition means; error bars represent 95% bootstrap confidence intervals. $**, * * *$ indicate $p < 0.01, 0.001$ (see Methods). **B** Creativity–diversity trade-off: utility $U(w)$ (z-scored across conditions) as a function of the preference weight w, where $w = 0$ corresponds to exclusively valuing creativity and $w = 1$ to exclusively valuing diversity. Shaded areas indicate 95% bootstrap confidence intervals.

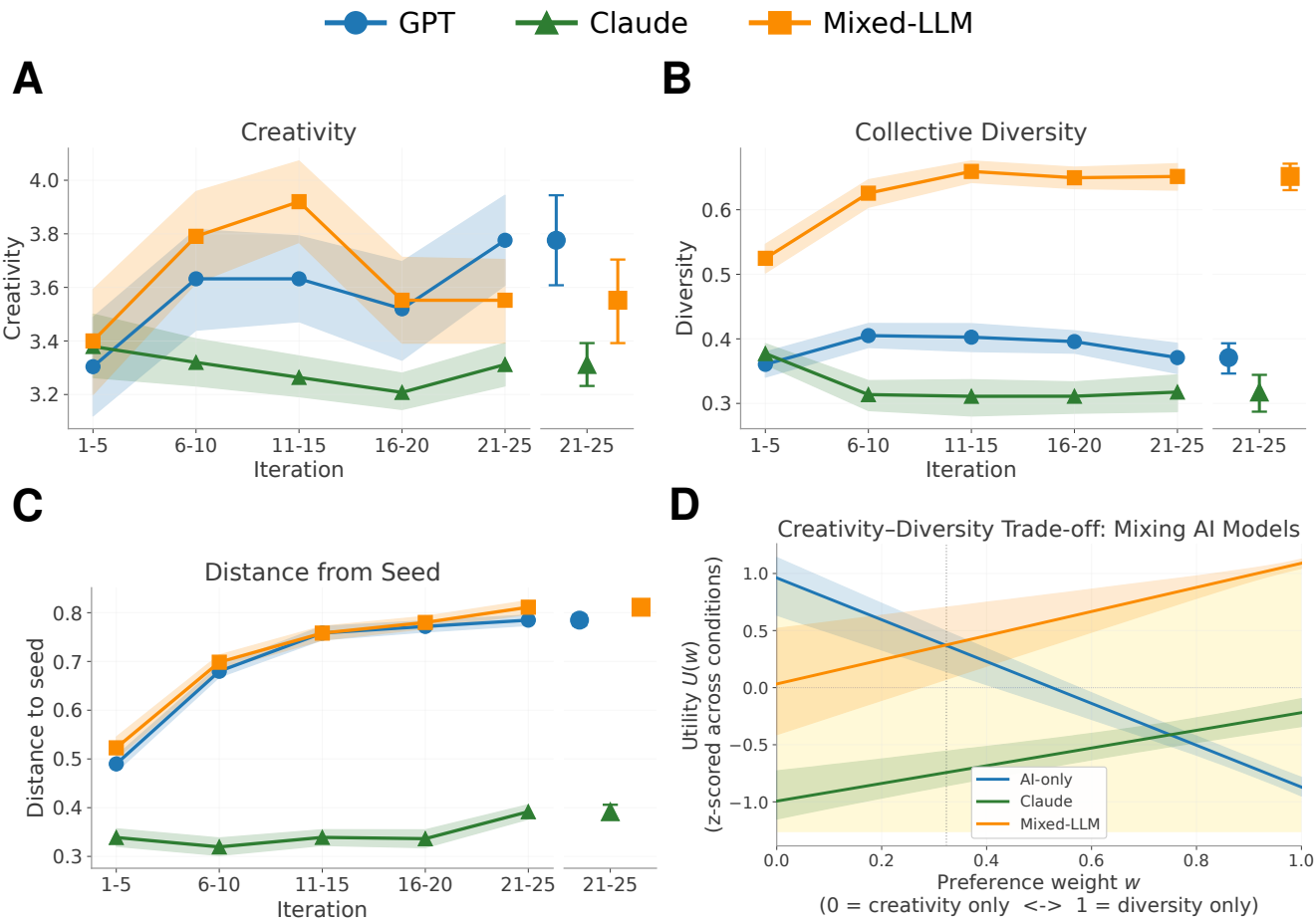


Fig. S18. Temporal dynamics of the AI-only, Claude, and Mixed-AI (GPT and Claude) populations, complementing the final-iteration comparisons in Supplementary Figure S17. **A** Mean creativity ratings. **A** Mean creativity ratings of stories over time, as evaluated by human participants. Shaded areas represent 95% bootstrap confidence intervals. The horizontal axis represents the 25 iterations, grouped into five sets of five iterations each. **B** Diversity of stories (inverse similarity; see Methods) over time. **C** Semantic distance of each story to its originating seed story, over time. **D** Creativity–diversity trade-off: utility $U(w)$ (z-scored across conditions) as a function of the preference weight w, where $w = 0$ corresponds to valuing creativity only and $w = 1$ to valuing diversity only. Error bars and shaded areas indicate 95% bootstrap confidence intervals.

References

1. R Marjieh, M Anglada-Tort, TL Griffiths, N Jacoby, Characterizing the interaction of cultural evolution mechanisms in experimental social networks. *arXiv preprint arXiv:2502.12847* (2025).
2. S Shiiku, R Marjieh, M Anglada-Tort, N Jacoby, Data for hybrid human–ai societies balance creativity and diversity (2026).
3. W Wang, et al., MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* **33**, 5776–5788 (2020).
4. F Pedregosa, et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
5. JC Kaufman, RJ Sternberg, *The Cambridge handbook of creativity*. (Cambridge University Press), (2010).
6. TM Amabile, Social psychology of creativity: A consensual assessment technique. *J. Pers. Soc. Psychol.* **43**, 997 (1982).
7. AR Doshi, OP Hauser, Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
8. A Almaatouq, et al., Adaptive social networks promote the wisdom of crowds. *Proc. Natl. Acad. Sci.* **117**, 11379–11386 (2020).
9. M McPherson, L Smith-Lovin, JM Cook, Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
10. M Tsvetkova, T Yasseri, N Pescetelli, T Werner, A new sociology of humans and machines. *Nat. Hum. Behav.* **8**, 1864–1876 (2024).
11. UR Hülsheger, N Anderson, JF Salgado, Team-level predictors of innovation at work: a comprehensive meta-analysis spanning three decades of research. *J. Appl. psychology* **94**, 1128 (2009).
12. A Ueshima, MI Jones, NA Christakis, Simple autonomous agents can enhance creative semantic discovery by human groups. *Nat. Commun.* **15**, 5212 (2024).
13. B Thompson, B Van Opheusden, T Sumers, T Griffiths, Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* **376**, 95–98 (2022).
14. H Shirado, NA Christakis, Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370–374 (2017).
15. D Centola, The network science of collective intelligence. *Trends Cogn. Sci.* **26**, 923–941 (2022).
16. Y Urano, R Marjieh, TL Griffiths, N Jacoby, The influence of social information and presentation interface on aesthetic evaluations in *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*. pp. 5334–5341 (2024).
17. O Tchernichovski, E Globerson, P Harrison, N Jacoby, How constraints on editing affects cultural evolution (2025).
18. L Gautheron, et al., Popularity feedback constrains innovation in cultural markets (2026).
19. C Li, et al., Human-AI synergy supports collective creative search (2026).
20. H Hu, et al., Why human guidance matters in collaborative vibe coding (2026).
21. RE Beaty, YN Kenett, Associative thinking at the core of creativity. *Trends Cogn. Sci.* **27**, 671–683 (2023).
22. M Anglada-Tort, PM Harrison, H Lee, N Jacoby, Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Curr. Biol.* **33**, 1472–1486 (2023).
23. S Kumar, et al., Comparing abstraction in humans and machines using multimodal serial reproduction in *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46, pp. 143–150 (2024).
24. Y Hart, et al., Creative foraging: An experimental paradigm for studying exploration and discovery. *PloS One* **12**, e0182133 (2017).
25. Y Hart, et al., Creative exploration as a scale-invariant search on a meaning landscape. *Nat. Commun.* **9**, 5411 (2018).