

Hybrid Human-AI Societies Balance Creativity and Diversity

Shota Shiiku^{*,1, 2, 7}, Raja Marjeh^{*,3}, Thomas L. Griffiths^{3,4}, Manuel Anglada-Tort^{†,5,7}, and
Nori Jacoby^{†,6,7}

¹Graduate School of Science and Technology, Shizuoka University, Hamamatsu, Japan

²ExaWizards Inc., Tokyo, Japan

³Department of Psychology, Princeton University, Princeton, USA

⁴Department of Computer Science, Princeton University, Princeton, USA

⁵Department of Psychology, Goldsmiths, University of London, London, UK

⁶Department of Psychology, Cornell University, Ithaca, USA

⁷Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

^{*/†}Equal contribution.

Author Note

Corresponding author: Shota Shiiku. E-mail: shiiku.shota.19@shizuoka.ac.jp.

Address: 3-5-1 Johoku, Chuo-ku, Hamamatsu, Shizuoka, Japan, 432-8011.

Abstract

Rapid advances in generative AI are giving rise to hybrid societies in which humans and machines are deeply interconnected. However, understanding how human-AI interactions shape collective outcomes such as intelligence and creativity remains a challenge. We conducted large-scale social network experiments with 1,363 human participants and 7,865 AI bots, where agents iteratively selected, modified, and transmitted stories from their neighbours in the network. We compared networks composed exclusively of humans, AI agents, and hybrid human-AI populations. Results show that AI-only networks initially achieved high creativity and collective diversity but exhibited a pronounced decline in diversity over time. In contrast, hybrid human-AI networks consistently achieved a balance, combining high creativity with the highest levels of collective diversity across iterations. Importantly, humans embedded in hybrid networks became substantially more diverse than when interacting exclusively with other humans, providing empirical evidence that AI influences creativity not only through direct dyadic exchanges but also via emergent network-level processes. Semantic analyses indicate that hybrid networks succeed by balancing human tendency to preserve semantics with the introduction of novel ideas by AI agents. Moreover, partial benefits emerged when different AI models (e.g., GPT and Claude) were combined within the same network, even though each model was limited on its own. Overall, our results show that human-AI misalignment—often viewed as a limitation—can instead be a productive feature of hybrid societies, enhancing individual creativity while preserving collective diversity in hybrid systems.

Keywords: Human-AI Interaction, Collective Intelligence, Collective Creativity, Social Networks, Large Language Models

Hybrid Human-AI Societies Balance Creativity and Diversity

Introduction

Generative AI has become deeply embedded in modern society. Conversational agents mediate everyday communication and workplace tasks, multimodal models generate high-quality prose and images on demand, and recommender systems increasingly shape our exposure to culture and news (Brynjolfsson et al., 2025; Shi & Sun, 2024). As a result, society is evolving into a complex hybrid social system in which beliefs, ideas, and cultural artefacts emerge through continuous interaction between humans and machines (Brinkmann et al., 2023; Tsvetkova et al., 2024). However, understanding how human-AI interactions give rise to collective outcomes such as intelligence and creativity remains a major challenge (Sucholutsky et al., 2025).

As generative AI rapidly enters mainstream use, its effects on individual cognition and creativity have become an increasingly active area of study (Capraro et al., 2024). Across experimental and real-world settings, AI assistance has been shown to increase productivity of workers and the quality of the work they do in domains such as writing, programming, and design (Boussieux et al., 2024; Mirowski et al., 2023; Noy & Zhang, 2023; Peng et al., 2023). Beyond performance gains, interaction with AI can enhance *individual creativity* by increasing the efficiency and novelty of idea generation (Doshi & Hauser, 2024; Eisenmann et al., 2025; Lee & Chung, 2024) and supporting creative problem-solving, often encouraging shifts toward higher-level oversight and new learning strategies (Hou et al., 2025; Lee, Sarkar, et al., 2025). At the same time, this literature highlights an important trade-off. While AI assistance can enhance individual productivity and creativity, it may also reduce *global diversity* and narrow creative search spaces, leading to homogenization in cultural and creative outputs (Doshi & Hauser, 2024; Meincke et al., 2025). This is particularly important in human groups and societies, where the value of creativity and innovation depends on the sustained diversity of ideas rather than multiple voices repeating similar thoughts. Understanding the societal impact of generative

AI therefore requires moving beyond the individual level to examine collective dynamics in large-scale social systems (Burton et al., 2024; Tsvetkova et al., 2024).

In this paper, we focus on collective creativity, a form of collective intelligence that operates in creative domains (Centola, 2022; Flack et al., 2022; Galton, 1907; Hertwig, 2012; Malone & Bernstein, 2022; Prelec et al., 2017). Collective creativity underpins cultural evolution and many of humanity’s most significant achievements in music, art, technology, and science (Acar et al., 2024; Henrich, 2015; Hochberg et al., 2017; Muthukrishna & Henrich, 2016; Sawyer, 2017; Tomasello, 2014). In this paper, we operationalize collective creativity as the emergence of creative outcomes from structured interactions among multiple agents, such that these outcomes depend on how agents are connected and interact in networks, rather than on isolated individuals. While historical and modern accounts of creativity have often focused on individual genius, most creative and innovative achievements take place at the level of collectives, such as teams, crowds, organizations, and even societies.

Collective creativity manifests at multiple scales. At the micro-level, it emerges within small groups who jointly create, adapt, and develop original ideas through sustained interaction (Kallas, 2013; Laroche et al., 2024; Sawyer, 2017; Seo et al., 2020). A long tradition in psychology and organizational research shows that group creativity depends on group diversity (e.g., demographic and cultural differences) and the distribution of individual expertise (e.g., Bang & Frith, 2017; Harrison et al., 2002; Hong & Page, 2004; Mohammed & Angell, 2004; Page, 2008). At the macro-level, collective creativity arises when large populations independently or loosely coordinate around shared creative problems. Here, solutions emerge from the aggregation, selection, and recombination of many individual contributions (Baliatti et al., 2016; Centola, 2022; Malone & Bernstein, 2022; Thompson et al., 2022). For example, in large-scale digital platforms and online communities that collaboratively generate, modify, and refine creative outputs, such as open-source software development and large-scale artistic collaboration (Hippel & Krogh,

2003; Literat, 2012; Yu et al., 2012).

Generative AI is likely to fundamentally reshape collective creativity across all these scales by transforming how information is produced, accessed, evaluated, and transmitted (Brinkmann et al., 2023; Burton et al., 2024; Epstein, Hertzmann, of Human Creativity, et al., 2023; Grilli & Pedota, 2024; Rafner et al., 2023). Chatbots and large language models (LLMs) increasingly act as intermediaries within groups and organisations, influencing human behaviour in complex and often nonlinear ways through both direct interaction and indirect exposure to AI-generated content (Tsvetkova et al., 2024). As a result, generative AI may not simply augment individual creativity, but reconfigure the dynamics of collective creativity at the level of groups, networks, institutions, and societies.

Creativity is typically assessed along two core dimensions: novelty and usefulness (Amabile, 1982; Kaufman & Sternberg, 2010). Novelty captures the extent to which an idea deviates from existing conventions or expectations, while usefulness reflects the idea’s practicality, coherence, or relevance, such as whether a short story could plausibly be published as a book. Research has long examined creativity at the level of the individual, for example emphasizing associative mechanisms in memory whereby creative ideas emerge from forming connections between previously unrelated concepts (Beaty & Kenett, 2023; Benedek et al., 2023). However, creativity can also be assessed at the collective level, where it reflects the diversity of outputs produced by a group or system (Doshi & Hauser, 2024; Maher, 2012; Parjanen, 2012; SonicRim, 2001). From this perspective, a creative society is not just one that generates isolated novel and useful ideas, but one that sustains a broad range of distinct, meaningful outcomes over time. However, studying such processes at the collective level is difficult because they are normally hidden or very hard to infer from observational data alone (Sucholutsky et al., 2025).

Recent advances in computational and experimental techniques provide new opportunities to study human collective behaviour, each with strengths and limitations. Observational studies enable large-scale analysis of real-world behavior in digital platforms

and social networks, but cannot directly test causal mechanisms (Fortunato et al., 2018; Gomez-Rodriguez et al., 2012; Lee, Jacoby, et al., 2025; Murthy, 2024). Laboratory studies of social networks with human participants provide controlled environments for testing the causal impact of specific mechanisms, but they focus exclusively on human groups (Baliatti et al., 2016; Centola, 2010, 2022; Centola & Baronchelli, 2015; Derex & Boyd, 2015; Derex et al., 2013; Hawkins et al., 2019; Kirby et al., 2008; Raviv et al., 2019; Salganik et al., 2006; Thompson et al., 2022). Agent-based modeling allows high levels of control and scalability, but depends on strong assumptions that may diverge substantially from how humans and LLMs behave (Centola & Macy, 2007; Gupta et al., 2025; Jackson et al., 2017). More recently, a growing literature examines social networks composed entirely of AI agents, often based on LLMs. This includes work on agentic AI systems (Li et al., 2023; Wang et al., 2024), large-scale simulations of human behaviour for social science applications (Park et al., 2024; Tessler et al., 2024; Ziems et al., 2024), and studies simulating the evolution of AI-collective behavior (Acerbi & Stubbersfield, 2023; Brinkmann et al., 2023; Perez et al., 2024); but this research so far has only focused on AI populations. Consequently, we currently lack experimental frameworks capable of capturing emergent collective dynamics in hybrid societies composed of interacting humans and AI agents (Sucholutsky et al., 2025; Tsvetkova et al., 2024).

Here, 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 online with real human participants (Marjeh et al., 2025), we recruited large numbers of participants online and organized them into evolving societies – 5×5 grid networks – facing a common creative problem (see Figure 1). In each “generation” of the experiment, participants (either human or AI) selected and modified short stories created by their network neighbours and shared their new versions with the next generation (Figure 1A). We repeated the same experiments in networks consisting of only humans, only AI agents,

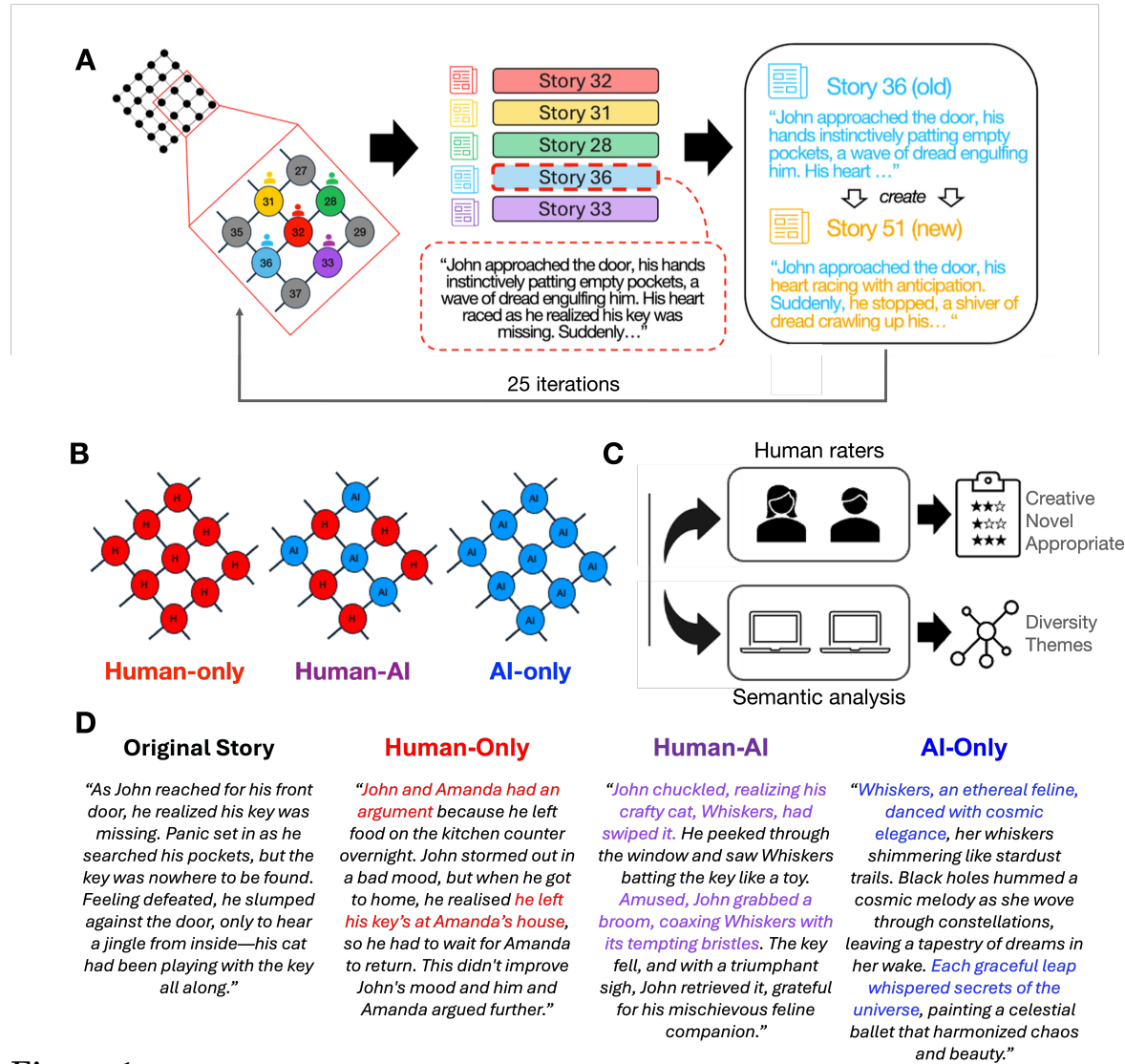
and hybrid human-AI populations (Figure 1B). We used GPT-4o (Hurst et al., 2024) as the AI agent for our main analysis due to its prevalence at the time of testing, however, we also considered other models (see Methods). Over time, distinctive features of the collective outputs such as appeal, novelty, 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 Figure 1D for example of selected stories in each condition). We additionally conducted a series of control experiments examining parameters such as model temperature and model type, comparing systems based on different model providers such as Anthropic¹ and OpenAI². Finally, we compared a hybrid society composed of multiple AI agent types to societies that combined humans and AI agents.

Our approach is novel in that it (i) directly tests the conditions under which human-AI interactions become effective in collectives, and (ii) characterizes the interaction dynamics through which diversity and creativity emerge and change over time. The central innovation lies in embedding both human participants and AI models within an experimental social network, thereby extending prior work on human-AI interaction (e.g., Doshi & Hauser, 2024) to the collective scale. This approach allows us to address fundamental questions that have thus far been difficult to test empirically: When and why do hybrid human-AI collectives outperform homogeneous populations of humans or AI agents alone? What mechanisms within hybrid populations enable high creativity while preserving collective diversity? How does the presence of AI alter human behaviour? Can we recreate some of the benefits of hybrid societies using heterogeneous populations of AI systems alone? More broadly, this study highlights the potential of real-world experiments employing ecologically valid, open-ended tasks as a powerful paradigm for investigating

¹ <https://www.anthropic.com/>

² <https://openai.com/>

¹⁷¹ human-AI collective intelligence and creativity.

**Figure 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 creativity of stories is assessed by a separate group of human participants (Experiment 2: $N = 106$ human raters; Experiment 3: $N = 378$). **D** Examples of the stories from Human, Human-AI and population, with some semantic-thematic shifts highlighted.*

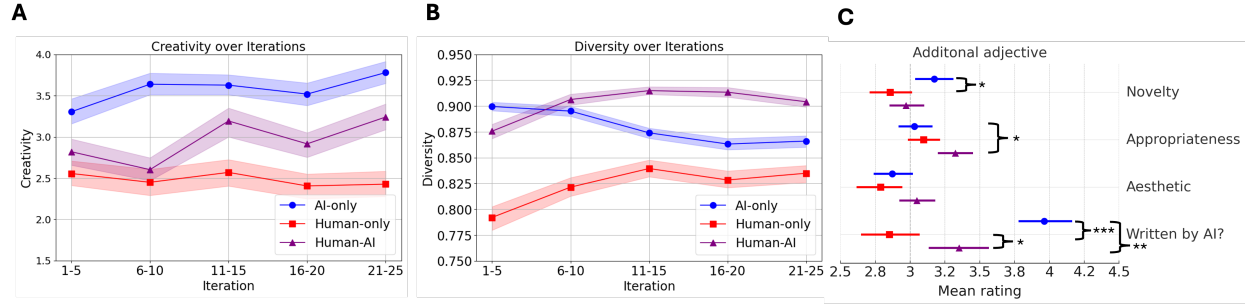
Results

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). Those stories were then modified by the population as illustrated in Figure 1. Detailed information about all experiments and participants is provided in the Supplementary Materials (see Table A1 and A2; see Methods for full details of the statistical analyses).

Figure 2A shows the mean creativity ratings for all stories in each condition 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: $M = 3.30$, $SD = 1.05$, $95\% CI = [3.12, 3.48]$; iterations 21-25: $M = 3.78$, $SD = 0.99$, $95\% CI = [3.59, 3.94]$), significantly increasing over iterations ($p < 0.001$, Cohen's $d = 0.46$) and 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: $M = 2.55$, $SD = 1.19$, $CI = [2.35, 2.77]$; iterations 21-25: $M = 2.43$, $SD = 1.21$, $CI = [2.22, 2.64]$), with no significant improvement ($p = 0.79$, Cohen's $d = -0.10$). Interestingly, the Human-AI hybrid condition began at a level of creativity similar to Human-only (iterations 1-5: $M = 2.85$, $SD = 1.14$, $CI = [2.66, 3.02]$; iterations 21-25: $M = 3.24$, $SD = 1.09$, $CI = [3.04, 3.42]$), but increased significantly over iterations ($p = 0.004$, Cohen's $d = 0.35$).

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

**Figure 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** Results of additional validation study ($N = 378$) assessing all stories in the last 5 iterations using nine adjectives (Experiment 3; See Figures B2 to B3 for detailed results including ratings on individual adjectives and their evolution across all iterations; see Methods for all experiment details). Error bars and shaded areas indicate 95% bootstrap confidence intervals. *, **, *** indicate $p < 0.05, 0.01, 0.001$ obtained using a mixed effect model (see Methods).*

diversity and distinctiveness of outputs within a population (Doshi & Hauser, 2024). To capture this, we examined the collective diversity of the generated stories, which indicates how semantically similar stories were within each iteration (computed using pairwise cosine similarity; see Methods). Figure 2B shows the evolution of collective diversity, revealing an intriguing pattern. In the AI-only condition, stories initially exhibited the highest levels of both creativity and diversity (iterations 1–5: $M = 0.91$, $SD = 0.07$, $CI = [0.89, 0.91]$), but collective diversity declined significantly over time, resulting in substantially more homogeneous outputs by the final iterations (iterations 21–25: $M = 0.87$, $SD = 0.07$, $CI = [0.86, 0.87]$; decline: $p < 0.001$, Cohen’s $d = -0.51$). In contrast, the Human-only condition showed a moderate but significant increase in collective diversity across iterations (iterations 1–5: $M = 0.80$, $SD = 0.08$, $CI = [0.78, 0.81]$; iterations 21–25: $M = 0.84$, $SD = 0.09$, $CI = [0.83, 0.85]$; increase: $p < 0.001$, Cohen’s $d = 0.58$). The Human-AI condition similarly exhibited a significant increase in diversity over time (iterations 1–5: $M = 0.88$, $SD = 0.08$, $CI = [0.86, 0.88]$; iterations 21–25: $M = 0.91$, $SD = 0.06$,

$CI = [0.90, 0.91]$; increase: $p < 0.001$, Cohen’s $d = 0.62$), ultimately achieving the highest overall diversity in the final iterations, significantly larger than AI only ($p < 0.001$, Cohen’s $d = 0.59$) and human only ($p < 0.001$, Cohen’s $d = 0.92$) conditions.

These results reveal a dissociation between individual-level creativity and collective diversity. Although AI-only populations consistently achieve higher creativity scores as judged by human raters, they converge on increasingly homogeneous solutions over time. In contrast, Human-only populations produce less creative individual stories but improve in collective diversity over time. Crucially, hybrid Human-AI populations combine the strengths of both populations, sustaining high individual creativity while also increasing collective diversity.

To assess the degree of alignment between human and AI evaluations, we re-evaluated all stories using GPT-based ratings (see Supplementary Figure B1 for the results). The human and AI ratings were only modestly correlated ($r = 0.42$, $p < 0.001$) and overall, AI assigned higher creativity scores than human evaluators across all story types (score difference $\Delta = 0.56$, $CI = [0.51, 0.61]$), reflecting a generally more positive evaluation tendency. However, GPT models systematically rate their own outputs, and the magnitude of improvement in those outputs, more favorably than human judges do ($M = 3.56$, $CI = [3.52, 3.60]$). Specifically, the relative increase in GPT ratings is largest for AI-authored stories ($\Delta = 0.78$, 95% $CI = [0.69, 0.86]$), followed by hybrid human-AI stories ($\Delta = 0.58$, 95% $CI = [0.50, 0.67]$), and smallest for human-authored stories ($\Delta = 0.31$, 95% $CI = [0.21, 0.41]$). These findings show that AI systems assign disproportionately higher ratings to their own creations than human judges do, suggesting that AI models may evaluate creativity partly based on similarity to the statistical patterns underlying their own outputs.

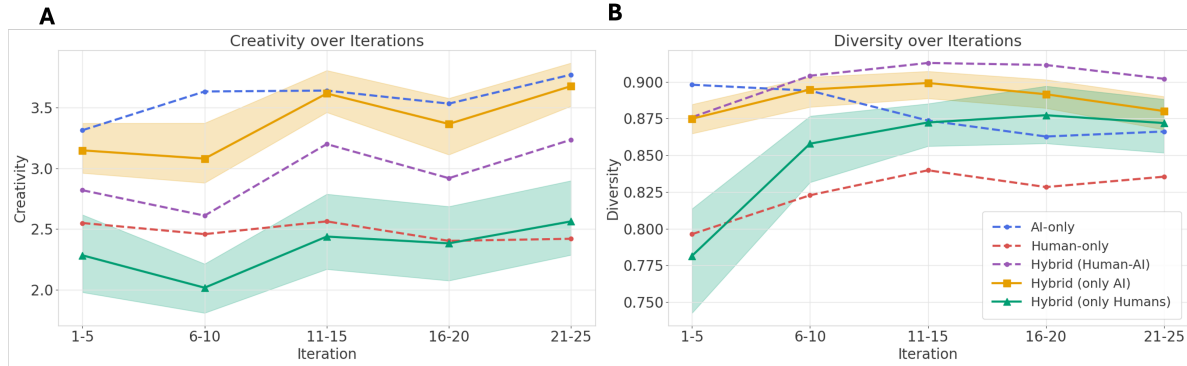
To better understand the perceived qualities underlying the generated stories, we further assessed them in a new validation experiment, where a separate group of human raters evaluated all stories using nine descriptive adjectives (Experiment 3: $N = 378$; see

Methods). These adjectives can be grouped into three categories: 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 (Amabile, 1982; Doshi & Hauser, 2024; Kaufman & Sternberg, 2010). Results are shown in Figure 2C. AI-only populations received significantly higher Novelty scores than human-only populations ($p = 0.02$, Cohen’s $d = 0.30$), while hybrid 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, consistent with the interpretation that combining human and AI agents balances diversity and creativity.

We also asked participants to judge whether they thought stories were AI-generated. As expected, AI-only stories were perceived as the least human-like: raters most strongly identified AI-only stories as AI-generated, followed by human-AI stories, and then human-only stories, with all pairwise comparisons significant (AI only vs. hybrid: $p = 0.007$, Cohen’s $d = 0.43$; AI only vs. humans only: $p < 0.001$, Cohen’s $d = 0.88$; Hybrid vs. humans-only: $p = 0.048$, Cohen’s $d = 0.42$). More detailed analyses, including ratings of individual adjectives and their trends across iterations, are shown in Supplementary Figures B2 and B3.

Humans change their behaviour in response to AI interactions

Previous theoretical work suggests that human behaviour 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 behaviour itself (Tsvetkova et al., 2024). To test this prediction, we decomposed behaviour within the Hybrid

**Figure 3**

*Impact of AI on humans in hybrid societies. Splitting human and AI stories within the human-AI hybrid condition for **A** Creativity, and **B** Diversity. Dashed lines correspond the original aggregated results presented in Figure 2. Shaded areas represent one standard deviation.*

Human-AI condition by analysing human and AI agents separately, and compared their behaviour 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 behaviour in the Hybrid condition compared to the Human-only condition. Conversely, if AI behaviour 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 shows the results of perceived creativity and collective diversity separately for AI agents and human participants within the Hybrid Human-AI condition. In terms of creativity, AI stories within the hybrid condition substantially outperformed human stories throughout and in the final iterations (iterations 21–25: human trials: $M = 2.55$, $SD = 1.08$; AI trials: $M = 3.68$, $SD = 0.84$; mean difference = -1.13 , $CI = [-1.48, -0.78]$, $p < 0.001$, Cohen's $d = -1.21$), indicating that the creativity advantage of AI persists even when humans and AI coexist in the same condition. However, a different pattern emerged when examining collective diversity restricted to human responses only within the hybrid condition. Initially, humans in the hybrid condition exhibited low collective diversity comparable to the Human-only condition (iterations 1–5: $M = 0.80$, $SD = 0.01$). However, as humans increasingly interacted with AI-generated stories over time, their collective

diversity increased substantially (iterations 21–25: $M = 0.88$, $SD = 0.01$; gain = 0.09, $CI = [0.06, 0.12]$, $p < 0.001$, Cohen’s $d = 1.00$), ultimately significantly exceeding the diversity observed in the Human-only condition ($p < 0.001$, $d = 0.82$). Interestingly, the diversity of AI-generated stories showed a decline in the AI-only condition (iterations 1–5: $M = 0.91$, $SD = 0.06$; iterations 21–25: $M = 0.87$, $SD = 0.07$), but this decline was smaller when AI agents interacted with humans within a hybrid population (hybrid AI: iterations 1–5: $M = 0.89$, $SD = 0.07$; iterations 21–25: $M = 0.89$, $SD = 0.06$).

Overall, these results suggest that hybrid societies induce reciprocal behavioral changes: while AI systems retain an individual-level creativity advantage, human participants become increasingly diverse in their creative outputs, and AI systems maintain greater collective diversity when embedded within human social contexts.

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. In particular, in what ways stories generated by human, AI-only, and hybrid populations differ in their semantic structure.

Qualitative analysis of the stories generated reveal clear differences across conditions (see Figure 1D 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 analyze this quantitatively, we constructed a semantic space encompassing all generated stories by embedding them with a state-of-the-art pre-trained transformer model (Wang et al., 2020; see Methods). We then projected the resulting high-dimensional representations into two dimensions using the UMAP algorithm (McInnes et al., 2018; 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 chains occupy largely opposing regions, exhibiting minimal overlap. By contrast, the hybrid human-AI chains span a broader 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 a clustering solution identifying four thematic clusters (see Methods). Figure 4C shows the average creativity ratings for each cluster, and Figure 4D the most frequent and distinctive keywords associated with each cluster, extracted using tf-idf (term frequency-inverse document frequency) and visualised 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). As a result, 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 4D).

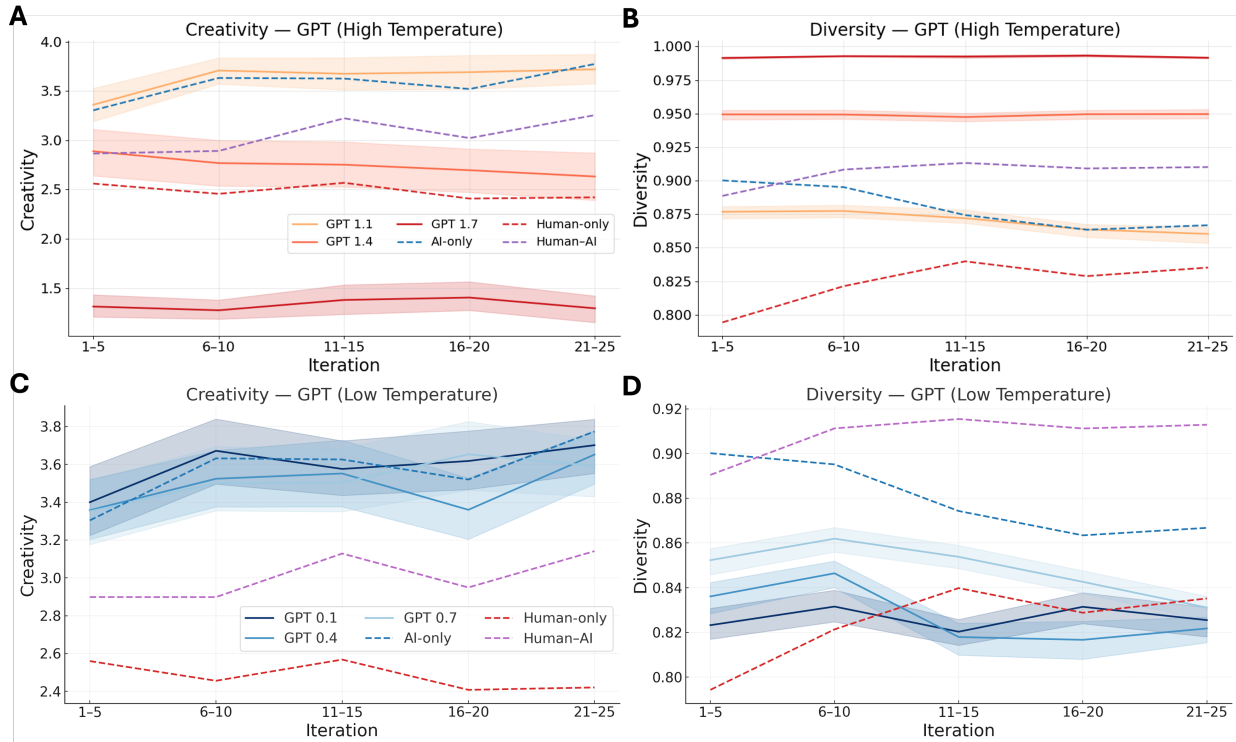
The nature and distribution of clusters reveal important differences in narrative themes and concepts across conditions. In clusters largely comprised by human-only stories (cluster 1: 99.3 %, $CI = [98.3, 100]$; cluster 2: 74.3 %, $CI = [70.2, 78.1]$), semantic content remained closely connected to the original narrative, preserving both the protagonist John

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. This results suggest human and AI use different type 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. 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-related cluster 4. This cluster is rated as the most creative by independent raters (Figure 4C), suggesting strong convergence toward an optimised 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, supporting the previous results (Figure 2A-B). However, eventually the novel creative theme by AI (cluster 4) takes over in both AI and human AI conditions, potentially because of its high value. Supplementary Figure B5 presents a parallel analysis based directly on tf-idf term weights, without relying on clustering, and reveals qualitatively similar dynamics.

Robustness of Results under Temperature, Model, and Prompt Manipulations

Since AI-generated outputs can be sensitive to model-specific parameters and prompting choices, it is important to assess whether these effects reflect general properties of human-AI interaction rather than the idiosyncrasies of a particular model configuration. We therefore conducted a series of control analyses varying the temperature parameter (see below), the underlying AI model, and the prompt, to test the robustness of the observed

**Figure 5**

Controlling for the effect of temperature on the dynamics of creativity and diversity in AI-only populations. We considered high (1.1, 1.4, 1.7; panels **A,B**) and low (0.1, 0.4, 0.7; panels **C,D**) temperature regimes (temperature values range between 0 – 2). Dashed lines correspond to results using the default temperature parameter of 1 (the results presented in Figure 2). Shaded areas represent one standard deviation.

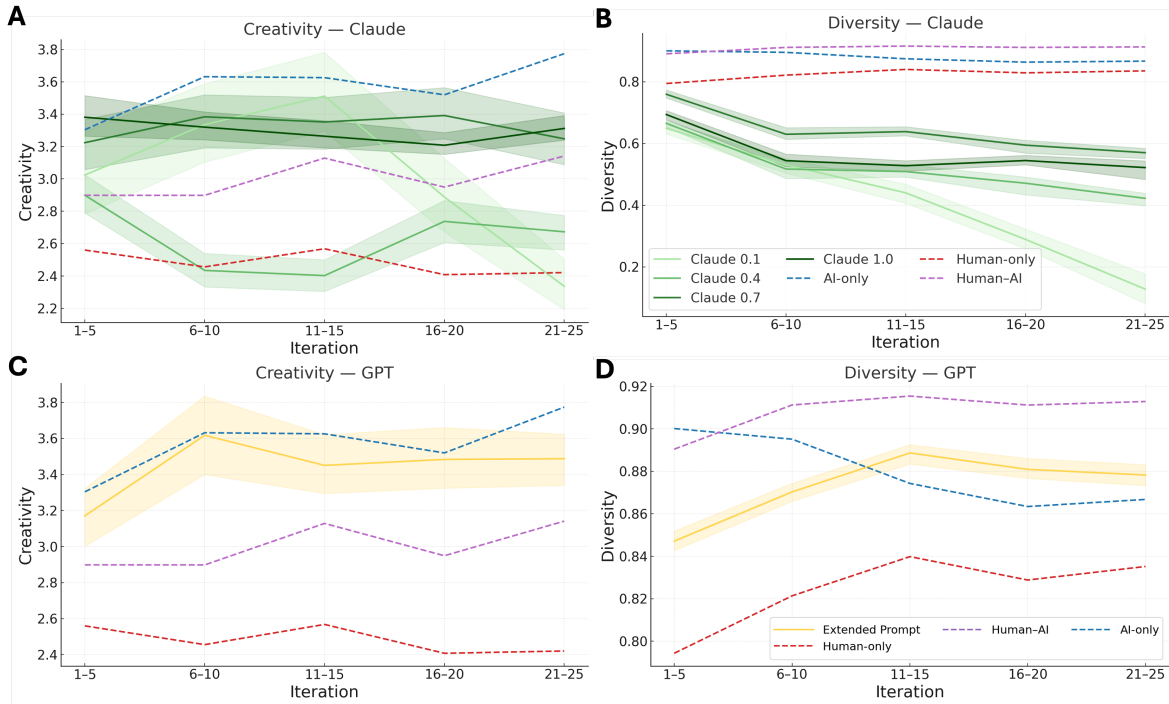
patterns.

First, the temperature parameter controls the stochasticity of the AI model’s text generation, ranging from 0 (deterministic output) to 2 (high randomness), with 1 as the default value. Because higher temperatures allow for greater variation in generated content, we tested whether adjusting this parameter could modulate the diversity (and potentially the creativity) of the resulting stories. The results are shown in Figure 5. For high temperature values (1.1, 1.4, 1.7; Figure 5A-B), the diversity of the AI-only population remains the highest (iterations 21-25: Mean score for GPT temperature 1.1: $M = 0.87$, $SD = 0.04$; GPT temperature 1.4: $M = 0.95$, $SD = 0.03$; GPT temperature 1.7: $M = 0.99$, $SD = 0.01$). However, this enhanced diversity comes at the cost of significantly

lower creativity scores (iterations 21-25: Mean score GPT temperature 1.1: $M = 3.72$, $SD = 0.83$; GPT temperature=1.4: $M = 2.63$, $SD=1.31$; GPT temperature=1.7: $M = 1.29$, $SD = 0.76$). Looking at stories generated with different temperature values (1.1, 1.4, and 1.7) reveals that as temperature increases, the generated narratives contain a higher proportion of random symbols and semantically unrelated words, resulting in reduced coherence and narrative appropriateness (see Supplementary Figure B6 for example stories generated at different temperature values). For low temperature values (0.1, 0.4, 0.7; Figure 5C-D), the diversity decreases significantly (iterations 21-25: Mean score GPT temperature 0.1: $M = 0.83$, $SD = 0.06$; GPT temperature 0.4: $M = 0.83$, $SD = 0.06$; GPT temperature 0.7: $M = 0.84$, $SD = 0.05$), whereas the creativity scores remained largely unchanged (Mean score GPT 0.1: $M = 3.70$, $SD = 0.88$, GPT 0.4: $M = 3.65$, $SD = 0.85$, GPT 0.7: $M = 3.58$, $SD = 0.85$). These results demonstrate that simply changing the temperature parameter is not sufficient to improve the balance between creativity and diversity, with the default value (1) providing an optimal tradeoff between variability and structure (maximizing creative performance without introducing excessive randomness).

Second, we assessed the robustness of our findings to additional experimental manipulations. In particular, we evaluated an alternative model (Claude-3.7-sonnet, version:2025-0219) across a range of temperatures (0.1–1.0). As shown in Figure 6A, performance for Claude was consistently below that of GPT (dashed line). Moreover, Claude exhibited very low diversity (Figure 6B): in many cases, the model generated identical stories, and this tendency intensified across iterations, leading to a progressive deterioration of the diversity score.

Third, because prompt engineering is known to influence AI-generated outputs, we also tested the sensitivity of our results to changes in the prompt. Specifically, we constructed a longer, more detailed prompt that incorporated all instructions provided to human participants in the initial explanation pages (see Methods). To assess robustness to this manipulation, we ran a version of the experiment in which the full set of human

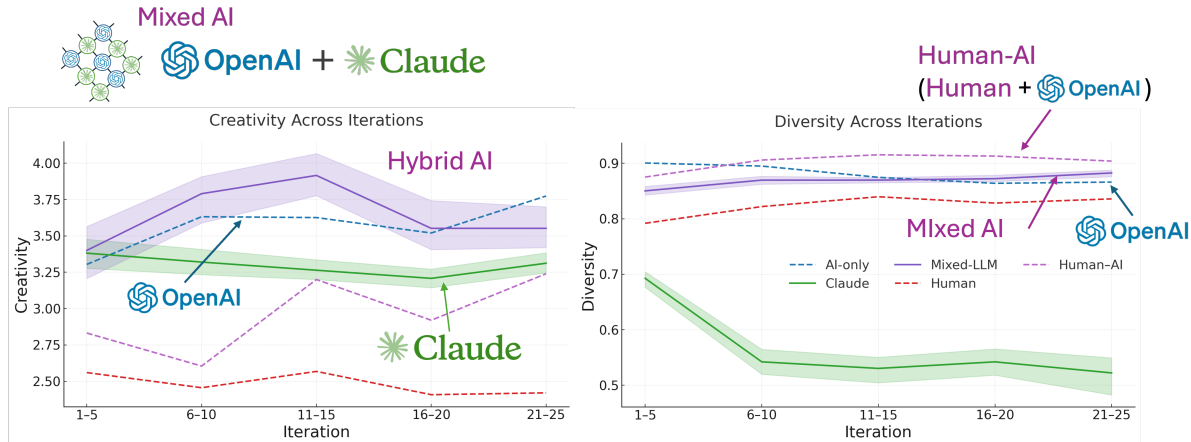
**Figure 6**

*Robustness analyses across models and prompts. **A-B** Results for an alternative model (Claude-3.7-sonnet-2025-0219) across temperature values 0.1–1.0 were consistently below GPT and showed reduced diversity that deteriorated over iterations. **C-D** Results for prompt control. Using an extended prompt containing the full human instructions did not substantially affect creativity in later iterations (21–25) but moderately influenced diversity.*

instructions was given to the AI model (Figure 6C). This manipulation did not substantially alter creativity scores in iterations 21–25 ($M = 3.47$, $SD = 0.93$, $\Delta = -0.30$, $p = 0.02$), but it did have a moderate effect on diversity relative to the original prompt ($\Delta = 0.01$, $p < .001$).

The Advantage of Population Diversity

The control analyses above (Figure 5) 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, is that the balance is achieved through conceptual diversity in the way human and AI agents apply a “creative” edit: humans, on the one hand, seek to maintain core elements of the original

**Figure 7**

Mixed AI models (GPT and Claude). Population creativity and diversity for an additional pure AI population (Claude) and a mixed GPT-4 and Claude population. Dashed lines corresponds to previous GPT-4, human, and human-AI conditions.

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 extended our analysis by introducing a second large language model alongside GPT-4o, namely Claude Version: Claude-3.7-sonnet-2025-0219, and examined creative dynamics in mixed-AI populations. As shown earlier (Figure 6A), this model alone does not provide substantial improvements in either creativity or diversity. This raises the question of whether a hybrid society combining Claude and GPT models might outperform each model individually.

The results for the mixed-AI condition, as well as the AI-only conditions (GPT-4o and Claude) are shown in Figure 7. In the Claude-only condition, creativity did not improve over time ($p = 0.91$, Cohen's $d = -0.16$), resembling the pattern observed in Human-only populations. However, stories in the Claude-only population exhibited a drastic decline in collective diversity over iterations ($p < 0.001$, Cohen's $d = -0.62$), in line with the decline observed in AI-only condition using GPT bots (see Figure 6B for additional temperature controls). In contrast, combining Claude and GPT-4 in a mixed population mitigated these limitations. The mixed-AI condition achieved final creativity

levels comparable to GPT-4 alone and higher than Claude alone (iterations 21–25: $M = 3.55$, $SD = 0.91$, $CI = [3.39, 3.71]$), while also maintaining greater diversity than either single AI population (iterations 21–25: $M = 0.89$, $SD = 0.08$, $CI = [0.88, 0.89]$; mixed vs. Claude: $p < 0.001$, Cohen’s $d = -1.48$; mixed vs. GPT-4: $p < 0.001$, Cohen’s $d = -0.22$). Note that diversity in the mixed-AI condition remained lower than that observed in the Human-AI hybrid population ($p < 0.001$, Cohen’s $d = 0.27$). Thus, population-level diversity arising from a combination of AI models can enhance creative performance while preserving diversity, mirroring (but not exceeding) the benefits observed in human-AI hybrid populations.

Discussion

We investigated how the interaction between humans and AI shapes collective creativity within experimental social networks. Using an iterative storytelling paradigm, we compared how different population compositions (human-only, AI-only, and hybrid human-AI) influence the emergence of creativity and collective diversity over time. This approach allowed us to causally test timely questions that were previously difficult to address empirically: when and why hybrid human-AI systems outperform homogeneous groups of humans or AI agents alone; how diversity and novelty are balanced within collectives; and how humans alter their behaviour in the presence of AI-generated content.

Our results reveal that hybrid human-AI populations consistently achieved a balance combining relatively high creativity with the highest levels of collective diversity across iterations (Figure 2). In contrast, AI-only populations initially reached high levels of creativity and diversity but exhibited a pronounced and systematic decline in diversity over time. Human-only populations, on the other hand, performed worst overall, showing the lowest levels of both creativity and diversity. These patterns were consistent across different dimensions of creativity, including novelty and appropriateness, and were robust to manipulations of model temperature (Figure 5), AI models (GPT vs. Claude), and prompt framing (Figure 6).

Individual and collective creativity

The decline in diversity observed in AI-only networks is particularly important and extends recent findings from studies of human-AI interaction in individual creativity tasks (Doshi & Hauser, 2024; Lee & Chung, 2024; Meincke et al., 2025). 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 (Brinkmann et al., 2023; Tsvetkova et al., 2024). 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. In our study, integrating humans and AI within the same network mitigated this issue. Hybrid populations generated stories that were both creative and structurally diverse, maintaining broad semantic coverage with small conceptual overlap.

Mechanisms of collective creativity

What mechanisms enabled hybrid human-AI populations to achieve this balance between creativity and diversity, outperforming both human-only and AI-only groups? The validation experiments, combined with computational semantic analyses of the stories (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—such as character identities (e.g., maintaining the protagonist, John) and central objects (e.g., “keys”)—thereby maintaining narrative continuity and coherence (see Supplementary Figure B5). 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 (Hart et al., 2018; Hills et al., 2015; Lazer & Friedman, 2007; Steele et al., 2021).

These findings are consistent with classical work on group creativity and collective intelligence, which shows that diversity in cognitive strategies, perspectives, and heuristics can enhance problem-solving and innovation (Centola, 2022; Page, 2008). They are 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 (Shirado & Christakis, 2017) or bots that aggregate information across network neighbors (Ueshima et al., 2024). Our contribution builds on this literature by identifying comparable mechanisms in open-ended, ecologically richer creative settings that map onto established forms of human activity (collaborative writing; Akoury et al., 2020; Du & Chilton, 2023; Lee et al., 2022; Venkatraman et al., 2025). 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 (Bommasani et al., 2022; Chen & Chan, 2024; Epstein, Hertzmann, the Investigators of Human Creativity, et al., 2023; Krakowski, 2025; Lindebaum & Fleming, 2024), 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 behaviour in the presence of AI

Another central question in our study concerns how the presence of AI alters human behaviour within social systems. Recent theoretical work (Sucholutsky et al., 2025; Tsvetkova et al., 2024) suggests that introducing AI agents into societies may reshape human perception, decision-making, and behavioural strategies over time. Empirical evidence from real-world domains supports this perspective. For example, studies in chess

and Go show that human players adopt strategies originally developed by superhuman AI systems (Gaessler & Piezunka, 2023; Shin et al., 2023), leading to lasting shifts in play styles and competitive norms. These findings suggest that AI can influence long-term human behaviour 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, although this effect was stronger in humans, it also extended to AI agents: the decline in diversity observed in AI-only networks was significantly attenuated when AI bots were embedded within hybrid populations. Overall, these findings suggest that hybrid societies induce reciprocal behavioural changes. Rather than exerting a one-directional influence, AI and humans co-shape each other’s behavioural trajectories through sustained interactions in social networks.

Agent diversity and alignment

Finally, 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 artificial agents, rather than integrating humans and AI. To test this, we constructed mixed-agent societies composed of two distinct LLMs (Claude and GPT). We found that when operating in isolation, each model exhibited a marked decline in diversity over time, replicating the convergence dynamics observed in AI-only populations. However, when embedded within a mixed-LLM network, their collective performance improved substantially in terms of diversity (Figure 7). 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 (Wang et al., 2024). 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. That said, the performance of the mixed-LLM population in our study did not reach the levels observed in human-AI hybrid populations, suggesting that diversity across model architectures alone may be insufficient to fully replicate the stabilising role played by humans.

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 (Binz et al., 2025; Hashemi & Macy, 2025; Kolluri et al., 2025). 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 (Fadaei et al., 2026; Hashemi & Macy, 2026; Holtz, 2026), 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 (Sucholutsky et al., 2023). 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 B1), 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 partially reproduced using heterogeneous AI agents alone (Figure 7). These findings suggest that misalignment can, in certain contexts, be advantageous: complementary skills and behavioral strategies may integrate to yield superior collective performance. Future work

should more precisely characterize the functional roles humans and AI can occupy in ecologically valid tasks—for example, collaborative code writing (Hu et al., 2026)—and determine when alignment versus diversity produces optimal outcomes.

Limitations and future directions

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. In our experiments, participants interacted within a bounded and relatively short-lived setting, whereas real-world societies involve repeated, long-term interactions in which individuals continuously adapt to one another. Another key limitation is the static structure of our simulated social network, which used a grid-based topology. Real social networks, such as groups of friends or work relationships, tend to have a modular structure with irregular patterns and evolving connections, allowing individuals to actively shape their interactions by forming and adjusting their social ties rather than being confined to static connections (Almaatouq et al., 2020; McPherson et al., 2001). Moreover, our paradigm avoids any direct communication between participants and masks agent information, such as whether one’s partner is a human or an AI. Prior research suggests that the awareness that one is interacting with an AI system may alter individual behavior, which in turn could influence collective dynamics (Tsvetkova et al., 2024).

Importantly, none of these constraints are inherent to our methodological approach. The experimental framework can be extended to incorporate more complex network topologies and to isolate specific social mechanisms, such as transmission and selection (Marjeh et al., 2025). Our ongoing work builds on these findings by introducing explicit social feedback and examining its influence on creativity (Gautheron et al., 2026; Tchernichovski et al., 2025; Urano et al., 2024). We are also extending the paradigm to tasks with objective evaluation criteria, which permit more precise computational analysis (Li et al., 2026), as well as to societally consequential domains such as collaborative coding (Hu et al., 2026). These directions will enable a broader understanding of how differing

task demands shape the dynamics of collective behavior.

Anoter important consideration in interpreting our results is the nature of the creative task itself. The storytelling task used in this study aligns with those traditionally employed in human creativity research (Beaty & Kenett, 2023; Kaufman & Sternberg, 2010). However, the task proved to be relatively easy for a large language model like GPT, which consistently outperformed humans alone both in creativity and diversity. This does not necessarily imply that AI is intrinsically more creative than humans; rather, it highlights the limitations of using tasks designed for human cognition as benchmarks for AI creativity. AI’s advantage may stem from its ability to rapidly generate highly novel yet semantically coherent text, a skill optimized by its training on vast datasets (Brinkmann et al., 2023; Burton et al., 2024). Conversely, human creativity and creative writing is influenced by memory constraints, cognitive biases, and adherence to implicit narrative conventions (Hayes, 2012).

Moving forward, creativity research should move beyond simple tasks that may be well-suited for assessing individual human creativity but fall short in evaluating AI capabilities or capturing the dynamics of human-AI synergy. For example, using multimodal creative tasks that do not only rely on language, such as music creation (Anglada-Tort et al., 2023) and drawings (Hart et al., 2017, 2018; Kumar et al., 2024). Moreover, future studies should focus on designing complex tasks that challenge both humans and AI in meaningful ways, allowing for a more nuanced understanding of collective creativity and underlying mechanisms.

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 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.

The experiments presented here represent an initial step toward an experimental framework for understanding the mechanisms and consequences of collective behaviour in hybrid societies. We show that collective creativity emerges from micro-level interactions within social networks that could not be predicted by human or machine behaviour 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

Data and Code Availability

All experiment and analysis code, as well as the human and AI data will be available upon publication.

Participants

A total of 1,363 participants took part in the study, with 879 contributing to the story writing experiments and 484 to the rating experiments. Demographic information for all participant groups is provided in Table A2. All participants were recruited online via Prolific³. Eligibility criteria required that participants were based in the UK and identified English as their native language. Participants provided informed consent in accordance

³ www.prolific.com

with the approved ethics protocol (Max Planck Ethics Council #202142), and were compensated at a rate of £9 per hour.

Online Experiment Implementation

PsyNet

The behavioural experiments were designed using PsyNet⁴ (Harrison et al., 2020), a framework for complex experiment design that builds on the Dallinger⁵ platform for experiment hosting and deployment. Participants completed the experiments online through a web-browser interface which communicates with a Python server that manages the timeline of the experiment.

Recruitment

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 chains and the required number of iterations per chain were reached. We ran 25 chains on a 5×5 square lattice, each evolving for 25 iterations. These parameters were determined through extensive piloting, as they were found to provide an optimal balance between generating sufficient breadth and depth of creative outcomes and maintaining practical experimental costs.

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. 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.

⁴ <https://www.psynet.dev>

⁵ <https://dallinger.readthedocs.io/en/latest/>

AI Models

We used two common large language models (LLMs), namely, GPT-4o (Hurst et al., 2024) by OpenAI (Version: September 3, 2024) and Claude⁶ by Anthropic (Version: February 19, 2025). These models were accessed using the official API services provided by OpenAI and Anthropic. For a summary of the number of API calls per experiment, see Supplementary Table A2. The 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 bot completes the different required tasks by performing suitable API calls to the specified LLM model. For prompting details, see LLM Prompting below.

Procedure

Story Writing Task

Each story writing experiment involved a 5×5 grid network (with periodic boundaries so that all nodes have exactly four neighbors). Each node in this network is associated with a short story. Initially, all nodes were initialized with 25 distinct seed stories. The original seed story was generated using GPT-4o. We then used the same model to produce 25 distinct close variations of this story. All stories used in this study are available in the OSF repository associated with this project. For illustration, the first seed 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.”

Upon providing consent, participants were randomly assigned to one location (node) in the network. They were then presented with five stories associated with that location and its

⁶ <https://claude.com/product/overview>

neighbors. Participants were instructed to choose the story that they liked most (“Please select your favorite version of the story from the options below.”), and then to modify it creatively (“Please creatively elaborate on the story, adding your own details and ideas”). Each participant contributed only once to the experiment, and the contributed story replaced the previous one in that location, becoming available to other participants. The process continued for 25 generations (or iterations), where a generation is completed when all nodes are updated once. We considered three experimental conditions: *human-only*, where all nodes were occupied by human participants, (2) *AI-only*, where all nodes were simulated using an LLM, and (3) *human-AI*, consisting of an equal distribution of human participants and LLM agents (50%). To minimize bias, participants were not informed that any of the stories may have been generated by AI agents.

Story Generation

We first prompted GPT-4o (OpenAI) to generate a small pool of candidate stories using the instruction: “Please provide a CSV file with 8 very short stories (4 sentences).” From the resulting outputs, we selected one story to serve as the base seed. We then prompted the same model with the full text of this selected story and requested: “Create 25 variations of the story below and output as CSV.” The model returned 25 distinct narrative variations, which were used to initialize the 5×5 network (one per node).

Validation Task

To evaluate the quality of the produced stories across conditions and iterations, we conducted rating experiments by an independent group of human raters. Each participant was presented with 20 randomly-selected stories from all experimental conditions and asked to rate their fit to a target concept of interest on 5-point scale. The main concept was creativity, and the 5-point scale ranged from 1 (not creative at all) to 5 (extremely creative) (“Rate the creativity of the story on a scale from 1 (not creative at all) to 5 (extremely creative)”). As before, participants were not informed whether stories were created by humans or AI.

In addition to creativity, we conducted two further validation experiments in which stories were rated on adjectives grouped into three dimensions: (1) Aesthetic evaluation (liking, interesting, and meaningful), (2) Novelty (novel, original, and surprising), and (3) Appropriateness (appropriate, feasible, and publishable). These dimensions were selected to capture the two core components of creativity typically assessed in the literature: novelty and usefulness (Amabile, 1982, 1983; Doshi & Hauser, 2024; Harvey & Berry, 2023). Novelty reflects the extent to which an idea departs from expectations, capturing originality, surprise, and rarity. Usefulness reflects the practicality and relevance of an idea—in this context, whether the story is coherent, meaningful, and could plausibly become a publishable narrative product if further developed. We also included ratings of aesthetic evaluation, as it plays a central role in the experience and evaluation of narrative creativity (e.g., stories can be technically sound yet aesthetically unengaging). Finally, raters indicated whether they believed each story was generated by a human or an AI model. Together, these measures provide a multidimensional evaluation of story quality and allow us to compare how creativity unfolds across iterations and across human-AI conditions.

Validation Task with AI Prompting

We examined whether human and AI evaluators differ in their ratings by administering a validation task using the following prompt (“Rate the creativity of the story on a scale from 1 (not creative at all) to 5 (extremely creative)”). Ratings produced by humans and by the AI were modestly correlated ($r = 0.42$, $p < 0.001$).

Prompt Control Experiment

To see the robustness of the results to changing the prompt, we created a version of the experiment where all the instructions for humans are provided to the AI (Figure 6). This manipulation was designed to test whether variation in the instruction prompt influences the LLM’s outputs, and whether our main results remain robust to such prompt-level changes. The prompt was in this case:

"Welcome to the Story Experiment! - instruction 1

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.

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.

Story Experiment - instruction 3

- You should replicate the story and 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.

- We will check that your reproduced story is of sufficient quality.

Failure to do so may result in termination of the experiment.

- The experiment takes about 5 minutes.

The experiment will start on the next page. Please pay close attention

to the instructions!"

Failing Criteria

Responses were required to meet predefined validity criteria to be included in the analyses. Specifically, responses had to contain at least 10 words, include at least one complete sentence, and share a minimum of five lexical items with the original story.

Responses failing to meet any of these criteria were failed until a new participants provide valid data.

Statistical Analysis

All analyses comparing experimental conditions and changes over time were conducted using non-parametric bootstrapping to quantify uncertainty. For each iteration group, participants were resampled with replacement, the corresponding set of stories was reconstructed, and the metric of interest was recomputed (e.g., creativity, diversity). This procedure was repeated 1,000 times, yielding a bootstrap distribution of estimates for each analysis.

From these distributions, we computed the mean trend estimate and the associated 95% confidence interval (2.5th and 97.5th percentiles). These estimates represent the expected change in the metric across iteration groups, with uncertainty derived directly from the empirical resampling distribution. Estimates whose confidence intervals did not include zero were interpreted as statistically reliable changes over time.

To compare overall performance across experimental conditions, we computed each condition’s mean score in the last 5 iterations of the experiment (iterations 21–25). For each bootstrap resample, we recomputed the condition means and then calculated pairwise differences between conditions (e.g., AI-only minus Human-only; AI-only minus Human-AI; Human-AI minus Human-only). For each comparison, we report the bootstrap mean difference and its 95% confidence interval. One-sided p -values were computed from the bootstrap distribution of differences as

$$p = \Pr(\mu_A - \mu_B \leq 0),$$

estimated as the fraction of bootstrap replicates in which the difference was ≤ 0 (testing the directional hypothesis that condition A outperforms condition B).

To quantify changes over time, we defined a gain score within each condition as the

800 difference between the mean in the final five iterations and the mean in the first five
 801 iterations:

$$\Delta = \bar{x}_{21-25} - \bar{x}_{1-5}.$$

802 For each bootstrap resample, we recomputed Δ for each condition and then compared
 803 gains by computing pairwise gain differences (e.g., $\Delta_{\text{AI}} - \Delta_{\text{Human}}$). We report bootstrap
 804 mean gain differences and 95% confidence intervals. One-sided p -values for gain
 805 comparisons were computed as

$$p = \Pr(\Delta_A - \Delta_B \leq 0),$$

806 estimated as the fraction of bootstrap replicates in which the gain difference was ≤ 0 .
 807 When testing whether a single condition exhibited improvement over time, we computed

$$p = \Pr(\Delta \leq 0),$$

808 corresponding to the directional hypothesis that the gain is positive.

809 To quantify the magnitude of change, we report Cohen’s d for all mean-based
 810 comparisons (including adjective ratings). For between-condition contrasts, Cohen’s d was
 811 computed as the standardized mean difference using the pooled standard deviation,

$$d = \frac{\bar{x}_A - \bar{x}_B}{s_p}, \quad s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}.$$

812 For within-condition changes over time, we report a standardized gain computed
 813 analogously from the early (1–5) and late (21–25) samples,

$$d_{\Delta} = \frac{\bar{x}_{21-25} - \bar{x}_{1-5}}{s_{p,\Delta}},$$

814 where $s_{p,\Delta}$ is the pooled standard deviation of the two time windows. Confidence intervals

for d were obtained by recomputing the same effect-size statistic within each bootstrap resample.

In Figure 3 we repeated the main analysis splitting the group of AI-human society to separate sub-groups of responses only by AI and only by humans the analysis was otherwise identical to the one of Figure 2.

Comparison of creativity dimensions across groups

To understand the group differences within each creativity dimension we performed the following analysis. For each of the three dimensions (novelty, aesthetics, and appropriateness), we fit mixed-effects models using `statsmodels` (Seabold & Perktold, 2010), with adjective (e.g., liking, interesting, meaningful) and group (AI, hybrid, human) as fixed effects and participant as a random effect.

We then applied Tukey’s Honestly Significant Difference (HSD) post hoc tests (Tukey, 1949) to assess group differences, which are reported in Figure 2C. The same procedure was used for the “written by AI” measure; however, because this measure included only a single adjective, the model contained only group as a fixed effect. Figure B2 shows the scores separated for each adjective, and Figure B3 shows the dynamics of the ratings across iterations (for Figure 2C we used the average over iterations 21-25).

Collective Diversity

We measured the diversity of stories in a population by computing their pairwise semantic similarity. Specifically, we embedded each story using a tf-idf (term frequency–inverse document frequency) vectorization approach (Leskovec et al., 2020) and calculated the pairwise cosine similarity between all story vectors within a predefined iteration groups. Diversity was operationalized as the inverse of the mean cosine similarity, with lower similarity indicating greater diversity.

Concretely, each story was represented as a tf-idf vector using `scikit-learn`’s `TfidfVectorizer` (Pedregosa et al., 2011), where the tf-idf weight of term t in story d is

defined as

$$\text{tfidf}(t, d) = \text{tf}(t, d) \cdot \log \frac{N}{\text{df}(t)},$$

where $\text{tf}(t, d)$ is the term frequency of t in d , $\text{df}(t)$ is the number of documents containing t , and N is the total number of documents in the set. Each story i is therefore represented as a vector $\mathbf{v}_i$ in the tf-idf space. We then computed the pairwise cosine similarity between all pairs of stories within a predefined iteration group using `scikit-learn`'s

`cosine_similarity`:

$$\cos(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|},$$

and obtained the mean similarity within group g by averaging over all unordered pairs (i, j) :

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

where n_g is the number of stories in group g . Diversity was operationalized as

$$D_g = 1 - S_g,$$

so that higher values indicate greater semantic diversity (lower average similarity).

Comparison of adjectives

Participant-level scores were computed by averaging the three adjectives within each dimension (Novelty, Appropriateness, Aesthetic) and then averaging these scores across iterations 21–25 for each participant and condition. We then fit a linear mixed-effects model (random intercept per participant) to these participant-level scores using `statsmodels`. The full model included condition, dimension, and their interaction as fixed effects:

$$\bar{y}_{p,k} = \beta_0 + \beta_{\text{cond}}(\text{condition}_p) + \beta_{\text{dim}}(k) + \beta_{\text{int}}(\text{condition}_p, k) + u_p + \varepsilon_{p,k},$$

where $k \in \text{Novelty, Appropriateness, Aesthetic}$ indexes the creativity dimension,
 $u_p \sim \mathcal{N}(0, \sigma_u^2)$ captures between-participant variability and $\varepsilon_{p,k} \sim \mathcal{N}(0, \sigma^2)$ is residual error.
 Models were fit using full maximum likelihood (ML) estimation, rather than restricted
 maximum likelihood (REML), to allow valid likelihood-ratio comparisons between models
 differing in their fixed effects.

To test for (i) an overall condition effect averaged across dimensions, (ii) a condition
 $\times$ dimension interaction, and (iii) any condition-related effect, we compared nested models
 via likelihood-ratio tests (LRTs) (Casella & Berger, 2024):

$$\chi^2 = 2(\ell_{\text{full}} - \ell_{\text{reduced}}),$$

with degrees of freedom equal to the difference in the number of fixed-effect parameters
 between models (with three conditions and three dimensions, $\text{df} = 2$ for the condition main
 effect, $\text{df} = 4$ for the interaction, and $\text{df} = 6$ for any condition-related effect).

Finally, to report interpretable pairwise group differences within each dimension, we
 performed Tukey HSD post hoc tests on the participant-level means $\bar{y}_{p,k}$ separately for each
 dimension. Significant comparisons are visualized in Figure 2C. Figure B2 presents ratings
 separated by adjective, and Figure B3 shows rating dynamics across iterations; all
 inferential statistics in Figure 2C are based on participant-level averages over iterations
 21–25.

For the “Do you think this was written by AI?” rating, we applied the same
 iteration filtering (21–25) and participant-level aggregation. Because this yielded one
 participant-level score per condition without a within-participant dimension factor, we
 tested for a condition effect using a one-way ANOVA implemented as an Ordinary Least
 Squares (OLS) model:

$$\bar{y}_p = \beta_0 + \beta_{\text{cond}}(\text{condition}_p) + \varepsilon_p,$$

followed by Tukey HSD post hoc tests to assess pairwise condition differences.

Semantic Analysis (Figure 4)

To characterize higher-level structure in the semantic space of stories, we pooled all stories from the AI-only, Human-only, and Human-AI conditions and embedded them in a shared sentence-level representation. Each story was encoded using a pretrained sentence transformer model (all-MiniLM-L6-v2; Wang et al., 2020), yielding a 384-dimensional embedding that captures distributional semantic similarity between stories. To visualize the semantic space of stories, we then used the UMAP embedding algorithm (McInnes et al., 2018) applied to the sentence-level embeddings (Figure 4A-B). UMAP was then applied to this similarity structure (package: `umap-learn`⁷, metric = euclidean, n_neighbors = 15, min_dist = 0.1, n_components = 2, random_state = 42), producing a two-dimensional embedding space in which stories positioned closer together share more overlapping semantic content, whereas distant points reflect semantically divergent stories. The resulting embedding provides a continuous representation of semantic proximity and allows visual inspection of global semantic structure across the entire corpus.

On this embedding space, we performed k -means clustering to identify groups of semantically similar stories. We evaluated cluster solutions for $k = 2 \dots 10$ by computing within-cluster sum of squares and applying the Kneedle algorithm (Satopaa et al., 2011) to identify the elbow point. The elbow was located at $k = 4$, which was used as the final clustering solution for all subsequent analyses and visualizations (Figure 4B-F).

To quantify creativity performance across semantic clusters, we calculated the average creativity ratings for each cluster (Figure 4C). To help interpretability, we computed word clouds for different story clusters (Figure 4D), using the package `wordcloud`⁸ with parameters max_words = 20. Finally, we computed the proportion of stories assigned to each of the four clusters over iterations. These proportions were then plotted within each condition to track how the distribution of semantic clusters evolved across conditions

⁷ <https://umap-learn.readthedocs.io/en/latest/>

⁸ <https://pypi.org/project/wordcloud/>

(Figure 4E).

Term Dynamics Visualization (Figure B5)

As a supporting analysis we analyzed leading terms (words that appear in stories) and changes in their frequency over the iterations (Figure B5). To do so, each story was first transformed into a tf-idf weighted vector representation and the terms with the highest mean tf-idf scores across the corpus were selected as global keywords.

To track the evolution of key terms across iterations, we first identified globally salient lexical items using tf-idf applied across the full corpus of stories (Human, AI-only, and Human-AI conditions combined; Figure B5B). Specifically, all stories were vectorized using a tf-idf representation with English stopwords removed, and the terms with the highest mean tf-idf scores across the corpus (17 terms) were selected as global keywords. This procedure ensured that term tracking was based on corpus-level salience rather than condition-specific frequency artifacts.

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 of the selected terms appeared with a non-zero tf-idf weight. The results are plotted in Figure B5

References

- Acar, O. A., Tuncdogan, A., van Knippenberg, D., & Lakhani, K. R. (2024). Collective creativity and innovation: An interdisciplinary review, integration, and research agenda. *Journal of Management*, 50(6), 2119–2151.
- Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44), e2313790120.
- Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N., & Iyyer, M. (2020). Storium: A dataset and evaluation platform for machine-in-the-loop story generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6470–6484.
<https://doi.org/10.18653/v1/2020.emnlp-main.525>
- Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaïd, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21), 11379–11386.
<https://doi.org/10.1073/pnas.1917687117>
- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997.
- Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of personality and social psychology*, 45(2), 357.
- Anglada-Tort, M., Harrison, P. M., Lee, H., & Jacoby, N. (2023). Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Current Biology*, 33(8), 1472–1486.
- Balietti, S., Goldstone, R. L., & Helbing, D. (2016). Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30), 8414–8419.

- Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, 4(8), 170193. <https://doi.org/10.1098/rsos.170193>
- Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7), 671–683.
- Benedek, M., Beaty, R. E., Schacter, D. L., & Kenett, Y. N. (2023). The role of memory in creative ideation. *Nature Reviews Psychology*, 2(4), 246–257.
- Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., . . . Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, 644(8078), 1002–1009. <https://doi.org/10.1038/s41586-025-09215-4>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022). On the opportunities and risks of foundation models. <https://arxiv.org/abs/2108.07258>
- Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2024). The crowdless future? generative ai and creative problem-solving. *Organization Science*, 35(5), 1589–1607.
- Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., et al. (2023). Machine culture. *Nature Human Behaviour*, 7(11), 1855–1868.
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative ai at work. *The Quarterly Journal of Economics*, 140(2), 889–942.
- Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., et al. (2024). How

large language models can reshape collective intelligence. *Nature Human Behaviour*, 8(9), 1643–1655.

- Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., ... Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6), pgae191. <https://doi.org/10.1093/pnasnexus/pgae191>
- Casella, G., & Berger, R. (2024). *Statistical inference*. Chapman; Hall/CRC.
- Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996), 1194–1197. <https://doi.org/10.1126/science.1185231>
- Centola, D. (2022). The network science of collective intelligence. *Trends in Cognitive Sciences*, 26(11), 923–941.
- Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112(7), 1989–1994.
- Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3), 702–734.
- Chen, Z., & Chan, J. (2024). Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 70(12), 9101–9117. <https://doi.org/10.1287/mnsc.2023.4863>
- Dere, M., Beugin, M.-P., Godelle, B., & Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, 503(7476), 389–391.
- Dere, M., & Boyd, R. (2015). The foundations of the human cultural niche. *Nature Communications*, 6(1), 8398.

- Doshi, A. R., & Hauser, O. P. (2024). Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290.
- Du, Y., & Chilton, L. (2023). Storywars: A dataset and instruction tuning baselines for collaborative story understanding and generation. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3044–3062. <https://doi.org/10.18653/v1/2023.acl-long.171>
- Eisenmann, T. F., Karjus, A., Sola, M. C., Brinkmann, L., Supriyatno, B. I., & Rahwan, I. (2025). Expertise elevates ai usage: Experimental evidence comparing laypeople and professional artists. *arXiv preprint arXiv:2501.12374*.
- Epstein, Z., Hertzmann, A., of Human Creativity, I., Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., et al. (2023). Art and the science of generative ai. *Science*, 380(6650), 1110–1111.
- Epstein, Z., Hertzmann, A., the Investigators of Human Creativity, Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., Mahari, R., Pentland, A., Russakovsky, O., Schroeder, H., & Smith, A. (2023). Art and the science of generative ai. *Science*, 380(6650), 1110–1111. <https://doi.org/10.1126/science.adh4451>
- Fadaei, F., Moran, J. C., & Yasseri, T. (2026). Gender dynamics and homophily in a social network of LLM agents. <https://arxiv.org/abs/2602.02606>
- Flack, J., Ipeirotis, P., Malone, T. W., Mulgan, G., & Page, S. E. (2022). Editorial to the inaugural issue of collective intelligence. *Collective Intelligence*, 1(1), 26339137221114179.
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, 359(6379), eaao0185.

- Gaessler, F., & Piezunka, H. (2023). Training with ai: Evidence from chess computers. *Strategic Management Journal*, 44(11), 2724–2750.
<https://doi.org/https://doi.org/10.1002/smj.3512>
- Galton, F. (1907). Vox populi. *Nature*, 75, 450–451.
- Gautheron, L., Marjeh, R., Conley, D. C., Frey, S., Rubin, H., Schneider, M. D., Tchernichovski, O., & Jacoby, N. (2026). Popularity feedback constrains innovation in cultural markets. <https://arxiv.org/abs/2602.09997>
- Gomez-Rodriguez, M., Leskovec, J., & Krause, A. (2012). Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4), 1–37.
- Grilli, L., & Pedota, M. (2024). Creativity and artificial intelligence: A multilevel perspective. *Creativity and Innovation Management*, 33(2), 234–247.
- Gupta, P., Zhong, Q., Yakura, H., Eisenmann, T., & Rahwan, I. (2025). The role of social learning and collective norm formation in fostering cooperation in llm multi-agent systems. *arXiv preprint arXiv:2510.14401*.
- Harrison, D. A., Price, K. H., Gavin, J. H., & Florey, A. T. (2002). Time, teams, and task performance: Changing effects of surface-and deep-level diversity on group functioning. *Academy of Management Journal*, 45(5), 1029–1045.
- Harrison, P., Marjeh, R., Adolphi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. *Advances in Neural Information Processing Systems*, 33, 10659–10671.
- Hart, Y., Goldberg, H., Striem-Amit, E., Mayo, A. E., Noy, L., & Alon, U. (2018). Creative exploration as a scale-invariant search on a meaning landscape. *Nature Communications*, 9(1), 5411.
- Hart, Y., Mayo, A. E., Mayo, R., Rozenkrantz, L., Tendler, A., Alon, U., & Noy, L. (2017). Creative foraging: An experimental paradigm for studying exploration and discovery. *PloS One*, 12(8), e0182133.

- Harvey, S., & Berry, J. W. (2023). Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity. *Academy of Management Review*, 48(3), 504–529.
- Hashemi, F., & Macy, M. W. (2025). Collective social behaviors in llms: An analysis of llms social networks [16 pages]. *Proc. SciSocLLM Workshop: Large Language Models for Scientific and Societal Advances (SciSocLLM @ KDD 2025)*.
<https://openreview.net/forum?id=oFWzp48kI8>
- Hashemi, F., & Macy, M. W. (2026). An empirical study of collective behaviors and social dynamics in large language model agents. <https://arxiv.org/abs/2602.03775>
- Hawkins, R. X., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences*, 23(2), 158–169.
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388.
- Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
- Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science*, 336(6079), 303–304.
- Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19(1), 46–54.
- Hippel, E. v., & Krogh, G. v. (2003). Open source software and the “private-collective” innovation model: Issues for organization science. *Organization science*, 14(2), 209–223.
- Hochberg, M. E., Marquet, P. A., Boyd, R., & Wagner, A. (2017). Innovation: An emerging focus from cells to societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1735), 20160414. <https://doi.org/10.1098/rstb.2016.0414>
- Holtz, D. (2026). The anatomy of the moltbook social graph.
<https://arxiv.org/abs/2602.10131>

- Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46), 16385–16389.
- Hou, J., Wang, L., Wang, G., Wang, H. J., & Yang, S. (2025). The double-edged roles of generative ai in the creative process: Experiments on design work. *Information Systems Research*. <https://doi.org/10.1287/isre.2024.0937>
- Hu, H., Marjieh, R., Collins, K. M., Li, C., Griffiths, T. L., Sucholutsky, I., & Jacoby, N. (2026). Why human guidance matters in collaborative vibe coding. <https://arxiv.org/abs/2602.10473>
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jackson, J. C., Rand, D., Lewis, K., Norton, M. I., & Gray, K. (2017). Agent-based modeling: A guide for social psychologists. *Social Psychological and Personality Science*, 8(4), 387–395.
- Kallas, C. (2013). *Inside the writers' room: Conversations with american TV writers*. Bloomsbury Publishing.
- Kaufman, J. C., & Sternberg, R. J. (2010). *The cambridge handbook of creativity*. Cambridge University Press.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681–10686.
- Kolluri, A., Wu, S., Park, J. S., & Bernstein, M. S. (2025). Finetuning LLMs for human behavior prediction in social science experiments. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 30084–30099.
- Krakowski, S. (2025). Human-AI agency in the age of generative AI. *Information and Organization*, 35(1), 100560. <https://doi.org/10.1016/j.infoandorg.2025.100560>

- Kumar, S., Marjeh, R., Zhang, B., Campbell, D., Hu, M. Y., Bhatt, U., Lake, B., & Griffiths, T. (2024). Comparing abstraction in humans and machines using multimodal serial reproduction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46, 143–150.
- Laroche, J., Bachrach, A., & Noy, L. (2024). De-sync: Disruption of synchronization as a key factor in individual and collective creative processes. *BMC Neuroscience*, 25(1), 67.
- Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4), 667–694.
<https://doi.org/10.2189/asqu.52.4.667>
- Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, 8(10), 1906–1914.
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the 2025 Conference on Human Factors in Computing Systems*, 1–22.
- Lee, H., Jacoby, N., Hennequin, R., & Moussallam, M. (2025). Mechanisms of cultural diversity in urban populations. *Nature Communications*, 16(1), 5192.
<https://doi.org/10.1038/s41467-025-60538-2>
- Lee, M., Liang, P., & Yang, Q. (2022). Coauthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. *Proceedings of the 9th ACM Symposium on Cloud and Edge Computing (hypothetical placeholder venue)*.
<https://doi.org/10.1145/3491102.3502030>
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge University Press.

- Li, C., Marjeh, R., Hu, H., Steyvers, M., Collins, K. M., Sucholutsky, I., & Jacoby, N. (2026). Human-AI synergy supports collective creative search. <https://arxiv.org/abs/2602.10001>
- Li, G., Hammoud, H., Itani, H., Khizbullin, D., & Ghanem, B. (2023). Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36, 51991–52008.
- Lindebaum, D., & Fleming, P. (2024). ChatGPT undermines human reflexivity, scientific responsibility and responsible management research. *British Journal of Management*, 35(2), 566–575. <https://doi.org/10.1111/1467-8551.12745>
- Literat, I. (2012). The work of art in the age of mediated participation: Crowdsourced art and collective creativity. *International Journal of Communication*, 6, 23–23.
- Maher, M. L. (2012). Computational and collective creativity: Who's being creative? *Proceedings of the Third International Conference on Computational Creativity (ICCC 2012)*, 67–71.
- Malone, T. W., & Bernstein, M. (2022). *Handbook of collective intelligence*. MIT Press.
- Marjeh, R., Anglada-Tort, M., Griffiths, T. L., & Jacoby, N. (2025). Characterizing the interaction of cultural evolution mechanisms in experimental social networks. *arXiv preprint arXiv:2502.12847*.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444. <https://doi.org/10.1146/annurev.soc.27.1.415>
- Meincke, L., Nave, G., & Terwiesch, C. (2025). ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour*, 9(6), 1107–1109.

- 1157 Mirowski, P., Mathewson, K. W., Pittman, J., & Evans, R. (2023). Co-writing screenplays
1158 and theatre scripts with language models: Evaluation by industry professionals.
1159 *Proceedings of the 2023 Conference on Human Factors in Computing Systems*, 1–34.
- 1160 Mohammed, S., & Angell, L. C. (2004). Surface-and deep-level diversity in workgroups:
1161 Examining the moderating effects of team orientation and team process on
1162 relationship conflict. *Journal of Organizational Behavior: The International Journal*
1163 *of Industrial, Occupational and Organizational Psychology and Behavior*, 25(8),
1164 1015–1039.
- 1165 Murthy, D. (2024). Sociology of Twitter/X: Trends, challenges, and future research
1166 directions. *Annual Review of Sociology*, 50(1), 169–190.
- 1167 Muthukrishna, M., & Henrich, J. (2016). Innovation in the collective brain. *Philosophical*
1168 *Transactions of the Royal Society B*, 371, doi:10.1098/rstb.2015.0192.
- 1169 Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of
1170 generative artificial intelligence. *Science*, 381(6654), 187–192.
- 1171 Page, S. (2008). *The difference: How the power of diversity creates better groups, firms,*
1172 *schools, and societies-new edition*. Princeton University Press.
- 1173 Parjanen, S. (2012). Experiencing creativity in the organization: From individual creativity
1174 to collective creativity. *Interdisciplinary Journal of Information, Knowledge &*
1175 *Management*, 7.
- 1176 Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P.,
1177 & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv*
1178 *preprint arXiv:2411.10109*.
- 1179 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,
1180 Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,
1181 Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn:
1182 Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

- Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. <https://arxiv.org/abs/2302.06590>
- Perez, J., Léger, C., Ovando-Tellez, M., Foulon, C., Dussauld, J., Oudeyer, P.-Y., & Moulin-Frier, C. (2024). Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882*. <https://arxiv.org/abs/2403.08882>
- Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532–535.
- Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative ai. *Nature Human Behaviour*, 7(11), 1836–1838.
- Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, 286(1907).
- Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854–856. <https://doi.org/10.1126/science.1121066>
- Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *2011 31st international conference on distributed computing systems workshops*, 166–171.
- Sawyer, K. (2017). *Group genius: The creative power of collaboration*. Basic books.
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.
- Seo, E., Kang, H., & Song, J. (2020). Blending talents for innovation: Team composition for cross-border r&d collaboration within multinational corporations. *Journal of International Business Studies*, 51(5), 851–885.
- Shi, Y., & Sun, L. (2024). How generative AI is transforming journalism: Development, application and ethics. *Journalism and Media*, 5(2), 582–594.

- 1209 Shin, M., Kim, J., Van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial
1210 intelligence can improve human decision-making by increasing novelty. *Proceedings*
1211 *of the National Academy of Sciences*, 120(12), e2214840120.
- 1212 Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global
1213 human coordination in network experiments. *Nature*, 545(7654), 370–374.
- 1214 SonicRim, L. S. (2001). Collective creativity. *Design*, 6(3), 1–6.
- 1215 Steele, L. M., Hardy, I., Jay H., Day, E. A., Watts, L. L., & Mumford, M. D. (2021).
1216 Navigating creative paradoxes: Exploration and exploitation effort drive novelty and
1217 usefulness. *Psychology of Aesthetics, Creativity, and the Arts*, 15(1), 149–164.
1218 <https://doi.org/10.1037/aca0000236>
- 1219 Sucholutsky, I., Collins, K. M., Jacoby, N., Thompson, B. D., & Hawkins, R. D. (2025).
1220 Using llms to advance the cognitive science of collectives. *Nature Computational*
1221 *Science*, 1–4. <https://doi.org/10.1038/s43588-025-00848-z>
- 1222 Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C.,
1223 et al. (2023). Getting aligned on representational alignment. *arXiv preprint*
1224 *arXiv:2310.13018*. <https://arxiv.org/abs/2310.13018>
- 1225 Tchernichovski, O., Globerson, E., Harrison, P., & Jacoby, N. (2025). How constraints on
1226 editing affects cultural evolution. <https://arxiv.org/abs/2502.16694>
- 1227 Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R.,
1228 Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., et al. (2024). AI can
1229 help humans find common ground in democratic deliberation. *Science*, 386(6719),
1230 eadq2852.
- 1231 Thompson, B., Van Opheusden, B., Sumers, T., & Griffiths, T. (2022). Complex cognitive
1232 algorithms preserved by selective social learning in experimental populations.
1233 *Science*, 376(6588), 95–98.
- 1234 Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

- 1235 Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans
1236 and machines. *Nature Human Behaviour*, 8(10), 1864–1876.
- 1237 Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*,
1238 99–114.
- 1239 Ueshima, A., Jones, M. I., & Christakis, N. A. (2024). Simple autonomous agents can
1240 enhance creative semantic discovery by human groups. *Nature Communications*,
1241 15(1), 5212. <https://doi.org/10.1038/s41467-024-49528-y>
- 1242 Urano, Y., Marjeh, R., Griffiths, T. L., & Jacoby, N. (2024). The influence of social
1243 information and presentation interface on aesthetic evaluations. *Proceedings of the*
1244 *46th Annual Meeting of the Cognitive Science Society*, 5334–5341.
- 1245 Venkatraman, S., Tripto, N. I., & Lee, D. (2025). Collabstory: Multi-LLM collaborative
1246 story generation and authorship analysis. *Findings of the Association for*
1247 *Computational Linguistics: NAACL 2025*, 3665–3679.
1248 <https://doi.org/10.18653/v1/2025.findings-naacl.203>
- 1249 Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). Mixture-of-agents
1250 enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.
- 1251 Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep
1252 self-attention distillation for task-agnostic compression of pre-trained transformers.
1253 *Advances in Neural Information Processing Systems*, 33, 5776–5788.
- 1254 Yu, L., Nickerson, J. V., & Sakamoto, Y. (2012). Collective creativity: Where we are and
1255 where we might go. *arXiv preprint arXiv:1204.3890*.
- 1256 Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large
1257 language models transform computational social science? *Computational*
1258 *Linguistics*, 50(1), 237–291.

Appendix A

Additional Experiment Information

Table A1*Overview of 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	271
Experiment 2	Validation	Human	—	—	1875	94	0
Experiment 3	Validation	Human	—	—	9375	469	0
Experiment 4a	Writing	GPT	25	25	3750	0	3750
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

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 4 compares GPT, Claude, and mixed GPT-Claude evolution. Experiment 2 and 3 are one-shot human evaluations with independent human judgments, and Experiment 5 is a validation experiment with GPT evaluations.

Table A2*Summary of participant counts and AI model API calls per experiment condition.*

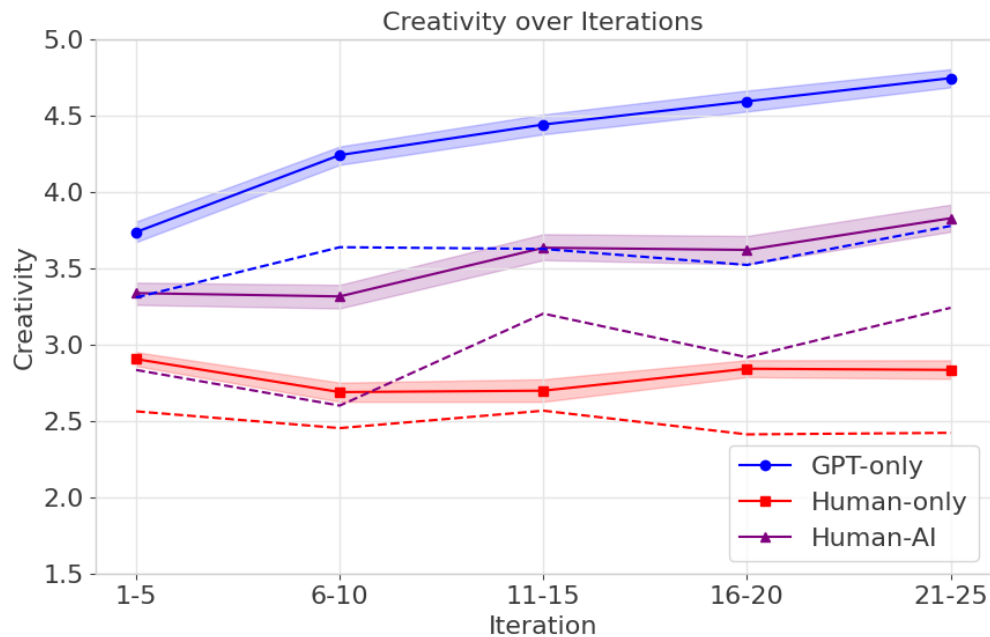
Experiment	Task	Condition	N (Failed)	N_{calls}	Age (Mean $\pm$ SD)	Sex (F/M)
Experiment 1	Writing	Human	625 (173)	0	41.53 ± 13.61	354 / 326
Experiment 1a	Writing	AI	0	625	–	–
Experiment 1b	Writing	Human-AI	254 (164)	271	43.15 ± 14.39	148 / 187
Experiment 2	Validation	Human	106 (2)	0	44.38 ± 14.27	40 / 57
Experiment 3	Validation	Human	378 (22)	0	46.60 ± 14.82	158 / 186
Experiment 4a	Writing	GPT	0	3750	–	–
Experiment 4b	Writing	Claude	0	2500	–	–
Experiment 4c	Writing	Mixed-LLM	0	625	–	–
Experiment 5	Validation	AI	0	94	–	–

All human participants were recruited via Prolific and were based in the United Kingdom.

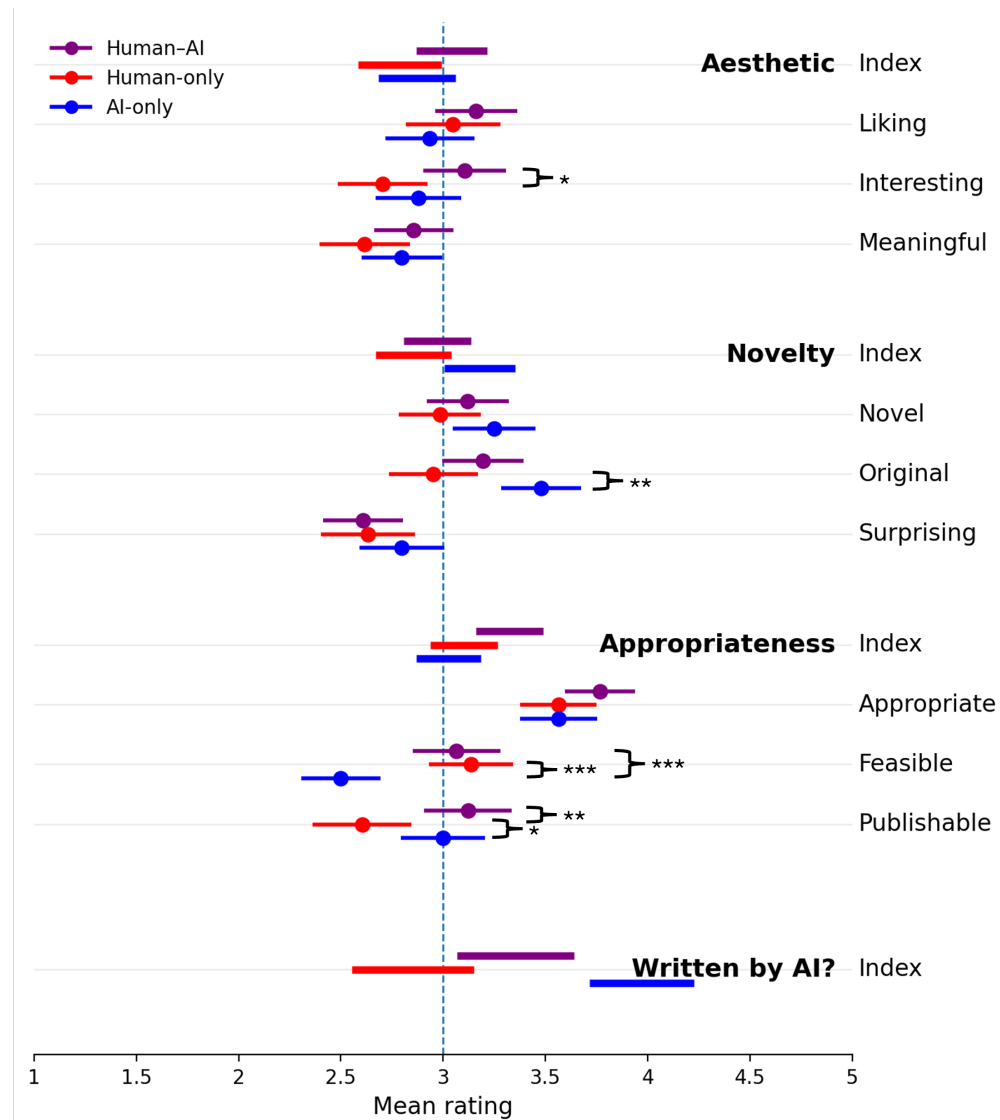
12 participants reported non-binary, prefer-not-to-say, or other gender responses
 (Experiment 1 Human: 8; Experiment 1 Human-AI: 3; Experiment 3: 1)

Appendix B

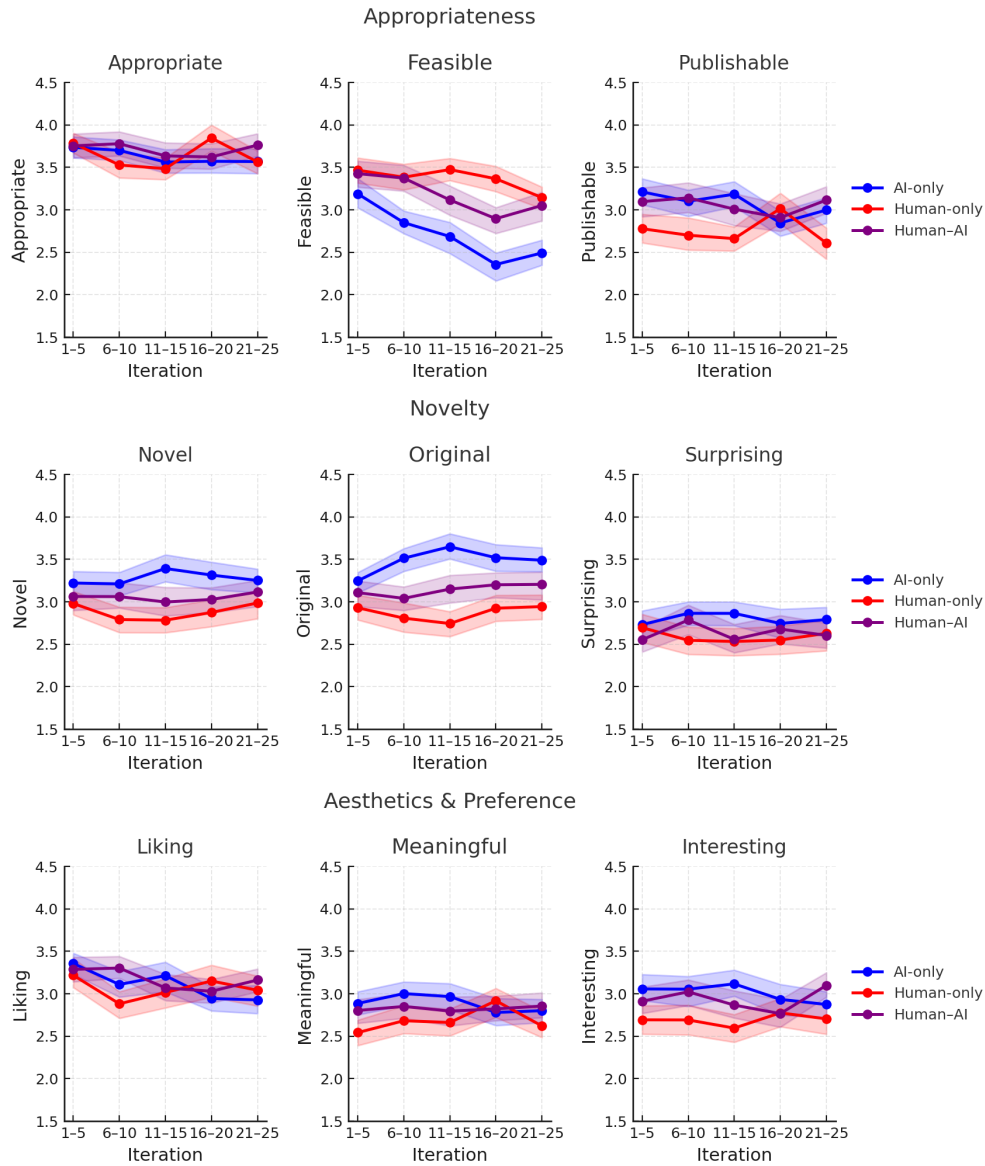
Additional figures

**Figure B1**

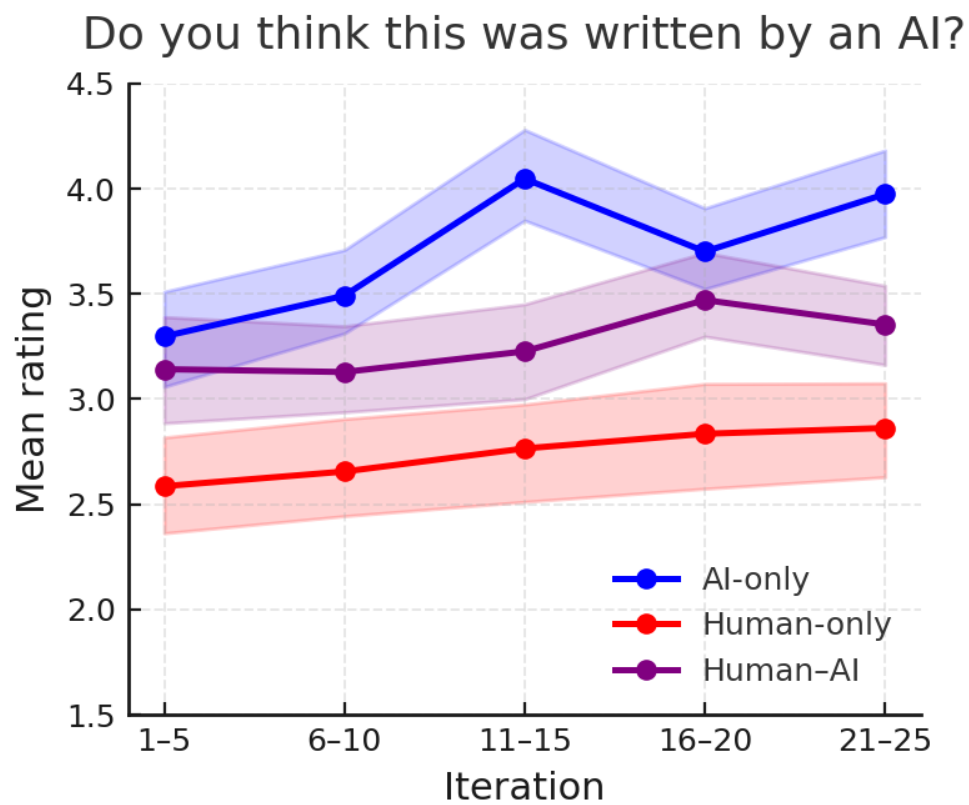
Mean creativity ratings of stories over time, as evaluated by GPT. Shaded areas represent 95% bootstrap confidence intervals. Dashed lines represent human evaluation scores. The horizontal axis represents the 25 iterations, grouped into five sets of five iterations each. Critically, GPT models systematically 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.

**Figure B2**

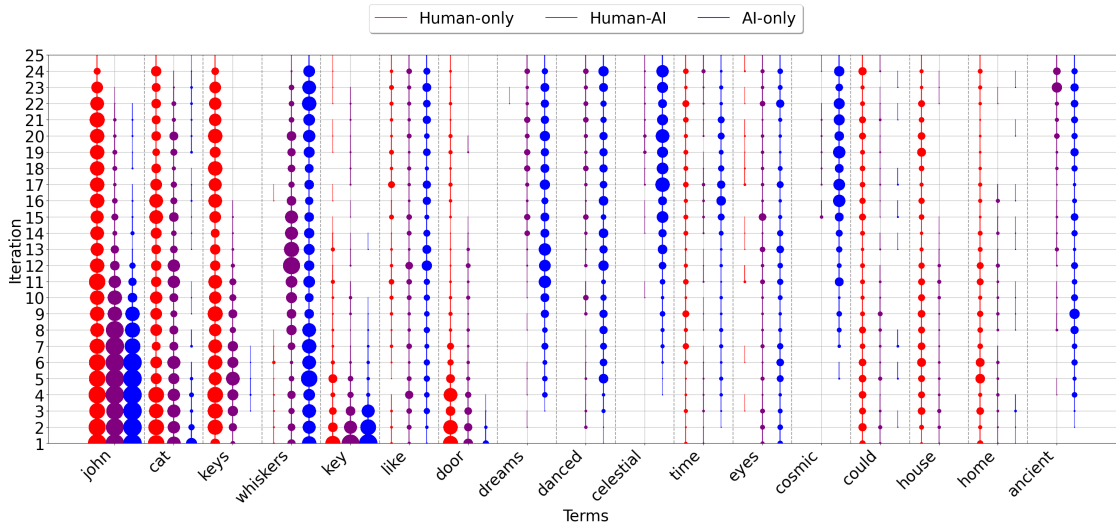
Detailed results including ratings on individual adjectives in the last 5 iterations. Error bars indicate one standard error. *, **, *** indicate $p < 0.05, 0.01, 0.001$ obtained using a mixed effect model (see Methods).

**Figure B3**

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

**Figure B4**

Detailed results including ratings on "Do you think this was written by an AI?" and its evolution across all iterations. Shaded areas represent 95% bootstrap confidence intervals.

**Figure B5**

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 suggests 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.

“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.”

“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”

“Phil Rodríguez knew he had broke
 through glass not noted cream
 AnSerrorible mural latterdays chalk
 line 梦date bathadows avenues
 earYDermo ðia chooses
 cǎnhñehynpř.seineri számct롬-
 wall.less capita 早餐 deinem
 references lawns–pairslandjér ben
 palaisquency ooptes grown
 looph ۛمابلی FluRR_fire hynrei du
 अरपन ज्यःॐॐ sépar_direct itfille
 lines_ŸSYdyaf iglesiasphivlað 손
 terms faucibus) Lou_zape Ješli
 ۛpärklich.rectangle shut
 WRONG STATSparse”

“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.”

[illegible]

“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.”

Figure B6

Examples of the stories from different temperature parameters (see Figure 5).