Creative or Uncreative Partner: Comparing Humans and AI in Collaborative Creative Tasks

Author name(s): Clin KY Lai, Simone A Luchini, Nina Lauharatanahirun, Roger Beaty
Affiliation: The Pennsylvania State University
Corresponding author address (full address is required):
140 Moore Building,
138 Fischer Road,
University Park,
16802
Corresponding author email address: clin.lai@psu.edu

Abstract:

Creativity is fundamentally a collaborative process. Yet as generative AI becomes increasingly integrated into creative work, understanding how AI reshapes collaboration has become critical. This pre-registered study directly compares human-human and human-AI collaboration dynamics across two creative tasks: the Alternative Uses Task (AUT) and creative short story writing. Participants were randomly assigned to pairs in either human-human (N = 68 pairs) or human-AI (GPT-4o; N = 72 pairs) conditions, with partners alternating turns as first responders to examine how initiation order shapes the creative process over time. Our findings reveal that the apparent "AI advantage" in creative collaboration is illusory, driven primarily by increased AI verbosity rather than enhanced creativity. Critically, collaboration with AI partners negatively impacted humans' own creative responses compared to human-human partnerships, with human-AI collaboration failing to enhance idea originality or diversity relative to human-human collaboration. Human partners demonstrated superior collaborative effectiveness that strengthened over time, indicating that current generative AI systems, while producing more verbose outputs, do not replicate the collective creativity characteristic of human-human collaboration. These results challenge assumptions about AI's creative potential, with direct implications for AI system design and collaborative creative practice.

Keywords: co-creativity, creativity, artificial intelligence, Human-AI team, human-computer interaction, social perception

Highlights:

- Human-AI collaboration does not enhance originality or diversity compared to human-human collaboration
- Human-AI pairs produce more verbose but not more creative ideas and stories
- Humans are better creative partners, especially as collaboration progresses

1. Introduction

With the rapid integration of generative AI (genAI) across society, a critical question arises: How does AI reshape collaborative creativity? Some scholars argue that genAI represents a new wave of innovation, capable of driving breakthroughs in arts, science, law, and beyond (Amabile, 2020; Choi & Schwarcz, 2023; Epstein et al., 2023; Rafner et al., 2023). Others, however, contend that genAI's inherent limitations – such as its tendency to reduce diversity of responses (Doshi & Hauser, 2024) and its inability to exhibit intentional, human-like creativity (Chamberlain et al., 2017; McCormack et al., 2019) – may stifle originality. More concerningly, critics warn that overreliance on AI could erode human cognitive abilities (Barr et al., 2015; Budzyń et al., 2025; Kosmyna et al., 2025).

Existing creativity research predominantly examines individual performance with versus without genAI, rather than comparing human-human with human-AI collaboration. As people increasingly collaborate with AIs rather than other people on creative tasks, it is crucial to understand potential costs and benefits of human vs. AI collaboration. This study therefore investigates a fundamental question: is working with a human or AI partner more beneficial for creativity? To this end, we paired participants with other participants or with GPT-4o and had them work on idea generation and story writing tasks, examining how different collaboration partners impact the creative process.

Background

Human collaboration, particularly between humans from diverse backgrounds, has a net positive in creativity and idea generation (Astutik et al., 2020; Ou et al., 2023; Tadmor et al., 2012; Xue et al., 2018). For instance, collaborative research has a greater impact than single-authored works, with 70% of citations concentrated in 30% of papers and single-authored works consistently receiving the fewest citations (Hsu & Huang, 2010).

However, what does collaboration mean when your partner is not human? Research by Amabile (2020) and Epstein et al. (2023) suggests how AI tools can augment creative processes in business and the arts by providing novel inspiration sources and streamlining iterative design work. Similarly, studies in legal and scientific fields (Choi & Schwarcz, 2023; Rafner et al., 2023) highlight genAI's capacity to accelerate research by rapidly synthesizing case law, generating hypotheses, and identifying patterns across massive datasets that would overwhelm human researchers. These capabilities suggest genAI may serve as a powerful collaborative partner that enhances human ingenuity.

GenAI and Cognition

However, genA also has the potential to constrain cognition. Doshi and Hauser (2024) provide empirical evidence that AI systems tend to cluster outputs around statistically common patterns, effectively reducing the diversity of responses compared to human-generated ideas. This convergence effect is compounded by what other researchers (Chamberlain et al., 2017; McCormack et al., 2019) discussed as a fundamental lack of intentionality in AI systems - while they can remix existing content impressively, they cannot originate truly novel concepts or understand cultural context in ways that drive transformational creativity (cf. Boden 2004). Together, these limitations suggest genAI may inadvertently homogenize creative outputs when used uncritically.

Neuroscience and behavioral research suggest that overreliance on AI may adversely impact human cognition and expertise. Cognitive offloading – where individuals take actions to reduce mental cognitive effort and rely on external sources (Risko & Gilbert, 2016) – is of particular concern. Cognitive offloading tends to lead to shallow retention of information in memory; for instance, heavy smartphone use correlates with 'cognitive miser' behavior (Barr et al, 2015), where users recall where to find information rather than the details (Skulmowski, 2023), promoting external reliance in humans (Ahmad et al., 2023). Over-reliance on AI may weaken cognitive foundations for expertise, such as schema-building and proceduralization (Oakley et al., 2025), and social thinking regions activate only during human interactions (Chaminade et al., 2012). Furthermore, specific uses of AI have been shown to degrade performance and creativity: endoscopists performed worse detecting adenomas when the AI was removed (Budzyń et al., 2025), and generative AI use reduces divergent thinking and metacognition in creative tasks compared to simply using Google (Kosmyna et al., 2025).

Collaborator in Creativity Outputs

The literature on human-AI collaborative creativity has largely focused on comparing human-AI dyads with humans working in isolation. Studies that directly compare human-human collaborative creativity to human-AI collaborative creativity are rare, representing a significant gap in our understanding of collaborative creative processes. In one notable study, researchers investigated whether AI (ChatGPT), the internet (Google Search), or another human serves as the most effective creative thinking partner (Tang et al., 2024). The findings revealed that human-human dyads experienced higher creative confidence and produced more creative ideas in the two alternate uses tasks with no difference in the problem-solving task. Humans felt that there was more cognitive load with working with another human but felt more creative confidence. On the other hand, AI made tasks feel easier for participants. Altogether, this suggests that the psychological and social dimensions of human collaboration may play crucial roles in creative processes that extend beyond mere task completion.

The research community remains divided on how to conceptualize AI's role in creativity, with some viewing it as only a supportive supplementary tool, others treating it more as an equal creative partner, and still others considering it to operate as an entirely autonomous creative agent divorced from human (Faiella et al., 2025). This is relevant when investigating creative collaboration and designing the interaction between humans and AI. For instance, Tang and colleagues (2024) did not alternate between humans and AI in a leadership role. Instead, researchers assigned human participants to always initiate the interaction (role A), while the generative AI was never assigned to be role A. Excluding AI from the leadership role introduces an asymmetry in the comparison conditions and limits the full spectrum of potential interaction patterns, particularly those where AI might serve as the primary ideation source with humans in a refining role. This limitation becomes especially significant when considering that the sequence of idea generation (i.e. stream of ideas that build upon each other) are interdependent (Toubia, 2006). A previous study has also found that while AI-only collaborators initially outperformed human-only and hybrid networks in creativity and diversity during a collaborative writing task, hybrid human-AI networks eventually surpassed AI-only ones in diversity due to humans' tendency to preserve narrative continuity, unlike AI agents. (Shiiku et al., 2025). Furthermore, the specific effects on human responses within these collaborative frameworks remain underexplored, such as whether AI collaboration makes human ideas more or less original.

The Present Study

The present study addresses three critical gaps in past literature: first, contributing to the limited body of work that directly contrasts human and AI collaboration; second, examining the balance of contributions between partners through a structured turn-based approach where AI leads; and third, analyzing the specific impact of genAI on human responses across different collaborative conditions.

Based on existing literature, we propose three hypotheses regarding creative outcomes in collaboration conditions. First, we hypothesize that AI collaborations will produce ideas with greater originality compared to human collaborations, as AI systems can generate novel combinations from vast training datasets (Amabile, 2020). Second, we predict that human collaborations will generate more diverse ideas than AI collaborations, as human collaborators bring varied experiences and cognitive starting points (Fink et al., 2012), while AI tends towards interaction patterns involving prompting and responsive agreement (Malmqvist, 2024). Third, we hypothesize that AI collaborations will result in higher elaboration AKA higher verbosity (Nijstad & Stroebe, 2006).

2. Methods

This study was pre-registered on the Open Science Framework (OSF), including the experimental protocol, primary hypotheses, sample size determination, and planned analytical approach. The preregistration document can be accessed at https://osf.io/tveaf.

Any deviations or exploratory analysis not on the registry are noted in the relevant sections or detailed in the supplementary information.

Participants

A total of 208 participants were recruited from Prolific for the study. The number of participants were based on a power analysis performed with the pwr package in R, based on a between-subjects design. For an independent two samples t-test with a small effect size (d = 0.35), a significance level of 0.05, and a power of 0.8 to minimize Type II errors, a minimum sample size of 65 pairs per group was required to detect a statistically significant difference. Upon taking part in the study, participants were randomly assigned to either a human-AI (72 pairs) or human-human (68 pairs) condition. In exchange for their participation for 30 minutes, participants received $6.

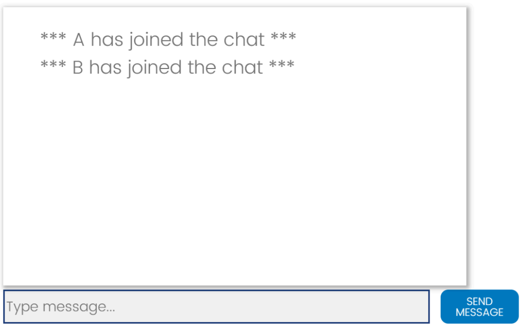
Materials and Procedure

Participants completed two collaborative tasks, a verbal creativity task, and several questionnaires. Upon clicking the study link, they were first presented with the consent page, and could only proceed after providing consent. The experiment was administered on Qualtrics, while the collaborative tasks were hosted on a custom external website. The matching of human participants was performed using the SMARTRIQS platform, a free, open-source tool for interactive Qualtrics experiments involving communication between multiple human participants or chatbots (Molnar, 2019).

Participants were instructed they would be taking part in a turn-taking collaborative creativity study. For collaborative tasks, participants were randomly assigned to either Role A or Role B, and that role persisted throughout the tasks. Participants assigned to Role A would initiate the turn-taking interaction, while Role B would go second. Prior to taking part in the collaborative tasks, participants were explicitly told whether they were assigned to the human-AI or human-human condition. They were also given a chance to practice using the platform with their partner by taking turns to complete the sentence ‘this’, ‘is’, ‘a’, ‘practice’. Role A would start with ‘this’ and alternate with B until the sentence is completed. Next, participants completed two trials of the Alternate Uses Task (AUT) and a short story writing task, taking turns to provide a response. After finishing the collaborative tasks, each participant individually completed a test of verbal creative ability; then, they answered a series of questions regarding their perception of their partner and their general perception of AI. Finally, they answered personality and demographic questions. All experimental tasks were administered on Qualtrics, while the matching of human participants was performed using the SMARTRIQS platform

ChatGPT-4o. The study used prompt engineering to engage ChatGPT-4o as a creativity collaborator in the human-AI condition, tasking the model with generating original ideas for both the Alternative Uses Task (AUT) and creative short story writing. This means prompts were carefully designed with explicit constraints and instructions, so that it served as a creative and useful collaborator in the relevant tasks. The model was accessed via the OpenAI API through PythonAnywhere, by hosting a custom chat interface that was designed to visually match the human partner condition (see Fig. 1).

Figure 1.
Chat interface



Note: This figure is a screenshot of the collaboration platform.

This interface was embedded into Qualtrics as an external website, ensuring a seamless and consistent user experience across both AI and human interaction modes. To standardize interaction timing and match it to the human-human condition, controlled response delays were implemented based on pilot testing. For the AUT, a 5-second delay was implemented between the participant's prompt and ChatGPT-4o's response, while a 10-second delay was imposed for the creative short story writing task. These delays ensured that variations in creativity outcomes were attributable to the nature of the collaboration rather than technical speed of the AI’s response.

Alternate Uses Task (AUT). The AUT measured divergent thinking by asking participants to generate as many creative uses as possible for an everyday object in 3 minutes (Guilford, 1967). Before starting, participants received instructions emphasizing turn-taking and collaboration. Participants were also urged to “be creative” when responding (Acar et al., 2020). The task began with the participant in Role A (or the AI in Role A) submitting an initial response. The participant in Role B would then contribute their own idea, after which it was again the turn of the participant in Role A, and so on. Participants completed the task twice, for two different objects: bottle and book. Originality of AUT responses was assessed using Cross-Lingual Alternate Uses Scoring (CLAUS), a transformer model fine-tuned to match human ratings, hosted on The Creativity Assessment Platform (CAP; Patterson et al., 2025).

Creative Short Story Task. In the creative short story task, participants construct a narrative based on a given set of three words ("stamp-letter-send") and alternate with their partner (human or AI) to develop a coherent story, contributing one sentence per turn. Participants were given a total of 10 minutes to complete the story. Divergent semantic integration (DSI;Johnson et al., 2022) was used as a calculation of the story’s creativity by computing the semantic distance between all pairs of words in the story text. A higher DSI score reflects greater integration of semantically diverse ideas within narratives and higher originality.

Verbal Creative Thinking Ability. To assess each individuals’ baseline level of verbal creative thinking ability, the Divergent Association Task (DAT) was used. In this task, participants are asked to generate 10 unrelated words (excluding proper nouns and technical terms) “that are as different from each other as possible, in all meanings and uses of the words” (Haase et al., 2025; Olson et al., 2021). Participants were given unlimited time to complete the task. Responses were analyzed using pairwise cosine distance comparisons using GloVe and a model that was pretrained on the Common Crawl corpus. The average semantic distance across all word pairs multiplied by 100 to yield the final DAT score. Higher scores indicate greater semantic divergence between words, reflecting stronger verbal creative ability. To account for the AI partner’s baseline creative capacity in our multilevel model, we derived a DAT score for ChatGPT-4o using an established protocol (Bellemare-Pepin et al., 2024; Cropley, 2023; Hubert et al., 2024). See SI 1 for further protocol details.

Perceptions of Partner, Task, and AI. Partner perceptions were assessed across multiple dimensions to capture participants' evaluations of their collaborative experience. Three single-item measures evaluated partner characteristics: creativity (1 = Not at all creative to 7 = Exceptionally creative), cooperativeness (1 = Not at all cooperative to 7 = Extremely cooperative), and ease of collaboration (1 = Extremely difficult to 7 = Extremely easy). Feelings of interpersonal connection were measured using a four-item scale adapted from previous research on human collaboration (Fowler et al., 2024), with participants rating their agreement (1 = Strongly disagree to 7 = Strongly agree) with statements including "I like my partner," "I value my partner," and "I feel connected to my partner." Additionally, participants indicated their perceived relationship closeness using the Inclusion of Other in Self scale (Aron et al., 1992), selecting from seven increasingly overlapping circle diagrams representing the degree of connection with their creative partner.

To verify participant attention and manipulation awareness, attention check questions were administered. Participants were asked to identify whether they were told they were working with another human participant or AI based on the study instructions, and separately to report their own perception of their partner's identity. An honesty check inquired whether participants used external internet or AI resources during the study, with assurance that responses would not affect their compensation. Participants' data were removed if they admitted to using external internet or AI resources not provided to them during the course of the study.

Finally, participants' general attitudes toward AI were assessed using an 11-item scale measuring two positive and negative dimensions of AI perception (Sundar, 2020). The scale included items evaluating AI's machine-like qualities such as "AI has machine-like precision," "AI is error free," and "AI has machine-like accuracy," as well as items assessing perceived human-like characteristics including "AI is able to detect human emotion," "AI has human-like subjective judgements," and "AI has human intuition." Additional items measured perceptions of AI rigidity, including "AI is unyielding," "AI is rigid," and "AI is mechanistic." All AI perception items were rated on a 7-point scale indicating how well each term described participants' thoughts about AI.

Personality and Demographics. Participants also completed brief personality and demographic questionnaires. Personality, in particular openness to experience, was measured as it tends to be positively associated with creativity. It was measured using a brief 10-item scale (Rammstedt & John, 2012) adapted from the Big Five Inventory (John & Srivastava, 1999), with items assessing key traits (e.g., "I see myself as someone who… is reserved", "…has an active imagination") rated on a 5-point Likert scale (1 = Strongly disagree to 5 = Strongly agree). The scale included items representing openness (e.g., artistic interests, imagination), extraversion (e.g., reserved, outgoing), conscientiousness (e.g., thoroughness, laziness), agreeableness (e.g., trusting, fault-finding), and emotional stability (e.g., relaxed, nervousness). Demographic information was collected via self-report, including age, gender, race/ethnicity, highest educational qualification, and whether English was their first language.

Data Analysis

Alternate Uses Task (AUT). Responses for the AUT tasks were aggregated (bottle and book). The lme4 and lmerTest packages in R were used to test the mixed-effects models for originality (hypothesis 1) and elaboration (hypothesis 3). Diversity was measured in two ways: Levene's text for the variance differences in AUT originality scores, and forward flow (hypothesis 2). All models controlled for individual creativity (DAT score) and accounted for nested random effects of pairs and participants.

Hypothesis 1 was tested by comparing the originality of ideas between AI and human collaborations, with AI expected to generate more novel combinations.

Hypothesis 2 predicts that human collaborations will produce more diverse ideas than AI collaborations due to their varied cognitive starting points. This will be assessed using forward flow analysis to quantify flexibility and semantic distance. Forward flow analysis calculates semantic distance (via cosine similarity) between consecutive ideas in a sequence, thereby indexing how conceptually distinct each new idea was from previous ones. To run the analysis, a multilingual sentence transformer model (paraphrase-multilingual-MiniLM-L12-v2) was used to quantify conceptual similarity between consecutive responses within each participant group. For each group, responses were sorted by running sequence and pairwise semantic distances were calculated between adjacent responses using cosine similarity of sentence embeddings. The semantic distance was computed as one minus the cosine similarity between consecutive response embeddings, where higher values indicate greater conceptual divergence, enabling quantification of semantic flow patterns within each participant's response sequence.

Hypothesis 3 expects AI collaborations to yield higher elaboration (word count), as AI can rapidly expand ideas with detailed explanations.

Creative Short Story Task. Pre-registered analyses focused on originality and diversity. Both the Human-Human and Human-AI diversity contribution score distributions violated normality (Shapiro-Wilk test: Human-Human $W = 0.856$, $p < 0.001$; Human-AI $W = 0.946$, $p = 0.004$). Given these violations, we used the non-parametric Mann-Whitney U test

(Wilcoxon rank-sum test) to compare central tendencies between conditions, rather than Welch's t-test which assumes normality. The two-sample Kolmogorov-Smirnov test compares the entire distributions (shape, location, and scale).

3. Results

Participant Demographics

Participant demographics, including age, race, gender, education, and language background, are summarized in Table 1.

Table 1

Demographics of participants in Study 1

	Overall	**AI**	**Human**	***p***
n	208	72	136	
Age (mean (SD))	35.18 (12.07)	36.46 (11.96)	34.51 (12.12)	0.269
Race (%)				0.516
African or African American	55 (26.4)	22 (30.6)	33 (24.3)	
Asian	6 (2.9)	2 (2.8)	4 (2.9)	
Caucasian	112 (53.8)	37 (51.4)	75 (55.1)	
Hispanic	9 (4.3)	5 (6.9)	4 (2.9)	
Mixed Race/Other	22 (10.6)	5 (6.9)	17 (12.5)	
Native American/Alaskan Native	2 (1.0)	0 (0.0)	2 (1.5)	
Prefer not to say	2 (1.0)	1 (1.4)	1 (0.7)	
Gender (%)				0.018
Female	114 (54.8)	42 (58.3)	72 (52.9)	
Male	88 (42.3)	25 (34.7)	63 (46.3)	
Other/Unspecified	6 (2.9)	5 (6.9)	1 (0.7)	
Education (%)				0.123
Associate (2-year college degree)	19 (9.1)	5 (6.9)	14 (10.3)	

Bachelor's (4-year college degree)	96 (46.2)	28 (38.9)	68 (50.0)	
Doctorate	5 (2.4)	4 (5.6)	1 (0.7)	
High School (or equivalent)	54 (26.0)	21 (29.2)	33 (24.3)	
Master's	26 (12.5)	12 (16.7)	14 (10.3)	
Vocational	8 (3.8)	2 (2.8)	6 (4.4)	
English as First Language = Yes (%)	202 (97.1)	69 (95.8)	133 (97.8)	0.713
Extraversion (mean (SD))	2.86 (1.01)	2.90 (1.01)	2.84 (1.01)	0.715
Agreeableness (mean (SD))	3.45 (0.93)	3.44 (0.95)	3.46 (0.93)	0.933
Conscientiousness (mean (SD))	3.89 (0.88)	3.88 (0.97)	3.90 (0.84)	0.864
Neuroticism (mean (SD))	2.72 (1.11)	2.77 (1.13)	2.69 (1.10)	0.64
Openness (mean (SD))	3.71 (0.94)	3.81 (0.99)	3.66 (0.91)	0.273

A chi-square test of independence showed a statistically significant association between gender and condition, $\chi^2(2, N = 208) = 8.07$ p = 018). However, this result was driven by the low-frequency 'Other/Unspecified' category (n = 6 total), which was higher in the AI collaborator condition (6.9%) than human collaborator (0.7%). No differences were observed for Female (58.3% vs. 52.9%) or Male (34.7% vs. 46.3%). Hence, we proceed with the assumption that there is no practical difference between the groups for the purposes of subsequent analysis.

Alternate Uses Task (AUT)

Originality. First, we tested hypothesis 1, whether human-human collaboration differed from human-AI collaboration in creativity. On average, pairs in the human-human condition produced a total of 10 responses (book: 9.81; bottle: 10.20) with 3.34 words per idea, while pairs in the human-AI condition produced 11.7 responses (book: 11.80; bottle: 11.50) with 12.7 words per idea. Humans in the paired group had an average DAT score of 75.4 while the humans in the AI group had a DAT score of 76.7.

Table 2 shows a summary of the effects of condition (human-human vs human-AI) and participant role (role A or role B) on the prediction of AUT response originality. Three nested multi-level models with random intercepts for group ID and group ID nested within response ID were fitted. All models included DAT as a scaled covariate to control for individual verbal creativity effects. Fluency is the total number of ideas for each pair.

The pre-registered model (with random slopes for condition by groupID and random intercepts for groupID and groupID:ResponseId) failed to converge, even after uncorrelating random effects. A reduced model with only the random slope for condition by groupID also failed. The final and successfully converging model include only random intercepts for groupID and the nested groupID by ResponseId term:

*originality ~ condition * role + dat_scaled +*
(1|groupID) + (1|groupID:ResponseId), data)

Table 2

Multi-Level Model Results for Alternative Uses Task Originality Predictions

Variable	Base	Covariates	Human-Only
Fixed Effects	***B* (SE)**	***B* (SE)**	***B* (SE)**
Intercept	0.569*** (0.005)	0.458*** (0.010)	0.431*** (0.014)
Condition (Human-Human)	-0.065*** (0.007)	0.010 (0.008)	0.029** (0.010)
Role B	0.018*** (0.004)	-0.004 (0.004)	-0.015 (0.009)
DAT (scaled)	0.021*** (0.002)	0.008*** (0.002)	0.002 (0.002)
Item (Bottle)	—	-0.007* (0.003)	-0.008* (0.004)
Fluency	—	0.001 (0.001)	0.003** (0.001)
Word Count	—	0.009*** (0.000)	0.009*** (0.001)
Condition × Role B	-0.021** (0.007)	0.002 (0.006)	0.012 (0.011)
Random Effects	***Variance***	***Variance***	***Variance***
Group ID	0.0004	0.0003	0.0004
Group ID:Response ID	0.0004	0.0002	0.0002
Residual	0.0079	0.0059	0.0068
Model Fit			
REML	-5750.7	-6579.2	-4248.2
N (observations)	2,970	2,970	2,066
N (groups)	139	139	137

Note. Base Model includes condition, participant role, and date as predictors. Covariates Model adds item type, fluency, and word count. Human-Only Model excludes AI-generated responses. Reference categories: Condition = AI-Human, Participant Role = A, Item = Brick. Dashes indicate variables not included in the model.
$p < .05$. *$p < .01$. **$p < .001$.

The base model (Model 1) revealed a significant main effect of condition. The Human-Human condition showed a negative effect ($B = −0.065$, $SE = 0.007$, $t = −9.69$, $p < .001$), while participant role B had a small but significant positive effect ($B = 0.018$, $SE = 0.004$, $t = 4.22$, $p < .001$). Additionally, a significant interaction between the Human-Human condition and participant role B was observed ($B = −0.021$, $SE = 0.007$, $t = −2.87$, $p = .005$), suggesting that the effect of condition varied depending on participant role. The scaled DAT score covariate (dat_scaled), measuring individual verbal creativity, was also significant (*B* = 0.021, *SE* = 0.002, *t* = 9.67, *p* < .001). These results indicate that being in the human-AI collaboration condition is, especially in role B, was associated with higher originality scores. However, scaled DAT score suggests that independent of the condition or role, higher verbal creativity also predicts higher originality scores.

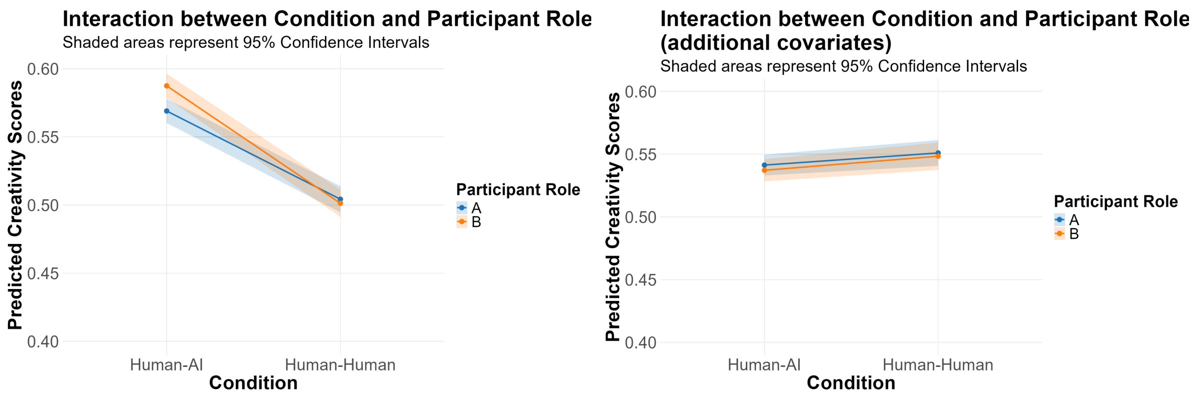
Model 2 incorporated additional covariates including item type, fluency, and word count to control for potential confounding variables: item type (book and bottle), fluency (number of ideas for each pair), and word count. The addition of these covariates substantially improved model fit (REML = -6579.2 vs. -5750.7). Notably, the inclusion of covariates eliminated both the main effect of condition (*B* = 0.010, *SE* = 0.008, *t* = 1.28, *p* = .201), and the condition × participant role interaction (*B* = 0.002, *SE* = 0.006, *t* = 0.24, *p* = .814). The scaled DAT score (dat_scaled) continued to have a positive effect (*B* = 0.0078, *SE* = 0.0019, *t* = 4.02, *p* < .001). Among additional covariates, word count emerged as a strong predictor (*B* = 0.009, *SE* = 0.0003, *t* = 31.61, *p* < .001), while item (Bottle) showed a slight negative effect (*B* = -0.0069, *SE* = 0.0029, *t* = -2.43, *p* = .015), fluency was marginally significant (*B* = 0.001, *SE* = 0.001, *t* = 1.80, *p* = .072). This implies that, after controlling for word count (a robust predictor) and other covariates, there was no difference between AI and human partners, there was no difference in whether one's partner was AI or human.

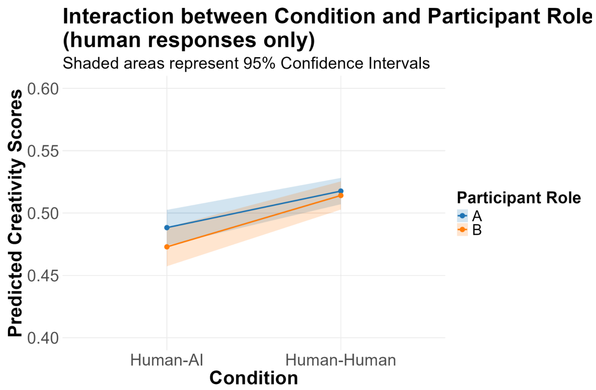
To specifically examine the effect of role in interactions in only human responses, we conducted an analysis excluding all AI-generated responses (Model 3). In this restricted sample, the condition effect re-emerged as significant (*B* = 0.0293, SE = 0.0104, *t* = 2.83, *p* = .005), with human-human pair showing more originality. Importantly, this effect was the opposite of the pre-registered analysis. However, the condition × participant role interaction remained non-significant (*B* = 0.0119, *SE* = 0.0106, *t* = 1.12, *p* = .263), suggesting that partner role did not substantially moderate the effect of condition. Among covariates, fluency (*B* = 0.0026, *SE* = 0.0009, *t* = 2.97, *p* = .003) became significant in this model, while DAT scores were no longer significant (*B* = 0.0019, *SE* = 0.0023, *t* = 0.80, *p* = .425).

Thus, contrary to the pre-registered findings, human collaborators were linked to higher originality when AI-generated responses were excluded, implying that AI responses may have previously masked underlying differences in human responses (see Fig. 2). To ensure that these results were not due purely because of comparison between two versus single person in the human-only model, the analysis was re-run with randomly selected humans in the human-human condition to compare with the human in the human-AI condition. Results were largely the same, with humans in the human condition scoring higher than the human in the AI partner condition, *B* = 0.027, *SE* = 0.011, *t*(175.5) = 2.518, *p* = .013).

Figure 2.

Interaction between condition and participant role for AUT originality score



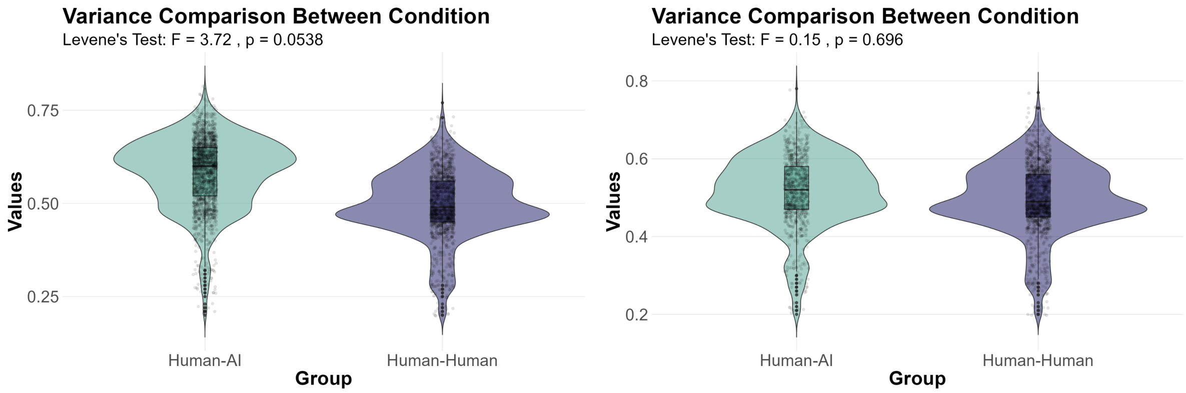


Diversity. Hypothesis 2 asks, does working with human vs AI creative partners produce more diverse ideas? To test this, two analyses were performed. First to examine the range of originality score, and second to use forward flow analysis to examine the semantic distance per idea from the previous. Levene's test for homogeneity of variance was used to examine the range of originality score.

For the first analysis using the median as the center, Levene's test indicated a marginal significant difference in the variance of scores between the two conditions, $F(1, 3036) = 3.72$, $p = .054$. However, removing the AI's responses indicated no difference, $F(1, 2132) = 0.15$, $p = .70$. This implies that the differences in the range of scores were predominantly driven by the AI's responses (see Fig. 3.).

Figure 3.

Comparing diversity of AUT originality scores between condition with score variance

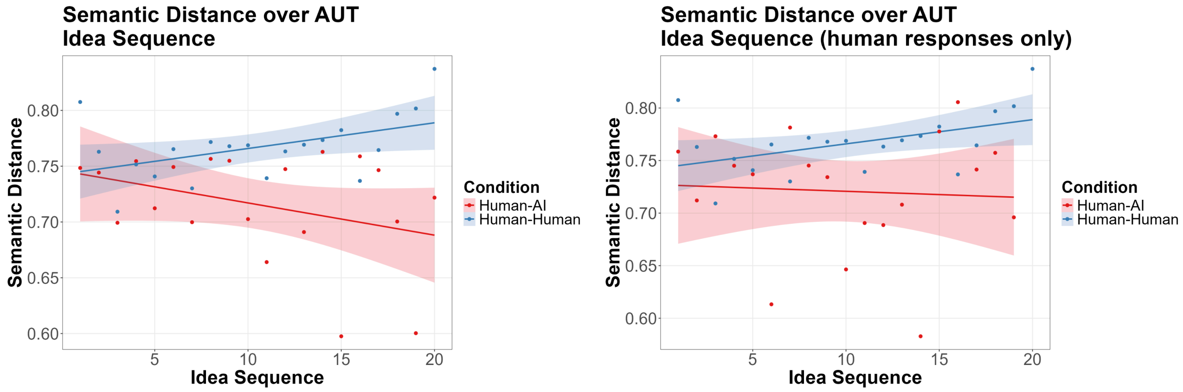


For the second analysis with forward flow, a linear regression analysis examined semantic distance across idea sequence position (1st idea to last idea produced) between collaboration conditions, with data truncated to the first 20 ideas per pair to ensure more equivalent comparison between the conditions. The model revealed a significant main effect of running idea sequence ($B = -0.002$, $SE = 0.001$, $t = -2.20$, $p = .028$) and a significant interaction between condition and running sequence ($B = 0.004$, $SE = 0.002$, $t = 2.43$, $p = .015$). As shown in figure 3, Human-AI pairs demonstrated decreasing semantic distance over time, while Human-Human pairs showed increasing semantic distance across the sequence. The main effect of the condition was not significant ($B = 0.003$, $SE = 0.013$, $t = 0.23$, $p = .818$), which indicates that while the overall semantic distance did not differ between conditions, the trajectory of idea diversity varied significantly, with human-AI pairs exhibiting more convergent thinking and human-human pairs showing more divergent thinking over time.

The results from the model that removed AI's response ideas are identical to the analysis that included the AI responses. The model revealed a significant main effect of running sequence ($B = -0.002$, $SE = 0.001$, $t = -2.20$, $p = .028$) and a significant interaction between condition and running sequence ($B = 0.004$, $SE = 0.002$, $t = 2.43$, $p = .015$). The main effect of the condition was not significant (*B* = 0.003, *SE* = 0.013, *t* = 0.23, *p* = .818). Human participants in Human-AI pairs demonstrated relatively stable semantic distance over time, while humans in Human-Human pairs showed increasing semantic distance across the sequence. These results suggest that the observed patterns reflect differences in human ideation processes over time between the two conditions.

Figure 4.

Comparing diversity of AUT responses between condition with semantic distance



Elaboration. Hypothesis 3 asks if AI partners were wordier than human partners. As indicated in the pre-registration, the linear mixed-effects model that examined elaboration (word count) with condition, participant role, and scaled DAT score as predictors initially failed to converge due to a gradient issue. The final simplified model, which removed the random slope for condition entirely and retained only random intercepts for groupID and the nested groupID:ResponseId structure, successfully converged:

*word count ~ condition * participant role + dat_scaled +*
(1|groupID) + (1|groupID:ResponseId)

Table 3

Multi-Level Model Results for Alternative Uses Task Elaboration (word count)

Variable	**Base**	**Covariates**	**Human-Only**
Fixed Effects	***B* (SE)**	***B* (SE)**	***B* (SE)**
Intercept	11.04*** (0.38)	14.35*** (0.73)	10.04*** (0.69)
Condition (Human-Human)	-7.08*** (0.55)	-9.13*** (0.63)	-5.04*** (0.64)
Role B	2.56*** (0.24)	2.57*** (0.24)	2.67*** (0.67)
DAT Scaled	2.12*** (0.14)	2.09*** (0.14)	-0.01 (0.13)
Item (Bottle)	—	0.71*** (0.18)	0.34** (0.11)
Fluency	—	-0.31*** (0.05)	-0.32*** (0.04)
Condition × Role B	-2.44*** (0.47)	-2.56*** (0.46)	-3.15*** (0.69)

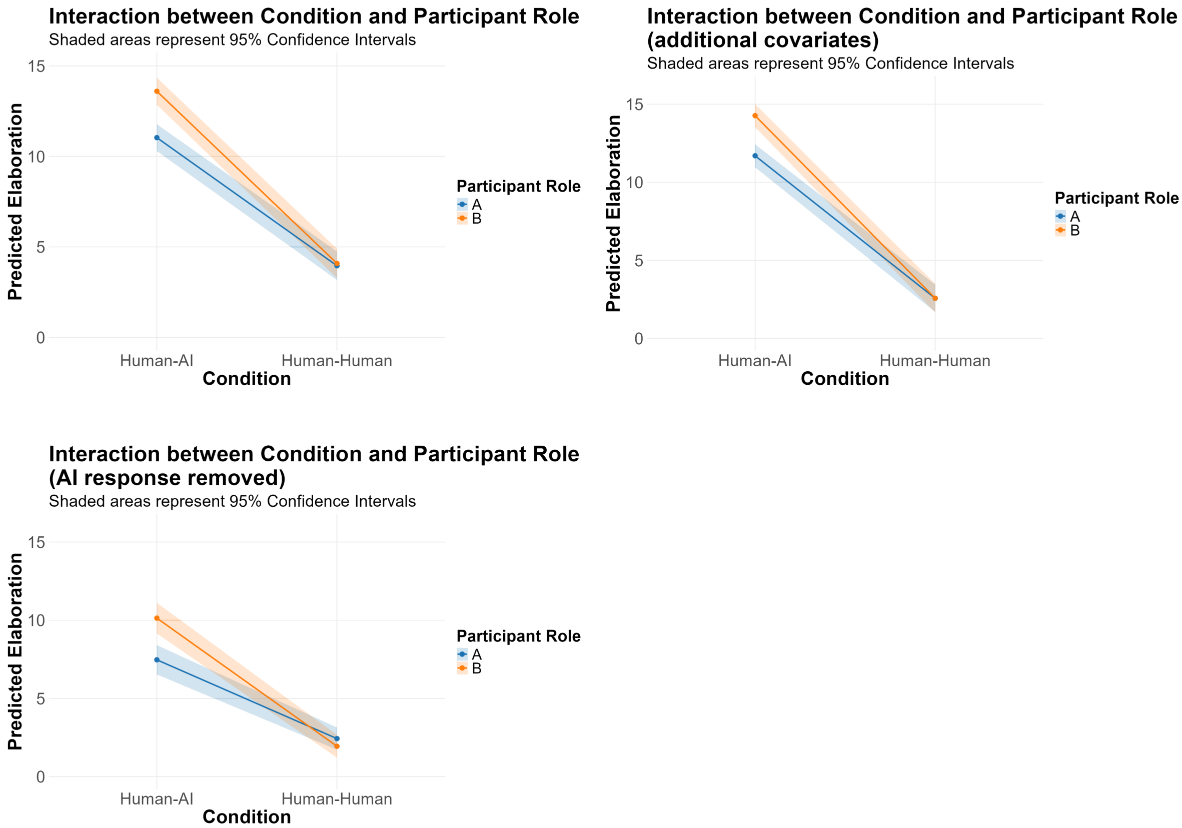
Random Effects			
Group ID	5.50 (2.35)	4.55 (2.13)	6.29 (2.51)
Group ID:Response ID	2.69 (1.64)	2.63 (1.62)	0.63 (0.79)
Residual	23.45 (4.84)	23.18 (4.82)	6.42 (2.53)
Model Fit			
REML	18125.2	18078.7	10124.2
Observations	2,970	2,970	2,066
N (Group)	139	139	137

Note. Base Model includes condition, participant role, and date as predictors. Covariates Model adds item type, fluency, and word count. Human-Only Model excludes AI-generated responses. Reference categories: Condition = AI-Human, Participant Role = A, Item = Brick. Dashes indicate variables not included in the model.
*p < .05. **p < .01. ***p < .001.

Results across all three variations of the model remain consistent (Table 3). The initial model (N = 2,970 observations) with condition, participant role, scaled DAT score, and condition x participant role interaction revealed significant effects for all predictors (ps < .001) with a negative Human-Human condition effect (*B* = -7.08, *SE* = 0.55, *t* = -12.79) and positive participant role effect (*B* = 2.56, *SE* = 0.24, *t* = 10.71). There were also significant positive effects for the scaled DAT score (*B*=2.12, *SE*=0.14, *t* = 14.73, *p*<.001) and a condition × role interaction (*B* = -2.44, *SE* = 0.47, *t* = -5.20, *p* <. 001).

When adding item type and fluency covariates (N = 2,970), the model showed comparable effects with significant negative fluency effects (*B* = -0.31, *SE* = 0.05, *t* = -5.88, *p* < .001) and positive Bottle item effects (*B* = 0.71, *SE* = 0.18, *t* = 3.97, *p* < .001). It maintained similar effects for the scaled DAT score (*B* = 2.09, *t* = 14.78, *SE* = 0.14, *p* < .001) while showing a slightly stronger interaction (*B* = -2.56, *SE* = 0.46, *t* = -5.51, *p* < .001). The human-only subset model (N = 2,066) showed stronger condition × role interaction (*B* = -3.15, *SE* = 0.69, *t* = -4.55, *p* < .001) and non-significant DAT score effects (*p* = .956). This means that there is a robust condition and role effect in the human-AI condition.

Figure 5.
Interaction between condition and participant role for elaboration of AUT responses

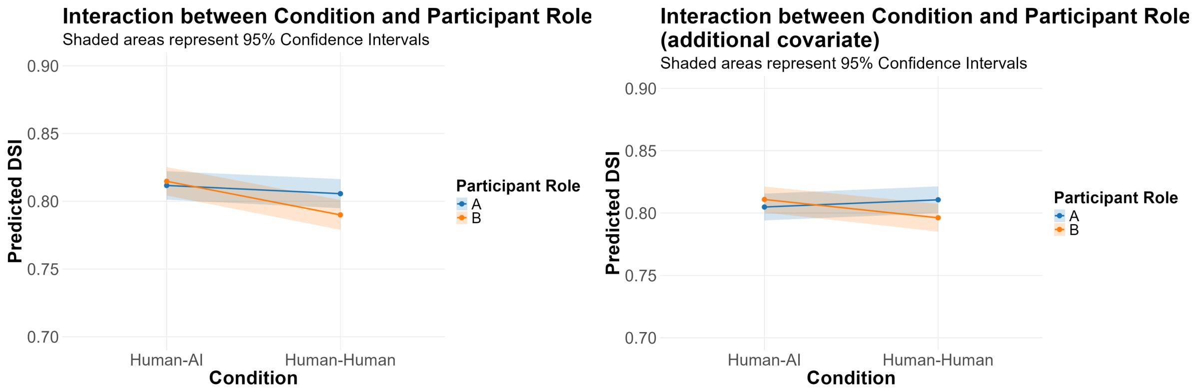


Creative Short Story Task.

Originality. We tested hypothesis 1 on whether human or AI collaborators different in originality performance on the creative short story task. Following the model indicated in the pre-registration, a linear mixed-effects model was conducted to examine the effects of condition and participant role on prediction accuracy, controlling for scaled DAT and accounting for random effects of group ID. The initial model revealed a significant intercept (B = 0.812, SE = 0.005, t = 151.86, p < .001), indicating high overall prediction accuracy. No significant main effects were found for condition (B = -0.006, SE = 0.008, t = -0.79, p = .431) or participant role (B = 0.003, SE = 0.007, t = 0.43, p = .667), and the scaled DAT covariate was not significant (B = 0.001, SE = 0.003, t = 0.26, p = .795). The interaction between condition and participant role approached significance (B = -0.019, SE = 0.010, t = -1.83, p = .070). However, when word count was added to the model, it emerged as a significant predictor (B = 0.0002, SE = 0.00006, t = 4.05, p < .001), with higher word counts associated with increased prediction accuracy. With the inclusion of word count, the interaction between condition and participant role became significant (B = -0.020, SE = 0.010, t = -2.03, p = .045), indicating that the effect of condition on prediction accuracy differed significantly depending on participant role, while main effects for condition (B = 0.006, SE = 0.008, t = 0.72, p = .472) and participant role (B = 0.006, SE = 0.007, t = 0.86, p = .393) remained non-significant (see Fig. 6).

Figure 6.

Interaction between condition and participant role for originality of short stories

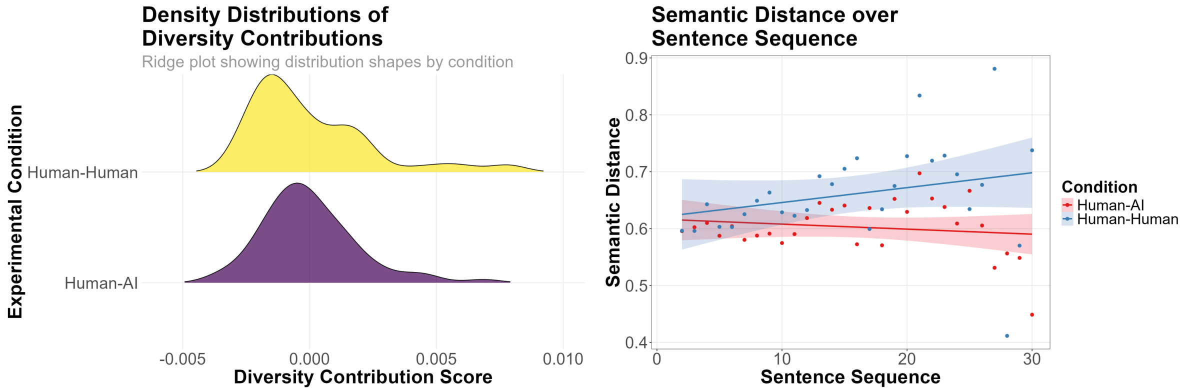


Diversity. We then tested hypothesis two in whether the diversity of stories differed between human and AI collaborators. The Mann-Whitney U test comparing diversity contribution scores between Human-AI (n = 72 pairs) and Human-Human (n = 68 pairs) conditions found no significant difference in central tendencies (U = 2674, p = 0.347). However, the Kolmogorov-Smirnov test examining overall distributional differences yielded a marginally non-significant result (D = 0.217, p = 0.061), suggesting potential differences in distribution shape or spread that warrant further investigation with larger sample sizes (see Fig. 7).

Exploratory forward flow analysis with the sequence of sentences in the creative short story task revealed no main effect of running sequence (B = 0.001, SE = 0.001, t = 1.70, p = .089) and a significant interaction between condition and running sequence (B = 0.003, SE = 0.002, t = 2.08, p = .038). As shown in figure 7, human participants in Human-AI pairs demonstrated relatively stable semantic distance over the sentence sequence, while humans in Human-Human pairs showed increasing semantic distance across sentences. The main effect of condition was not significant (B = 0.005, SE = 0.016, t = 0.28, p = .777). The overall model was significant (R^2 = .014, $F(3, 1762)$ = 8.60, p < .001). These results parallel the AUT idea forward flow analysis, indicating how partner type could shape semantic progression in both constrained (AUT) and open-ended (story-telling) creativity tasks.

Figure 7.

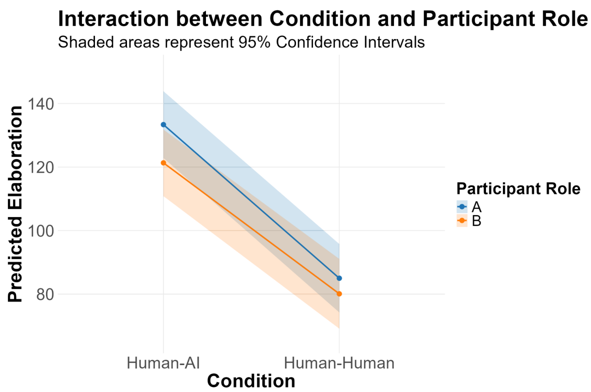
Comparing diversity of short stories between condition



Elaboration. Hypothesis 3 asks if elaboration is different with a human or AI collaborator. As indicated in the pre-registration, a separate linear mixed-effects model was conducted to examine elaboration (word count), operationalized as story word count, with condition and participant role as predictors, controlling for scaled DAT and accounting for random effects of group ID. The model revealed a significant intercept (B = 133.36, SE = 5.34, t = 24.97, p < .001), indicating substantial baseline word count. A significant main effect of the condition emerged (B = -48.38, SE = 7.68, t = -6.30, p < .001), with human-human interactions producing significantly fewer words compared to the AI condition. Additionally, there was a significant main effect of participant role (*B* = -12.04, *SE* = 5.59, *t* = -2.15, *p* = .033), indicating that role B participants used fewer words than role A participants. The scaled DAT covariate was also significant (*B* = 8.44, *SE* = 2.61, *t* = 3.24, *p* = .001), suggesting that word count increased for participants with higher verbal creativity. However, the interaction between condition and participant role was not significant (B = 7.14, *SE* = 8.05, *t* = 0.89, *p* = .377), indicating that the effect of condition on elaboration did not differ significantly between participant roles.

Figure 8.

Interaction between condition and participant role for elaboration of short stories



4. Discussion

The present study contributes to our understanding of collaborative creativity by providing evidence for the advantages and limitations of human-human versus human-AI collaboration. Three key hypotheses were examined: 1) Human-AI collaboration produces more original responses, 2) Human-Human collaboration yields higher diversity, and 3) Human-AI collaboration produces more elaborate outputs. For hypothesis 1, a consistent finding is that human-AI collaboration appears beneficial to originality, until we control for verbosity, then humans had worse performance. For hypothesis 2, human-human pairs showed greater semantic diversity over time, and whether a human or AI collaborator starts first impacted the originality of the creative short stories and their elaboration. Finally for hypothesis 3, human respondents in the AI collaboration wrote longer responses than participants in the human collaboration. Collectively, our results point to a potential tradeoff: AI collaboration may increase output volume but reduce originality and diversity compared to human partnership.

Hypothesis 1: Originality

AUT. Supporting our first hypothesis, human-AI collaboration seemed to generate more original ideas than human-human collaboration, but this benefit disappeared – and even reversed – once the verbosity of the ideas was factored in. Initial analyses with AUT responses suggest AI collaboration enhanced originality, consistent with findings that AI systems can generate more novel ideas than humans (e.g., Hubert et al., 2024), and previous research indicating that AI collaboration increases the originality of creative products (Hubert et al., 2024; B. C. Lee & Chung, 2024; Luchini et al., 2025; Urban et al., 2024). However, this apparent advantage disappeared when controlling for response length, fluency, and item-level effects. The initial "AI advantage" appears to be an artifact of AI's tendency to produce longer, more elaborate responses. This is aligned with literature highlighting AI's ‘elaboration bias’ when scored by another AI (Tang et al., 2024), where longer responses are scored more favorably and higher on originality.

Indeed, when excluding AI contributions, human-human collaboration was shown to produce more original responses than humans in the human-AI collaboration. One possibility is that participants in the human-AI collaboration engaged in cognitive offloading – a reliance on external systems to reduce mental effort and engagement. Generating creative responses is inherently effortful, and humans, as cognitive misers, often seek to minimize cognitive load when possible (Barr et al., 2015). Individuals across all working memory capabilities engage in cognitive offloading as a strategy (Morrison & Richmond, 2020). When AI provides readily available suggestions, humans may default to less effortful engagement in the task. This aligns with recent work indicating that frequent genAI usage is associated with weaker critical thinking abilities, worse learning outcomes, and even de-skilling (Budzyń et al., 2025; Gerlich, 2025; Kosmyna et al., 2025; H.-P. (Hank) Lee et al., 2025; Oakley et al., 2025; Skulmowski, 2023) and reduced recall of specific information with enhanced memory of where to access them instead (Sparrow et al., 2011). A recent Harvard Business Review coined the term 'workslop', when individuals use genAI to offload cognitive work to colleagues downstream, leading to a cascade of effortful decoding, inference, and complex decision-making (Niederhoffer et al., 2025). Creativity requires engaging in both associative and executive processes that are slow and effortful, and participants acknowledge that using AI makes tasks feel easier (Tang et al., 2024).

Creative short stories. In creative short stories, no difference was found in originality, but starting first was associated with being more creative (role A) only in the human-human collaboration condition. The different findings between AUT and creative short stories likely reflect differences in their cognitive demands. The AUT requires generating multiple discrete ideas that can be evaluated independently for novelty (Guilford, 1967). In contrast, creative storytelling, like discourse production, demands some level of narrative coherence and thematic unity, requiring ideas to work together as an integrated whole rather than as separate creative outputs (Kintz et al., 2016; Marini et al., 2005; Patel et al., 2024).

Human collaborators will relinquish creative agency to other humans, but not to AI. This is evidenced by a study where Role B participants produced less original content only when partnered with a human, effectively taking a "backseat," while their output remained stable when partnered with an AI. This divergence is likely due to perceived social differences that only activate during human-to-human collaboration, as AI is not treated as a true social peer. Neuroscientific evidence shows reduced activation in mentalizing networks (Chaminade et al., 2012; Montag et al., 2023), relevant for social-cognitive processes, when dealing with AI, indicating they're not processed as social partners. Behavioral data confirms lower engagement and physiological arousal compared to human-to-human interactions (Lim & Reeves, 2010). Unlike the AUT where an individual's responses do not necessitate building upon or interacting with each other, a collaborative creativity writing task requires one to take the lead in guiding the narrative direction. Furthermore, humans resist the AI taking a leading role in turn-taking tasks (Winston & Magerko, 2017) and base their trust and willingness to collaborate on their perception of the AI's agency (Sundar & Liao, 2023). These findings underscore that in creative collaborative tasks, factors like the partner's perceived social status and the order of creative initiation significantly influence human willingness to contribute or defer.

Hypothesis 2: Diversity

Our second hypothesis that human collaboration will be more diverse was supported. While we observed a wider range of originality scores for human-AI pairs in the AUT at first, this was, once again, driven by the presence of AI's responses, as the effect disappeared when excluding AI responses from the analysis. In both the AUT and creative short story, forward flow analysis showed increasing semantic jumps in human collaboration. A potential explanation is that the current AI systems may lack the capacity to elicit the type of conceptual leaps that leads to progressively divergent idea evolution over time. A recent study supports this finding in that AI usage trends towards homogeneity over time, calling it a 'creative scar inked in the temporal creativity trajectory' (Zhou et al., 2026). This may be related to the AI's tendency towards excessive agreement with its users, which is also known as 'sycophancy' (Malmqvist, 2024)

When a collective story is written by multiple individuals, questions arise about equivalence of contribution, engagement, and leadership. However, systematic reviews on generative AI show a consistent theme: human agency and ownership are still central to directing any creative work (Heigl, 2025). In collaborative storytelling contexts such as with role-playing games, players have been found to show higher verbal creativity (Chung, 2013), suggesting that routine structured and engaged collaborative creativity can enhance individual creative capacities. In such collaborative story-telling, individuals may pick the elements that other players have put forward and develop them further, or choose to ignore certain parts of them. On the other hand, LLMs function more as a monolithic, non-curating entity that does not have creative intent. For a narrative to be cohesive and compelling, the author must propel the story forward in a series of logical steps of cause and effect. Because LLMs operate through probability-based generation and lack creative intentionality or understanding outside their training data (Bender & Koller, 2020; Chamberlain et al., 2017; Felin & Holweg, 2024; Hicks et al., 2024; McCormack et al., 2019), for the story to be cohesive, humans would need to act as the essential curators, editors, and intentional directors of the narrative, providing the cohesive thread of cause and effect and novel creative direction.

Hypothesis 3: Elaboration

Lastly, our third hypothesis received strong support for both AUT and creative short stories: AI as a creative partner greatly increased the length of human responses, even when analyzing only human contributions. This effect was particularly pronounced for Role B in the AUT and Role A in the creative stories. This pattern may be explained through the anchoring effect, a well-established cognitive bias where initial information serves as a reference point that influences subsequent judgments and behaviors (Furnham & Boo, 2011; Tversky & Kahneman, 1974). In the context of creative collaboration, AI responses may have provided extensive elaboration that served as response length anchors, prompting human participants to match or exceed this standard of length. At first glance, this seems to contradict humans' reluctance to cede narrative control, but the two stem from different psychological mechanisms. The anchoring effect boosts output quantity/elaboration, not control, with participants often mistaking length for quality.

The study's findings yield three principal practical implications. First, in contexts aimed at fostering creativity and learning, the use of AI requires careful consideration, as human-AI collaboration can promote cognitive offloading, thereby diminishing original cognitive engagement and potentially undermining learning outcomes despite enhancing output elaboration. Second, the evidence affirms the enduring value of human-human collaboration for creative endeavors, as it sustains greater semantic diversity and mitigates the trend toward creative homogeneity observed with AI over time. Finally, and crucially, these insights demand that organizational leaders eschew a blanket "AI everywhere" mandate and instead formulate explicit, strategic policies for its use; relegating such decisions to individual employees without guidance reflects a lack of discernment and risks the indiscriminate application of AI to unsuitable tasks, ultimately privileging output volume at the expense of true originality and creative diversity.

Limitations and Future Directions

Several limitations of the present study warrant consideration and suggest important avenues for future research. First, our findings are limited to specific creative tasks (AUT and creative storytelling) and may not generalize to other forms of collaborative creativity such as problem-solving, design thinking, or artistic creation. Future studies should examine these dynamics across a broader range of creative domains to establish the generalizability of our findings. Second, while we observed different effects for human-AI and human-human partnerships, our study did not directly examine the underlying cognitive mechanisms that may explain these differences. Future research should investigate the role of shared mental models – cognitive representations of task requirements, partner capabilities, and collaborative processes – in mediating the effectiveness of different collaboration types

(Cannon-Bowers et al., 1993; Mathieu et al., 2000). Understanding how humans develop and maintain shared mental models with AI partners compared to human partners could provide crucial insights into the mechanisms underlying our observed effects. Third, our study focused on immediate creative outputs rather than longitudinal creative development, leaving questions about how these collaboration effects evolve over extended periods. Future longitudinal studies could examine whether the cognitive offloading effects we observed in human-AI collaborations persist, adapt, or intensify with prolonged interaction, and whether shared mental models become more sophisticated or remain fundamentally limited in human-AI contexts.

5. Conclusion

This study reveals that human-AI collaborative creativity produces fundamentally different outcomes compared to human-human collaborations. Our findings demonstrate that AI collaborations do not guarantee enhanced originality; instead, controlling for verbosity results in worse originality in humans. This suggests cognitive offloading in human participants, where individuals rely on AI-generated content to reduce their own mental effort and creative engagement. While humans produce longer responses when working with AI partners, this may be due to anchoring effects from AI's elaborate outputs. In contrast, human-human collaborations promote continuous exploration of semantic space over time and foster hierarchical relationship dynamics that encourage sustained cognitive engagement. Notably, humans appear unwilling to yield agency in collaborative story-telling with AI, whereas in human collaborations, they more readily let the starting partner take the lead. These findings suggest that the apparent benefits of AI collaboration may come at the cost of reduced human creative engagement, highlighting the importance of designing AI tools that promote active cognitive engagement rather than passive reliance. As AI systems become increasingly prevalent in creative workflows, understanding these cognitive offloading effects will be crucial for preserving human expertise.

Funding:
This project was funded by the APA Division 10 microgrant. R.E.B. is supported by grants from Amazon and the National Science Foundation [DRL-240078; DUE-2155070].

Declaration of competing interest:
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements:
We would like to acknowledge the assistance of Jimmy Pronchick, lab manager at the CNCL, for the human-AI platform.

References

Acar, S., Runco, M. A., & Park, H. (2020). What should people be told when they take a divergent thinking test? A meta-analytic review of explicit instructions for divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 39–49. https://doi.org/10.1037/aca0000256

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M. K., Irshad, M., Arraño-Muñoz, M., & Ariza-Montes, A. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, *10*(1), 311. https://doi.org/10.1057/s41599-023-01787-8

Amabile, T. M. (2020). Creativity, Artificial Intelligence, and a World of Surprises. *Academy of Management Discoveries*, *6*(3), 351–354. https://doi.org/10.5465/amd.2019.0075

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Astutik, S., Susantini, E., Madlazim, Nur, M., & Supeno. (2020). The Effectiveness of Collaborative Creativity Learning Models (CCL) on Secondary Schools Scientific Creativity Skills. *International Journal of Instruction*, *13*(3), 525–538.

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior*, *48*, 473–480. https://doi.org/10.1016/j.chb.2015.02.029

Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., Bengio, Y., & Jerbi, K. (2024). *Divergent Creativity in Humans and Large Language Models* (No. arXiv:2405.13012). arXiv. https://doi.org/10.48550/arXiv.2405.13012

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault

(Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., … Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, *0*(0). https://doi.org/10.1016/S2468-1253(25)00133-5

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In *Individual and group decision making: Current issues* (pp. 221–246). Lawrence Erlbaum Associates, Inc.

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2017). Putting the Art in Artificial: Aesthetic Responses to Computer-Generated Art. *Psychology of Aesthetics, Creativity, and the Arts*, *12*. https://doi.org/10.1037/aca0000136

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00103

Choi, J. H., & Schwarcz, D. (2023, August 13). *AI Assistance in Legal Analysis: An Empirical Study* [SSRN scholarly paper]. https://doi.org/10.2139/ssrn.4539836

Chung, T. (2013). Table-top role playing game and creativity. *Thinking Skills and Creativity*, *8*, 56–71. https://doi.org/10.1016/j.tsc.2012.06.002

Cropley, D. (2023). Is artificial intelligence more creative than humans? : ChatGPT and the Divergent Association Task. *Learning Letters*, *2*, 13–13. https://doi.org/10.59453/ll.v2.13

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. https://doi.org/10.1126/sciadv.adn5290

Epstein, Z., Hertzmann, A., & THE INVESTIGATORS OF HUMAN CREATIVITY. (2023). Art and the science of generative AI. *Science (New York, N.Y.)*, *380*(6650), 1110–1111. https://doi.org/10.1126/science.adh4451

Faiella, A., Zielińska, A., Karwowski, M., & Corazza, G. E. (2025). Am I Still Creative? The Effect of Artificial Intelligence on Creative Self-Beliefs. *The Journal of Creative Behavior*, *59*(2), e70011. https://doi.org/10.1002/jocb.70011

Felin, T., & Holweg, M. (2024). Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, *9*(4), 346–371. https://doi.org/10.1287/stsc.2024.0189

Fink, A., Koschutnig, K., Benedek, M., Reishofer, G., Ischebeck, A., Weiss, E. M., & Ebner, F. (2012). Stimulating creativity via the exposure to other people's ideas. *Human Brain Mapping*, *33*(11), 2603–2610. https://doi.org/10.1002/hbm.21387

Fowler, Z., Palombo, D. J., Madan, C. R., & O'Connor, B. B. (2024). Collaborative imagination synchronizes representations of the future and fosters social connection in the present. *Proceedings of the National Academy of Sciences*, *121*(25), e2318292121. https://doi.org/10.1073/pnas.2318292121

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, *15*(1), Article 1. https://doi.org/10.3390/soc15010006

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Haase, J., Hanel, P. H. P., & Pokutta, S. (2025). *S-DAT: A Multilingual, GenAI-Driven Framework for Automated Divergent Thinking Assessment* (No. arXiv:2505.09068). arXiv. https://doi.org/10.48550/arXiv.2505.09068

Heigl, R. (2025). Generative artificial intelligence in creative contexts: A systematic review and future research agenda. *Management Review Quarterly*. https://doi.org/10.1007/s11301-025-00494-9

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hsu, J., & Huang, D. (2010). *Correlation between impact and collaboration*. https://doi.org/10.1007/s11192-010-0265-x

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), 3440. https://doi.org/10.1038/s41598-024-53303-w

John, O. P., & Srivastava, S. (1999). *The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*.

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., van Hell, J., Kennedy, E., Sullivan, G. F., Taylor, C. L., Ward, T., & Beaty, R. E. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01986-2

Kintz, S., Fergadiotis, G., & Wright, H. H. (2016). Aging effects on discourse production. In *Cognition, language and aging* (pp. 81–106). John Benjamins Publishing Company. https://doi.org/10.1075/z.200

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (No. arXiv:2506.08872). arXiv. https://doi.org/10.48550/arXiv.2506.08872

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, *8*(10), 1906–1914. https://doi.org/10.1038/s41562-024-01953-1

Lee, H.-P. (Hank), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025, April 1). *The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers*. CHI 2025. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/

Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies*, *68*(1–2), 57–68. https://doi.org/10.1016/j.ijhcs.2009.09.008

Luchini, S., Pronchick, J., Ceh, S., Kaufman, J., Johnson, D., Rafner, J., & Beaty, R. (2025). *The Roles of Idea Generation and Elaboration in Human-AI Collaborative Creativity*. https://doi.org/10.31234/osf.io/xm2f5_v2

Malmqvist, L. (2024). *Sycophancy in Large Language Models: Causes and Mitigations* (No. arXiv:2411.15287). arXiv. https://doi.org/10.48550/arXiv.2411.15287

Marini, A., Boewe, A., Caltagirone, C., & Carlomagno, S. (2005). Age-related Differences in the Production of Textual Descriptions. *Journal of Psycholinguistic Research*, *34*(5), 439–463. https://doi.org/10.1007/s10936-005-6203-z

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *The Journal of Applied Psychology*, *85*(2), 273–283. https://doi.org/10.1037/0021-9010.85.2.273

McCormack, J., Gifford, T., & Hutchings, P. (2019). Autonomy, Authenticity, Authorship and Intention in Computer Generated Art. In A. Ekárt, A. Liapis, & M. L. Castro Pena (Eds.), *Computational Intelligence in Music, Sound, Art and Design* (pp. 35–50). Springer International Publishing. https://doi.org/10.1007/978-3-030-16667-0_3

Molnar, A. (2019). How to implement real-time interaction between participants in online surveys: A practical guide to SMARTRIQS. *The Quantitative Methods for Psychology*, *16*(4), 334–354.

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B. W., & Becker, B. (2023). Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging. *Personality Neuroscience*, *6*, e3. https://doi.org/10.1017/pen.2022.5

Niederhoffer, K., Kellerman, G. R., Lee, A., Liebscher, A., Rapuano, K., & Hancock, J. T. (2025, September 22). AI-Generated “Workslop” Is Destroying Productivity. *Harvard Business Review*. https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity, https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity

Nijstad, B. A., & Stroebe, W. (2006). How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups. *Personality and Social Psychology Review*, *10*(3), 186–213. https://doi.org/10.1207/s15327957pspr1003_1

Oakley, B., Johnston, M., Chen, K.-Z., Jung, E., & Sejnowski, T. J. (2025). *The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI* (No. arXiv:2506.11015). arXiv. https://doi.org/10.48550/arXiv.2506.11015

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25), e2022340118. https://doi.org/10.1073/pnas.2022340118

Ou, X., Goldschmidt, G., & Erez, M. (2023). The effect of disciplinary diversity on design idea generation in dyadic teams. *Design Studies*, *86*, 101184. https://doi.org/10.1016/j.destud.2023.101184

Patel, T., Macpherson, S. E., & Hoffman, P. (2024). *Balancing Act: A Neural Trade-off Between Coherence and Creativity in Spontaneous Speech* (p. 2024.12.23.630053). bioRxiv. https://doi.org/10.1101/2024.12.23.630053

Patterson, J., Pronchick, J., Panachanadikar, R., Fuge, M., van Hell, J., Miller, S., Johnson, D., & Beaty, R. (2025). *CAP: The Creativity Assessment Platform for Online Testing and Automated Scoring*. OSF. https://doi.org/10.31234/osf.io/zeyh3_v2

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. https://doi.org/10.1038/s41562-023-01751-1

Rammstedt, B., & John, O. P. (2012). *Big Five Inventory-10* [Dataset]. https://doi.org/10.1037/t01744-000

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Shiiku, S., Marjieh, R., Anglada-Tort, M., & Jacoby, N. (2025). *The Dynamics of Collective Creativity in Human-AI Social Networks* (No. arXiv:2502.17962). arXiv. https://doi.org/10.48550/arXiv.2502.17962

Skulmowski, A. (2023). The Cognitive Architecture of Digital Externalization. *Educational Psychology Review*, *35*(4), 101. https://doi.org/10.1007/s10648-023-09818-1

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science (New York, N.Y.)*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Sundar, S. S. (2020). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Sundar, S. S., & Liao, M. (2023). Calling BS on ChatGPT: Reflections on AI as a Communication Source. *Journalism & Communication Monographs*, *25*(2), 165–180. https://doi.org/10.1177/15226379231167135

Tadmor, C. T., Satterstrom, P., Jang, S., & Polzer, J. T. (2012). Beyond Individual Creativity: The Superadditive Benefits of Multicultural Experience for Collective Creativity in Culturally Diverse Teams. *Journal of Cross-Cultural Psychology*, *43*(3), 384–392. https://doi.org/10.1177/0022022111435259

Tang, M., Hofreiter, S., Werner, C. H., Zielińska, A., & Karwowski, M. (2024). “Who” Is the Best Creative Thinking Partner? An Experimental Investigation of Human–Human, Human–Internet, and Human–AI Co-Creation. *The Journal of Creative Behavior*, *n/a*(n/a). https://doi.org/10.1002/jocb.1519

Toubia, O. (2006). Idea Generation, Creativity, and Incentives. *Marketing Science*, *25*, 411–425. https://doi.org/10.1287/mksc.1050.0166

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, 105031. https://doi.org/10.1016/j.compedu.2024.105031

Winston, L., & Magerko, B. (2017). Turn-Taking with Improvisational Co-Creative Agents. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *13*(1), Article 1. https://doi.org/10.1609/aiide.v13i1.12931

Xue, H., Lu, K., & Hao, N. (2018). Cooperation makes two less-creative individuals turn into a highly-creative pair. *NeuroImage*, *172*, 527–537. https://doi.org/10.1016/j.neuroimage.2018.02.007

Zhou, Y., Liu, Q., Huang, J., & Li, G. (2026). Creative scar without generative AI: Individual creativity fails to sustain while homogeneity keeps climbing. *Technology in Society*, *84*, 103087. https://doi.org/10.1016/j.techsoc.2025.103087