
Incentives Shape How Humans Co-Create with Generative AI

Nathanael Jo*

MIT

Cambridge, MA 02139

nathanjo@mit.edu

Manish Raghavan

MIT

Cambridge, MA 02139

mragh@mit.edu

Abstract

1 Generative AI is quickly becoming an integral part of people’s everyday workflows.
2 Early evidence has shown that while generative AI can increase individual-level
3 productivity, it does so at the cost of collective diversity, potentially narrowing the
4 set of ideas and perspectives produced.
5 Our research stands in contrast to this concern: through a pre-registered randomized
6 control trial, we show that *incentives* mediate AI’s homogenizing force in a creative
7 writing task. Participants rewarded for originality produce collectively more diverse
8 writing than those rewarded for quality alone. This divergence is driven not by
9 abandoning AI, but by how participants use it: those incentivized for originality
10 incorporate fewer AI suggestions verbatim, relying on the model more selectively
11 for brainstorming, proofreading, and targeted edits. Our results reveal that the
12 effects of generative AI depend not only on the technology itself, but also the
13 behavioral strategies and incentive structures surrounding its use.

14 1 Introduction

15 As generative AI becomes woven into creative workflows, a troubling pattern has emerged: AI
16 assistance appears to homogenize outputs, narrowing the diversity of ideas and expression even when
17 humans retain final control [31, 29, 2, 5, 17, 1]. This has led to concerns that widespread AI adoption
18 will compress the creative landscape. We argue that this conclusion is premature because existing
19 findings have two shortcomings that limit the extent to which we can extrapolate to the real world.

20 First, **incentives matter**. In controlled experiments, participants without proper incentives may
21 simply accept the first AI-generated suggestion because they want to complete the task quickly. This
22 is not necessarily a good proxy for real-world settings. For example, even if prospective students use
23 AI to help write their college admissions essays, we might expect them to invest substantial effort
24 to stand out within a highly competitive applicant pool. With stronger incentives for differentiation,
25 human judgment might mitigate or overcome the homogenizing force of AI [19].

26 Second, much of the evidence on homogenization comes from experimental setups with limited
27 **interactivity** between humans and AI. Most studies restrict participants to a single mode of use:
28 brainstorming [2, 5] or text completion [17, 1]. But real-world AI use is open-ended and iterative.
29 Studying only one modality risks overstating the homogenizing effect, since it prevents users from
30 exercising the kind of sustained agency that characterizes genuine **co-creation**.

31 **The present work: The impact of incentives on co-creation.** Our work addresses these shortcom-
32 ings by embedding a richer model of co-creation within a randomized trial ($n = 200$). Participants
33 write a 250-350 word short story with or without access to AI (GPT5-Mini), interacting freely with a
34 general-purpose assistant however they see fit. This design is substantially more interactive than prior
35 work, and yields rich longitudinal data on participants’ strategies and writing trajectories throughout

*Corresponding author.

the 25-minute session. We also vary *incentives* across participants: half are rewarded for high-quality work, and half for originality. While neither condition exactly matches real-world conditions, the *contrast* between them offers a directionally informative window into how market forces mediate human-AI co-creation. The design and analysis plan were pre-registered on OSF.²

Our Results. We find that while AI suggestions initially make outputs more similar, diversity is mostly restored as people further revise and edit the model’s suggestions. Crucially, incentives play a key role; when participants are rewarded for originality, they adjust their behaviors (via adopting the AI suggestions less), yielding greater diversity. These findings reveal that human-AI interaction is a strategic, adaptive process, where outcomes reflect not just model capabilities, but how users allocate effort and trust in response to the task. Designing AI systems—and incentives around them—thus plays a central role in determining whether co-created outputs reflect the values we care about.

2 Experiment Design and Data

Task and Recruitment. Participants were instructed to write a short story of 250–350 words using a browser-based rich-text editor. After revealing the task and instructions (Fig 5 and 6), participants were given at most 5 minutes to brainstorm and outline their short story (Fig 7). This stage gave participants time to scaffold their ideas so that they could begin writing with a concrete plan, which was important for enabling them to complete their stories within the limited time available. After the brainstorming phase, they had at most 20 minutes to write their story (see Fig 8). We recruited $n = 200$ participants from Prolific, with 50 participants per group. Participants received \$5 for completing the study, which took on average 25 minutes.

Treatment Groups. We run a pre-registered, between-subjects experiment in which participants are assigned via block randomization to one of four conditions in a 2×2 factorial design.

Factor 1: AI Usage. In the **SELF** conditions, participants wrote entirely on their own. In the **AI** conditions, participants had access to an integrated writing assistant designed to mimic general-purpose LLMs (see Figure 8). Note that we did not provide the AI tool during the brainstorming phase because there exist numerous studies investigating the impact of AI on creative ideation [5, 13, 16].

Factor 2: Task Incentives. We also randomized participants into a condition incentivizing **originality (O)**, and another rewarding **technical quality (T)**. In **O**, participants were told that their bonus is a function of originality and uniqueness, while in **T**, participants’ bonuses were based on a grade score accounting for organization and technique. See Figure 6 for details on the grading and bonus structure, and Appendix B.5 for details on the payment of bonuses.

Datasets and Preprocessing. We took snapshots of the text editor every 5 seconds, resulting in a rich temporal dataset of each participant’s progress throughout the task. In addition to the temporal dataset, we also collected information about participants’ characteristics, such as their ethnicity, age, sex, education level, and experience with AI tools. Appendix Table 3 shows a balance table of participant demographics for each group.

AI Usage. Those with access to an AI tool ($n = 100$) on average took 3.57 conversation turns. Many of them were **editing** requests (38%; condensing, restructuring, partial edits), then **drafting** (30%; asking for full stories), **creative consulting** (14%; asking opinions), **mechanical assistance** (3%; spelling or grammar checks), and others (15%). The vast majority of participants (91%) asked for a full draft at least once. This does not suggest that they will use these drafts (we quantify adoption later in Section 5), but they at least observed what the AI suggested. Given the importance of AI to draft, we flagged all AI suggestions that resulted in a valid draft, which we later use in our analyses.

3 Diversity of Stories

A central question we ask is how stories become more or less diverse when participants are given access to an AI tool. All of the analyses in this section were pre-registered.

3.1 Methods

Diversity/Similarity Metrics. Diversity of natural language text is hard to measure since language is incredibly high-dimensional. Our view is that diversity in natural language is a latent construct;

²<https://tinyurl.com/4mn7av7j>

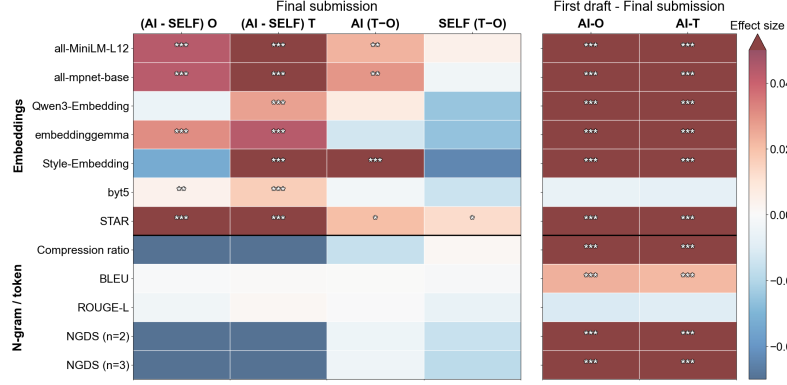


Figure 1: Effect sizes from difference in means tests (Welch t-test, one-sided) with significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, over the final submissions (left) and between the final submission and the first valid draft the AI suggested (right). A higher value means that the first group in the difference-in-means test is more homogeneous. All rows are similarity metrics except NGDS, which is a diversity metric. As such, we flip the comparison for NGDS so that the effect sizes are directionally consistent. See Table 5 for specific values and multiple hypothesis correction.

85 instead of taking a particular view of diversity, we pre-specify a suite of diversity metrics, each
 86 providing correlated aspects of text diversity, see Table 1.

Name	Description
Embeddings (using cosine similarity)	
all-MiniLM-L12-v2 ^a	One of the most popular lightweight embedding models
all-mpnet-base-v2 ^b	One of the most popular lightweight embedding models
Qwen3-Embedding-0.6B [35]	Lightweight (<1B) but high-performing on embedding benchmarks ^c
embeddinggemma-300m [26]	Lightweight (<1B) but high-performing on embedding benchmarks ^c
Style-Embedding [28]	Intended to capture style more than content (for author attribution)
STAR [10]	Intended to capture style in social media text
byt5 [34]	Breaks into byte-level instead of token-level data (might capture style)
N-gram/token based	
Compression ratio [22]	Size of original file $\div$ size of compressed file (e.g., via gZip)
BLEU [18]	Geometric mean of n -gram precision between two texts
Rouge-L [15]	Harmonic mean of longest common subsequence between two texts
n -gram diversity score* [14]	$\sum_{n=1}^N (\# \text{ unique } n\text{-grams} \div \# \text{ all } n\text{-grams})$

^a<https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>

^b<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

^c<https://huggingface.co/spaces/mteb/leaderboard>

Table 1: Table of similarity metrics used, in that higher similarity means reduced diversity. * n -gram diversity score is the only diversity measure, where a higher score means more diversity.

87 For each participant i in cell $g \in \{\text{AI}, \text{SELF}\} \times \{\text{O}, \text{T}\}$, let X_i be their text. Let $S(X_i, X_j) \in \mathbb{R}_+$
 88 be a similarity function. Define the participant-level outcome as the leave-one-out average similarity
 89 within the same cell g : $Y_i^{(g)} := \frac{1}{n_g - 1} \sum_{j \in g, j \neq i} S(X_i, X_j)$.

90 **Hypothesis Tests.** We construct two hypotheses: **(H1)** Less diversity under AI, and **(H2)** More
 91 diversity when originality is rewarded. From the data, we get an estimate of diversity by taking
 92 $\bar{Y}^{(g)} = \sum_{i \in g} Y_i^{(g)}$. We then conduct a two-sample difference in means test for each metric in Table 1
 93 using Welch’s t -test. Since these metrics are statistically correlated reflections of the same hypothesis,
 94 we do not adjust for multiple hypothesis testing. However, even with Benjamini-Hochberg adjustment
 95 for multiple hypothesis testing [3], the results remain consistently significant.

96 3.2 Results

97 Figure 1 (left) reports the results of the hypothesis tests over participants’ final submissions. In
 98 Figure 1 (right), we also display the difference in means between the first full draft that the AI
 99 produces versus the final submission. This difference gives a glimpse on how participants change and
 100 edit substantially before being satisfied with their endproduct.

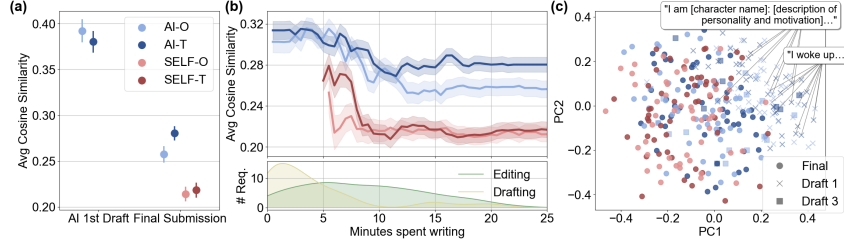


Figure 2: (a) Average cosine similarity of embeddings across randomized groups, for both the final submission and the first full draft the AI produces (when available). (b) [Top] Average cosine similarity of embeddings across randomized groups through time spent in the session. We only include text that has at least 200 words for every timestep to ensure that the embeddings capture enough information about the story. [Bottom] Number of requests over time spent in the session. (c) First and second principal component of the embeddings across groups, for both the drafts that AI produces (when available) and the final submissions. All plots use the embedding model all-MiniLM-L12-v2. See Appendix D.3 for other metrics as robustness checks.

101 **On n -gram/token based measures.** We note that n -gram-based similarity metrics start with having
 102 large effects at the first-draft stage, but virtually disappears at final submission. In contrast, embedding-
 103 based measures remain statistically significant. This pattern suggests that participants manage to
 104 diversify phrasing when they edit, but deeper narrative structure and style remain more homogeneous
 105 when AI assistance is involved. We are generally more interested in characterizing higher-order
 106 features of style and narrative, and thus restrict our attention to the embedding-based metrics.

107 **Hypothesis 1.** We reject the null hypothesis in both incentive conditions (O and T). Across most
 108 embedding-based metrics, AI-assisted outputs are significantly more similar than human-only outputs
 109 at final submission (two leftmost columns). We also find that the first drafts provided by AI are
 110 much more similar than the AI-assisted final submissions (right subfigure). This result indicates that
 111 participants often do not like the first AI draft and exert significant effort to express their own voice,
 112 but there remains a lingering homogenizing effort at final submission.

113 **Hypothesis 2.** In the **AI condition**, we reject the null (third column from left). This indicates that
 114 the incentive to be original was strong enough to induce those with AI access to make attempts to
 115 differentiate themselves. In Section 5, we will unpack the strategies participants take to achieve
 116 diversity. In the **SELF condition**, we fail to reject the null. One interpretation is that, absent a
 117 homogenizing tool, participants already write in their own styles, and the distinction between “quality”
 118 and “originality” incentives may be difficult to operationalize in practice.

119 4 Characterizing Writing Trajectories

120 In the previous section, we provided an initial look at how stories evolve from the first AI draft to the
 121 final submission. We now examine these trajectories in greater depth, asking *when* and *in what ways*
 122 stories shift throughout the session. From this point onward, our analyses are exploratory.

123 4.1 Methods

124 **Diversity through time.** We apply the same similarity metrics from Table 1, now evaluated at fixed
 125 time intervals. Taking $t = 0$ as the start of the writing session, we use snapshots of each participant’s
 126 text at 30-second intervals until the session concludes. If a participant finishes before some time
 127 $t = T$, we use their final submission for all subsequent time points. To ensure that embeddings
 128 meaningfully capture story content, we restrict attention to text containing at least 200 words.

129 **Qualitative structure of stories.** We also ask whether different groups produce qualitatively distinct
 130 types of stories. For example, even if AI-assisted stories are less diverse, they may still occupy the
 131 same overall semantic space as **SELF** stories—or they may form a distinct cluster. To investigate this,
 132 we embed all stories and perform principal component analysis (PCA) on the embedding vectors.

133 4.2 Results

134 For clarity, we focus on results using the all-MiniLM-L12-v2 embedding model, though the
 135 qualitative patterns hold across alternative similarity measures (see Appendix D.3).

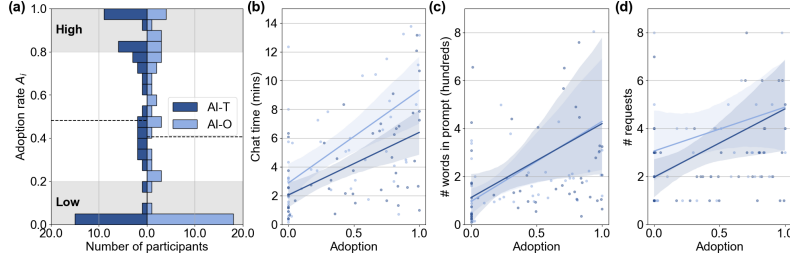


Figure 3: (a) Histogram of adoption score A_i for both the **AI-T** and **AI-O** groups. (b) Time spent writing AI prompts by adoption and group. (c) Number of words in AI prompts by adoption and group. (d) Number of AI requests by adoption and group.

136 **Diversity through time.** Figure 2(a) reports average similarity across conditions for both the first
 137 AI draft and the final submission. These results mirror the findings from the previous section. Most
 138 strikingly, first AI drafts are highly homogeneous, but by the final submission stage, similarity in
 139 **AI-O** drops to 0.26, which is substantially closer to the SELF baseline.

140 Figure 2(b) traces similarity over time. Diversity increases steadily throughout the session, which is
 141 expected because participants express more of their personal style as the session goes on. Importantly,
 142 **AI-O** and **AI-T** begin to diverge around 12 minutes after. This divergence occurs at an inflection
 143 point where enough participants shift from asking for drafting suggestions to actively editing with or
 144 without AI (bottom subfigure). We examine these behavioral strategies in greater detail in Section 5.

145 **Qualitative structure of stories.** Figure 2(c) visualizes the top two principal components of story
 146 embeddings. AI-generated drafts form a distinct cluster, separated from both AI-edited submissions
 147 and human submissions. Many AI drafts share similar structures; for example, starting the story with
 148 “I woke up in...”. This finding indicates that **there exists a recognizable AI voice that is both highly**
 149 **homogeneous and distinct from human writing.** However, across successive drafts, AI outputs
 150 move closer to the distribution of final human submissions, suggesting that repeated interaction
 151 allows users to shape AI outputs away from the default style.

152 5 Differential Strategies in Using AI

153 Apart from characterizing the trajectories of participants’ stories, we ask how do participants’
 154 strategies in using AI differ so that they achieve greater diversity in the **AI-O** condition?

155 5.1 Methods

156 **Measuring adoption of AI suggestions.** To summarize the complex and heterogeneous trajectories
 157 participants take throughout the session, we measure the adoption rate of AI suggestions A . As we
 158 show later, this measure is highly informative of participants’ overall strategies.

159 Our approach is based on lexical attribution. First, we extract n -grams for each final submission,
 160 with $n = 10$. For each n -gram, we check whether it appears in any AI-generated output within the
 161 interaction transcript. If a match is not found, the n -gram is *not AI-attributable*. If a match is found,
 162 we find the earliest AI suggestion in which the n -gram appears and check the corresponding user
 163 prompt. If the n -gram also appears in the corresponding prompt, then the n -gram is *not AI-attributable*
 164 because it was originally authored by the user. Otherwise, it is AI-attributable. See Fig 12 for a
 165 representative illustration of AI attribution. Adoption is then defined as: $A = \frac{\# \text{ AI-attributable } n\text{-grams}}{\# \text{ total } n\text{-grams}}$.
 166 We conduct multiple robustness checks to our adoption metric choices in Appendix C.1.

167 5.2 Results

168 We first examine how participants’ adoption rates differed, and then show how these strategic
 169 differences account for the diversity patterns from the prior section.

170 **Adoption as a driver for diversity.** Figure 3(a) shows the distribution of the adoption measure
 171 A_i for participants in the **AI-O** and **AI-T** conditions. In both groups, adoption is bimodal: most
 172 participants either anchor heavily on AI suggestions or make minimal use of them. Few participants
 173 fall in between these extremes. Relative to **AI-T**, the **AI-O** condition exhibits a clear shift toward

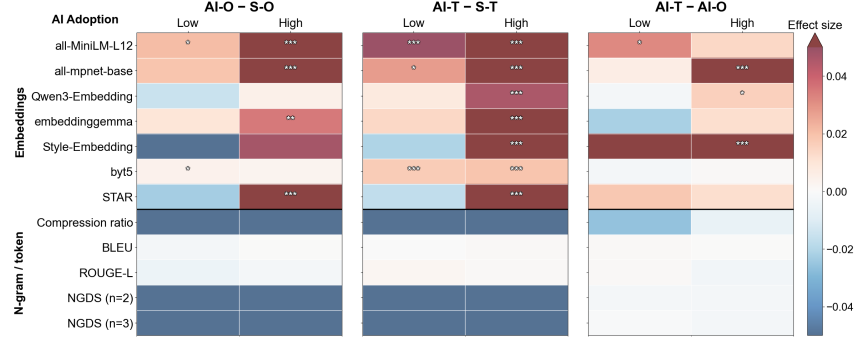


Figure 4: Effect sizes from difference in means tests (Welch t-test, one-sided) with significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, over the final submissions stratified by low and high adoption among the AI-assisted group. See Table 6 for specific effect sizes.

lower adoption. More **AI-O** participants adopt a low-adoption strategy ($R_i < 0.2$), and fewer adopt a high-adoption strategy ($R_i > 0.8$). This pattern suggests that incentives to diversify encourage participants to distance their final submissions from AI-generated text, possibly because they have priors that AI outputs are relatively homogeneous or unlikely to stand out.

This modest shift in adoption explains the gap in diversity between **AI-O** and **AI-T**. In Figure 4, we re-estimate the difference-in-means tests after stratifying AI-assisted participants into low- and high-adoption groups. Among low-adoption participants, final submissions are nearly indistinguishable in diversity from those in the **SELF** condition. In contrast, high-adoption participants produce substantially more similar outputs, yielding large and statistically significant gaps relative to **SELF**.

Low-adoption users still engage with AI. Participants in the low-adoption group still interacted with the AI multiple times, averaging just under three prompts per session (Figure 3(d)) and spending roughly three minutes composing these prompts (Figure 3(b)). These users are more likely to request full drafts than high-adoption users, but make far fewer subsequent attempts to engage with the AI. This pattern suggests that low-adoption participants made genuine attempts to use the AI, but ultimately chose not to incorporate its suggestions—perhaps due to strong preferences over narrative direction or voice, or skepticism that the AI could further improve their story. See Appendix D.1 for additional discussion of participants’ attitudes toward AI usage.

6 Discussion

The role of AI in co-creation. Our findings show that the homogenizing effect of AI is shaped by the conditions under which people use it. When given a general-purpose AI tool, many participants used diverse interaction strategies and heavily revised AI outputs. This differs from settings where AI is limited to text completion [17, 1] or idea generation [2, 5].

At the same time, our results highlight incentives as an important but underexplored variable in studies of human-AI co-creation. Prior work has largely tested AI’s homogenizing effect on participants with weak or no incentives to differentiate—a setting that may systematically overstate homogenization relative to real-world contexts where stakes are high. Researchers studying human-AI collaboration should therefore account for how incentive structures shape the behaviors they observe.

Beyond the lab, our findings suggest that competitive incentives can serve as a practical lever for promoting differentiation. Strong incentives arise naturally in some markets: college admissions, for instance, creates pressure for applicants to stand out, which our results suggest may partly counteract AI’s homogenizing effect. In sum, aligning reward structures with the values we care about may be more effective than restricting AI use outright.

Limitations and Future Work. We note several limitations and avenues for future work. First, our experiment captures behavior within a single-session interaction and therefore cannot speak to how individuals’ relationships with AI evolve over time. An important open question is whether sustained use of AI might eventually help users better articulate and refine their own creative voice. Beyond temporal dynamics, our findings are specific to a particular model of co-creation: a general-purpose chat interface used alongside a text editor. While this reflects a common mode of AI-assisted writing today, it may not represent how AI will be integrated into creative workflows in the long run.

References

- [1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2025.
- [2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425, 2024.
- [3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
- [4] Sudapa Chompunuch and Todd Lubart. Ai as a helper: Leveraging generative ai tools across common parts of the creative process. *Journal of Intelligence*, 13(5):57, 2025.
- [5] Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.
- [6] Jennifer Haase and Sebastian Pokutta. Human-ai co-creativity: Exploring synergies across levels of creative collaboration. *arXiv preprint arXiv:2411.12527*, 2024.
- [7] Joshua Holstein, Patrick Hemmer, Gerhard Satzger, and Wei Sun. When thinking pays off: Incentive alignment for human-ai collaboration. *arXiv preprint arXiv:2511.09612*, 2025.
- [8] Chenyu Hou, Gaoxia Zhu, and Vidya Sudarshan. The role of critical thinking on undergraduates’ reliance behaviours on generative ai in problem-solving. *British Journal of Educational Technology*, 56(5):1919–1941, 2025.
- [9] Hsuan-Che Brad Huang. Unlocking creativity with artificial intelligence (ai): Field and experimental evidence on the goldilocks (curvilinear) effect of human–ai collaboration. *Journal of Experimental Psychology: General*, 2025.
- [10] Javier Huertas-Tato, Alejandro Martín, and David Camacho. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, 296:111867, 2024.
- [11] Simran Kaur, Sara Salimzadeh, and Ujwal Gadiraju. Incentive-tuning: Understanding and designing incentives for empirical human-ai decision-making studies. *arXiv preprint arXiv:2601.15064*, 2026.
- [12] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.
- [13] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025.
- [14] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312, 2023.
- [15] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [16] Kelsey Medeiros, David H Cropley, Rebecca L Marrone, and Roni Reiter-Palmon. Human-ai co-creativity: Does chatgpt make us more creative? *The Journal of Creative Behavior*, 59(2): e70022, 2025.
- [17] Vishakh Padmakumar and He He. Does writing with language models reduce creative diversity? In *ICLR*, 2024.

- [18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [19] Manish Raghavan. Competition and diversity in generative ai. *arXiv preprint arXiv:2412.08610*, 2024.
- [20] Raluca Rilla, Tobias Werner, Hiromu Yakura, Iyad Rahwan, and Anne-Marie Nussberger. Recognising, anticipating, and mitigating llm pollution of online behavioural research. *arXiv preprint arXiv:2508.01390*, 2025.
- [21] Mark A Runco and Selcuk Acar. Divergent thinking as an indicator of creative potential. *Creativity research journal*, 24(1):66–75, 2012.
- [22] Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*, 2024.
- [23] Robert J Sternberg. Do not worry that generative ai may compromise human creativity or intelligence in the future: It already has. *Journal of Intelligence*, 12(7):69, 2024.
- [24] Shuhua Sun, Zhuyi Angelina Li, Maw-Der Foo, Jing Zhou, and Jackson G Lu. How and for whom using generative ai affects creativity: A field experiment. *Journal of Applied Psychology*, 2025.
- [25] E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*, 1966.
- [26] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.
- [27] Feifei Wang, Nuo Li, Alan CK Cheung, and Gary KW Wong. In genai we trust: An investigation of university students’ reliance on and resistance to generative ai in language learning. *International Journal of Educational Technology in Higher Education*, 22(1):59, 2025.
- [28] Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26. URL <https://aclanthology.org/2022.repl4nlp-1.26/>.
- [29] Emily Wenger and Yoed Kenett. We’re different, we’re the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*, 2025.
- [30] Sean J Westwood. The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47):e2518075122, 2025.
- [31] Fan Wu, Emily Black, and Varun Chandrasekaran. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*, 2024.
- [32] Suqing Wu, Yukun Liu, Mengqi Ruan, Siyu Chen, and Xiao-Yun Xie. Human-generative ai collaboration enhances task performance but undermines human’s intrinsic motivation. *Scientific Reports*, 15(1):15105, 2025.
- [33] Zhuohao Wu, Danwen Ji, Kaiwen Yu, Xianxu Zeng, Dingming Wu, and Mohammad Shidujaman. Ai creativity and the human-ai co-creation model. In *International Conference on Human-Computer Interaction*, pages 171–190. Springer, 2021.
- [34] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.
- [35] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

310 A Other Related Work

311 **Co-creation and Diversity.** Our work is closely related to a growing literature on homogenization
312 in generative AI systems. Prior work has shown that AI models, when prompted independently,
313 tend to produce highly similar outputs [31, 29]. Beyond studying AI-generated text in isolation,
314 an increasing body of research examines settings in which humans and AI co-create to varying
315 degrees. For example, Doshi and Hauser [5] show that using AI for creative ideation leads to more
316 homogeneous writing, while Agarwal et al. [1] and Padmakumar and He [17] find similar effects when
317 AI is used for text completion. Together, these results suggest that homogenization arises not only
318 from fully AI-generated content, but also from human exposure to AI suggestions. Our work builds
319 on this line of research by studying homogenization in a more natural co-creation setting that allows
320 for iterative prompting, editing, and selective adoption of AI outputs. We also explicitly manipulate
321 incentives for diversity, allowing us to examine how users strategically adjust their interaction with
322 AI in response to the task environment. Most closely related is Kosmyna et al. [12], which also
323 studies a relatively unrestricted co-creation setting. However, they focus on studying brain function
324 and cognitive debt rather than, as we do, on how incentives shape user strategies and the diversity of
325 outputs.

326 **AI assistance, creativity, and reliance.** A large body of work studies how AI assistance shapes
327 human creativity. Much of the cognitive science and human-computer interaction literature advocates
328 using generative AI to *enhance* creative processes rather than to replace human effort [33, 6, 4].
329 Creativity in these settings is often operationalized through divergent thinking tasks, where success is
330 measured by the ability to generate a wide range of novel ideas [21], see e.g., Torrance tests [25].
331 Empirically, recent studies paint a mixed picture. Sun et al. [24] conduct a field experiment and find
332 that AI assistance is most beneficial for workers who already engage deeply with their tasks. Other
333 work suggests that AI alone or human-AI teams can yield short-term gains in novelty or productivity,
334 but may undermine independent creative performance once assistance is removed [13, 9, 16, 23].
335 Our work complements these findings by adopting a broader notion of creativity that extends beyond
336 ideation to include writing style, narrative structure, and thematic exploration in creative writing.

337 A recurring concern in this literature is the risk of overreliance on AI to achieve tasks, whether
338 that be creative or otherwise. While human-AI collaboration can improve immediate performance,
339 recent evidence suggests it may reduce intrinsic motivation or engagement when the tool is no longer
340 available [32]. Related work also documents forms of cognitive offloading, in which users defer
341 critical thinking or decision-making to AI systems [12, 27, 8]. Although our study does not examine
342 long-term effects of AI use, we observe clear short-term costs of overreliance: participants who
343 heavily anchor on AI-generated drafts produce less diverse stories.

344 **Incentives in human-AI collaboration.** Recent work has increasingly emphasized the role of
345 incentives in human-AI decision-making. A central critique is that many studies overlook how
346 strongly incentive structures shape participant behavior and, consequently, study outcomes [11]. Most
347 related to our work is Holstein et al. [7], which shows that altering incentives can substantially reduce
348 overreliance on AI in decision-making tasks. To our knowledge, our study is the first to explicitly
349 manipulate incentive structures in a human-AI *co-creation* setting involving generative AI tools.

B Experiment Details

B.1 Randomization

Originality (O)	Technical Quality (T)
<p>“You are participating in a short story competition. There are thousands of submissions, so your goal is to stand out as much as possible. Find your voice and be as creative as possible! The short story should be around 250–350 words.”</p> <p><i>Meta instruction:</i> “Your bonus will be determined based on originality and uniqueness.”</p>	<p>“You are starting an Intro to Writing class. Your first assignment is to create a 250–350 word short story. Your goal is to get an A by submitting a high-quality piece of work.”</p> <p><i>Meta instruction:</i> “Your bonus will be determined based on the grade you receive.”</p> <p><i>Grading is based on a rubric, see Figure 6.</i></p>

Table 2: Instructions given to the originality (O) and technical quality (T) conditions.

	AI-O	SELF-O	AI-T	SELF-T
<i>Continuous variables</i>				
Prolific Approvals	1211.38 (62.21)	1171.52 (57.80)	1193.66 (61.10)	1165.34 (61.04)
Age	40.74 (1.90)	39.86 (1.86)	41.32 (1.93)	40.04 (1.75)
AI Experience (general)	2.30 (0.14)	1.94 (0.16)	1.90 (0.15)	2.04 (0.14)
AI Experience (writing)	1.64 (0.17)	1.08 (0.16)	1.46 (0.17)	1.06 (0.15)
<i>Binary variables</i>				
Student	0.14 (0.05)	0.22 (0.06)	0.21 (0.06)	0.11 (0.05)
Sex (Male)	0.36 (0.07)	0.34 (0.07)	0.32 (0.07)	0.42 (0.07)
Residence (USA)	0.34 (0.07)	0.42 (0.07)	0.44 (0.07)	0.36 (0.07)
Primary Language (English)	0.94 (0.03)	0.96 (0.03)	0.92 (0.04)	0.96 (0.03)
<i>Ethnicity</i>				
Asian	5	5	4	2
Black	7	8	3	5
Mixed	1	3	5	1
Other	1	1	3	1
White	36	33	35	41
<i>Employment Status</i>				
NIPW	5	4	8	5
Full time	25	23	17	24
Other	8	6	8	9
Part time	11	14	9	10
Unemployed (looking)	1	3	8	2

Table 3: Balance table of participant demographics by group. For continuous and binary variables, we display the mean and standard error (in parentheses). All groups consist of $n = 50$ participants. “NIPW” stands for Not in Paid Work. AI Experience is on a Likert scale from 0 to 4, with 0 being not at all familiar and 4 being extremely familiar with AI tools.

B.2 Interface Design

In this section, we explain the interface design in the online experiment.

B.2.1 Instruction Phase (Window 1)

Participants start with an initial instruction window, where they see different instructions depending on their randomized group, see Figure 5. If they were randomized into **SELF**, they were informed that any use of external AI tools are forbidden. If they were randomized into **AI**, they were told to only use the AI tool provided to them.

Instructions

Participant: ANON · Group: AI-DIV · Session: 5577c664-f52e-4c6a-99b8-8aef93715aff

Welcome! You'll complete a short creative writing task for a research study. Our goal is to understand how people write stories with the use of AI, in particular how it might contribute to less diverse stories. Please read the instructions carefully.

Instructions:

- Be thoughtful and authentic in your responses!
- **Stay on the session.** Once you leave the session, you will not be able to return or start a new session.
- You are provided an in-app AI tool to aid in your creative writing task; in fact you are **required to use the AI tool when prompted**. However, oftentimes the AI tool will not achieve the quality you desire. We encourage you to use the AI tool as a stepping stone and as a helper, but **ultimately you are responsible for your own story**. Please do not use external AI tools that we do not provide to you.

☐ I have read and understand the instructions.

[Begin Study](#)

Instructions

Participant: ANON · Group: SELF-CONV · Session: b5e1fc19-ca64-4c1f-a151-a96e4bc601a

Welcome! You'll complete a short creative writing task for a research study. Our goal is to understand how people write stories without the use of AI, in particular how people brainstorm, write first drafts, and refine their stories. Please read the instructions carefully.

Instructions:

- Be thoughtful and authentic in your responses!
- **Stay on the session.** Once you leave the session, you will not be able to return or start a new session.
- **Do not use external AI tools** (e.g., Google, ChatGPT). We will be monitoring for prohibited AI usage using keystroke data, attention checks, and your final submission. **If you are caught using AI, we will return your submission.**

☐ I have read and understand the instructions.

[Begin Study](#)

Figure 5: [Top] Instructions for **AI** group; [Bottom] Instructions for **SELF** group

359 B.2.2 Prompt Phase (Window 2)

360 After the instruction window, they are given their writing task, see Figure 6. The only difference
 361 in the instructions is in how they are graded; those in the **O** group are assessed by uniqueness and
 362 originality, while those in **T** are assessed by the technical qualities of their writing. Since writing
 363 quality tends to be vague, we included a grading rubric to make clear that participants will be assessed
 364 by their organization (narrative flow) as well as technique (grammar, spelling, and punctuation). We
 365 emphasized that we will not be grading for creativity or uniqueness.

366 B.2.3 Brainstorming phase (Window 3)

367 Participants are then brought to a brainstorming window, where they are given five minutes to
 368 brainstorm ideas and outline a story, see Figure 7. This window does not differ across groups.
 369 To better guide participants, we divided the outline into “Main Character”, “Setting”, “Conflict”,
 370 “Resolution”, and “Plot”. We also gave nudges that the story should be in first-person, take place
 371 over no more than one day, and should center on a decision or dilemma. While these were presented
 372 as instructions, they were intended to place guardrails on the seemingly open-ended task, since
 373 participants only had 20 minutes to write a story. Participants had to at least write 20 words in the
 374 outline, and spend at least one minute in this window to make sure that they are participating in good
 375 faith.

376 B.2.4 Writing phase (Window 4)

377 After brainstorming, participants are given 20 minutes to write their short story. We include the text
 378 from the outline in the previous window to help participants recall their writing plan. Those in **SELF**
 379 are given just a plain-text editor, while those in **AI** are given a side-by-side window where left is a
 380 plain-text editor and right is an AI conversation box. The AI assistant was gpt5-mini with reasoning

Prompt

You are participating in a short story competition.

There are thousands of submissions, so **your goal is to stand out** as much as possible. Find your voice and be as creative as possible! The short story should be 250-350 words.

In other words, your bonus will be determined based on originality and uniqueness.

Top % Originality	Bonus
Top 2%	\$7.00
Top 2-10%	\$4.00
Top 10-25%	\$2.00

In the next step, you will be given at most 5 minutes to brainstorm and outline your story plan. Then, you will have at most 20 minutes to write your story. You do not need to use the entire allotted time. After multiple periods of inactivity, we will flag your submission for manual review.

Please read the prompt carefully. You can proceed in 0 seconds.

Continue

Prompt

You are starting an Intro to Writing class.

Your first assignment is to create a 250-350 word short story. Your goal is to get an A by submitting a high-quality piece of work.

Grading Rubric

Your story will be evaluated on two aspects:

1. Organization (Narrative Flow)	(4 pt) Excellent: Clear beginning, middle, and end with smooth transitions (3 pt) Competent: Logical flow with minor structural issues (2 pt) Basic: Some organization but lacks coherence (1 pt) Does not meet expectations: Disorganized or unclear structure
2. Technique (Grammar, Spelling, Punctuation)	(4 pt) Excellent: No errors; demonstrates mastery (3 pt) Competent: Few minor errors that don't impede understanding (2 pt) Basic: Several errors that occasionally distract (1 pt) Does not meet expectations: Frequent errors that impede understanding
Creativity	Note: You are not graded on creativity or your voice . While creativity and uniqueness are often contributors to good pieces of writing, for now we are only asking for a technically well-written story.

Your bonus will be determined based on the grade you receive:

Point Range	Grade	Bonus
8 points	A	\$2.50
6-7 points	B	\$1.00
5 points or lower	C or lower	None

In the next step, you will be given at most 5 minutes to brainstorm and outline your story plan. Then, you will have at most 20 minutes to write your story. You do not need to use the entire allotted time. After 30 seconds of inactivity, we will flag your submission for manual review.

Please read the prompt carefully. You can proceed in 0 seconds.

Continue

Figure 6: [Top] Prompt and rubric for **O** group; [Bottom] Prompt and rubric for **T** group

Step 1 • Brainstorm

Story Guidelines

To help spark ideas and make your story easier to shape, we've added a few gentle guidelines to focus your creativity:

- Write the story from a **first-person** point of view.
- The story should take place over **no more than one day**.
- The story should center on a **decision or dilemma**.

Outline your story plan below. Use as many of the boxes as you find necessary. Remember, your goal is to **get the highest grade possible!** Remember, the story should be **250-350 words**.

Quick Ideas

Jot down as many ideas for a story you have.

Outline

Main Character

Who is your main character? What are their traits? Additionally, what is your character's goal in the story? What do they want?

Who is your main character?

Setting

Where does this story take place? When does this story take place? What time period does the story occur over (1 year? 1 day? 20 minutes?)

Where and when does the story take place?

Conflict

What is the main conflict of this story? How does this conflict prevent the character from getting what they want?

What problem drives the story?

Resolution

How does the story end? Does the main character achieve their goal? Why or why not?

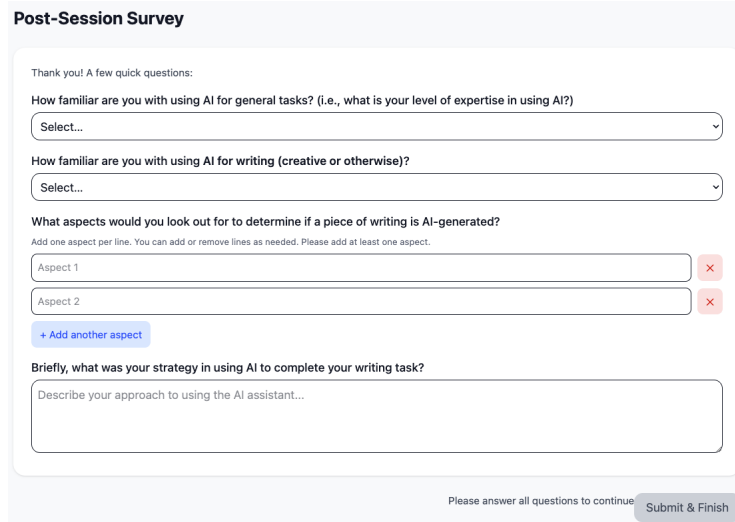
How is the conflict resolved?

Plot

Now, write out the events of the story. How does the story get from beginning to end? What happens? How is the resolution reached?

Summarize the main events or structure.

Figure 7: Brainstorming and outlining page.



Post-Session Survey

Thank you! A few quick questions:

How familiar are you with using AI for general tasks? (i.e., what is your level of expertise in using AI?)
 Select...

How familiar are you with using AI for writing (creative or otherwise)?
 Select...

What aspects would you look out for to determine if a piece of writing is AI-generated?
 Add one aspect per line. You can add or remove lines as needed. Please add at least one aspect.

Aspect 1 ×

Aspect 2 ×

[+ Add another aspect](#)

Briefly, what was your strategy in using AI to complete your writing task?
 Describe your approach to using the AI assistant...

Please answer all questions to continue [Submit & Finish](#)

Figure 9: Post-session survey questions.

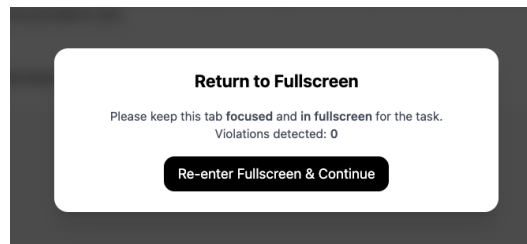


Figure 10: Warning sign instructing users to enter full screen to continue the study.

381 set to medium, and we capped usage to 15 calls/prompts for each session. While they are not required
 382 to use any of the AI suggestions, participants are required to prompt the AI tool at least once.

383 After 20 minutes, if the text in the text editor is not within 250-350 words, participants have a 5 minute
 384 grace period, during which they are instructed to quickly make edits to adhere to the word limit.
 385 After the 5 minute grace period, their session immediately ends and the responses are automatically
 386 submitted. Only 11 submissions did not adhere to the word limit, which we still include in the final
 387 analysis because they were reasonably within the word limit.

388 B.2.5 Post-Survey (Window 5)

389 Finally, we ask three questions to all participants, and an additional question for those randomized to
 390 **AI** (see Figure 9):

- 391 1. How familiar are you with using AI for general tasks? (0-4 Likert scale)
- 392 2. How familiar are you with using AI for writing (creative or otherwise)? (0-4 Likert scale)
- 393 3. What aspects would you look out for to determine if a piece of writing is AI-generated?
- 394 4. (Unique to **AI** group) What was your strategy in using AI to complete your writing task?

395 B.3 Attention Checks and Safeguarding Against AI Use

396 Ensuring the fidelity of an online experiment has been the topic of much debate recently, especially
 397 since participants can access a generative tool externally [20, 30]. This is particularly concerning in
 398 our setting where the **SELF** group serves as an important human baseline. We took great steps to
 399 safeguard against external AI use, and in this section we discuss the strategies we employed.

Instructions

Participant: ANON - Group: SELF-DIV - Session: 31491b95-b254-4fd7-97cb-bfa187339675

Welcome! You'll complete a short creative writing task for a research study. Our goal is to understand how people write stories without the use of AI, in particular how people brainstorm, write first drafts, and refine their stories. Please read the instructions carefully.

Instructions:

- Be thoughtful and authentic in your responses!
- Stay on the session.** Once you leave the session, you will not be able to return or start a new session.
- Do not use external AI tools** (e.g., Google, ChatGPT). We will be monitoring for prohibited AI usage using keystroke data, attention checks, and your final submission. **If you are caught using AI, we will return your submission.**

☒ I have read and understand the instructions.

Attention Check

Based on the instructions above, what is the policy regarding AI usage in this study?

☐ Zero tolerance - No AI usage allowed at any point
 ☒ AI usage is only allowed when specifically provided by the study
 ☐ Any AI usage is allowed throughout the study

⚠ Attention Check Failed

You have failed to correctly identify the AI usage policy after multiple attempts. This indicates that you may not have read the instructions carefully enough.

Unfortunately, we cannot proceed with the study if participants do not understand the key requirements.

Please [click this link](#) to go back to Prolific and return the study. Alternatively, copy and paste this code: **C246RY97**.

Your **bonus** will be determined based on the **grade** you receive:

Point Range	Grade	Bonus
8 points	A	\$2.50
6-7 points	B	\$1.00
5 points or lower	C or lower	None

In the next step, you will be given at most 5 minutes to brainstorm and outline your story plan. Then, you will have at most 20 minutes to write your story. You do not need to use the entire allotted time. After 30 seconds of inactivity, we will flag your submission for manual review.

Based on the prompt, what is your bonus determined by?

☐ Originality and uniqueness
 ☐ The grade I receive

Begin Brainstorm

Figure 11: Attention checks for the Instruction phase (top) and the Prompt phase (bottom). Participants are not allowed to continue the session if they fail to answer both questions correctly.

Global strategies. Participants are asked to enter full screen mode. If they exit from full screen or switch windows while on full screen, they will be met with a warning as in Figure 10, which displays the number of violations so far in the session. If the participant reaches 5 violations, the session automatically ends and their submission will not be counted toward the final analysis. Whenever they are booted off the platform, the same participant ID cannot restart the session.

We also employ an **attention mechanism** throughout the session. The window has listeners to keep track of mouse and keyboard activity, though these are not saved. After 30 seconds of consecutive mouse or keyboard inactivity, participants are given a warning saying that they should stay engaged by moving their mouse or typing. In each session, they are given 6 such warnings before their session automatically ends and their submission will not be counted toward the final analysis. After the second-to-final warning, the warning persists in the window to make sure that participants know the risk of being banned from the platform. Participants are given up to 6 chances to ensure that they can take breaks (e.g., to grab water), or for slack if they are thinking for extended periods of time. If, however, there is minutes of inactivity at any given point, the session automatically ends.

Window-specific strategies. In each window, we also employ specific strategies to make sure participants are paying attention and engaged.

- Instruction phase (Window 1): Participants are asked to acknowledge that they read the instructions, and are given a quick attention check to make sure they have read the AI policy (see Figure 11, top). If they fail, their session will immediately end and they cannot restart.
- Prompt phase (Window 2): Similarly, participants are asked a question about how their writing is assessed (see Figure 11, bottom). If they fail, their session will immediately end and they cannot restart.
- Brainstorming phase (Window 3): Participants cannot copy paste anything onto the textboxes. The attention mechanism is also in place in this window.
- Writing phase (Window 4): Participants in **SELF** cannot copy paste anything onto the textbox, while those in **AI** can only copy paste within their window in case they want to integrate the AI suggestions into their own writing. The attention mechanism is also in place in this window.

Post-hoc strategies. Even after the extensive strategies employed during the session, we conducted one final analysis by leveraging the 5-second snapshots of each participant’s writing. For every participant in the **SELF** group, we analyzed the extent to which they backtracked or stopped writing—indications that they hesitate and revise—consistent with a realistic human writing process. Only one participant failed our threshold because they barely backtracked or stopped; We cannot, however, determine whether the participant is an incredibly confident writer or using external AI tools on the side, and so we decided to pay the participant but exclude their results from the study.

In all, the strategies we employed are quite conservative, and we managed to filter out 29 Prolific workers. In fact, a handful of participants had made good effort attempts but were still banned from the platform; we made sure to go through each complaint and pay each participant accordingly for their time. These mechanisms cannot, however, exclude the possibility that an extremely sophisticated participant is using an external AI tool on another device, and is writing the story with uncertainty and pause. We believe that this is an unlikely scenario to begin with, and even if they exist, are rare enough not to significantly change our overall findings.

B.4 Recruitment Details

In order to obtain high-quality submissions, we applied the following inclusion criteria: (1) Resides in the United States or United Kingdom; (2) Primary language is English or fluent in English; (3) At least 18 years old; (4) At least 500 previously completed tasks on Prolific; (5) Historical approval rate of at least 98%. We sought to recruit individuals who are experienced and have a good track record in Prolific, and can write confidently in English due to the strict time constraint.

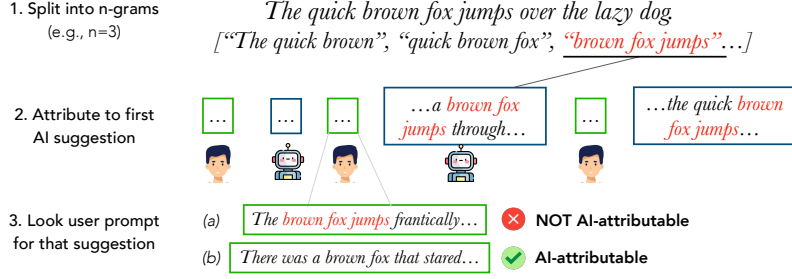


Figure 12: Example of the AI attribution procedure to construct the adoption metric.

B.5 Payment Details

Participants are paid \$5 for completing the study. On average, participants took around 25 minutes to complete, resulting in an hourly pay of around \$12. We also gave a fixed bonus of \$1 for every completed work; this is to deliver on the promise in the prompts that promised to pay bonuses based on the quality of their work. However, instead of going through each story and making a (subjective) decision on quality, we decided to give a fixed payment of \$1. Since this payment was done well after participants completed the study, it has no impact on the incentives that the prompt were intended to have. Note also that we gave partial payments for edge cases where participants failed some attention check, but already invested sufficient time into the task and were determined to have done the task in good faith.

C Methods Details

C.1 Measuring Adoption on AI-Generated Text

For each participant i , we observe a final written submission F_i and a sequence of AI-generated suggestions $\{S_{ik}\}_{k=1}^{m_i}$ produced during the interaction. These AI generations may take different forms, including full drafts, partial rewrites, or shorter textual suggestions. Our goal is to quantify the extent to which the participant’s final submission is anchored on AI-generated content.

We operationalize this notion via a single adoption measure,

$$A_i \in [0, 1],$$

which captures the fraction of the final submission that can be attributed to AI-generated text. Intuitively, A_i is high when large portions of the final submission closely resemble AI output, and near zero when the final submission is largely authored independently.

To construct A_i , we consider two complementary attribution schemes: a conservative *n-gram-based* approach that captures near-verbatim reuse, and a more permissive *embedding-based* approach that additionally captures paraphrased reuse. In both cases, the core idea is to identify which portions of the final submission can be plausibly traced to any AI generation, and to measure their total mass. We have introduced the *n-gram* based metric in the main text; here, we introduced the embedding-based approach, which we use as a robustness check for our main reliance measure.

C.1.1 N-gram based

Let $F_i = (f_1, \dots, f_{T_i})$ denote the token sequence of the final submission, where T_i is the total number of tokens. For each AI suggestion S_{ik} , we extract all contiguous *n*-grams (for a fixed *n*, e.g., $n = 10$). We then scan the final submission for maximal contiguous spans of tokens that appear verbatim in S_{ik} .

When an *n*-gram in the final submission appears in multiple AI generations, we attribute it to the *earliest* AI generation in which it appears. However, to avoid attributing content that originated from the participant rather than the AI, we further check whether this *n*-gram already appears in the corresponding user prompt that produced that AI generation. If so, the *n*-gram is not counted as AI-attributable because the participant originally generated it. Let W_i denote the number of final

tokens that are AI-attributable under this criterion. The n-gram-based adoption measure is then defined as

$$A_i = \frac{W_i}{T_i}.$$

C.1.2 Embedding-Based Reliance (Paraphrased Reuse)

The n-gram approach is conservative and does not capture paraphrased reuse. We therefore define a second reliance measure based on local semantic alignment using sentence embeddings. We segment the final submission into a sequence of spans (e.g., sentences), $\{F_{iu}\}_{u=1}^U$, where each span covers a contiguous set of tokens. Each AI suggestion S_{ik} is segmented in the same way. Let $e(\cdot)$ denote a sentence embedding function.

For each final span F_{iu} , we compare it sequentially to the AI generations in the order they were produced. Specifically, we identify the *first* AI generation k for which F_{iu} exhibits sufficiently high semantic similarity to any span in A_{ik} , as measured by cosine similarity between embeddings. Formally, we consider the smallest index k such that

$$\max_v \cos(e(F_{iu}), e(S_{ikv})) \geq \tau,$$

for a fixed high threshold τ .

To avoid attributing content that originated from the participant rather than the AI, we further check whether F_{iu} is already present in the corresponding user prompt that produced generation S_{ik} , using the same embedding-based similarity criterion (and optionally additional lexical guardrails). If F_{iu} matches the prompt, it is treated as human-originated and is not counted as AI-attributable.

A final span is considered AI-attributable if it satisfies the above criterion for some AI generation and does not match the corresponding user prompt. These criteria are chosen to capture close semantic reuse, including paraphrasing, while avoiding attribution to generic topical similarity or high-level advice.

Let m_u denote the token length of span u , and let

$$W_i^{\text{emb}} = \sum_{u=1}^U m_u \cdot \mathbf{1}\{F_{iu} \text{ is AI-attributable}\}$$

be the total number of final tokens attributable to AI under the embedding-based criterion. We define the embedding-based reliance measure as

$$A_i^{\text{emb}} = \frac{W_i^{\text{emb}}}{T_i}.$$

C.1.3 Robustness of adoption metric

In the main text, we used $n = 10$ -gram as the basis of our adoption metric. In Figure 13, we show that our choice is $n = 10$ is highly correlated ($\rho \geq 0.94$) with other choices: lower n , aggregated over multiple choices of n , or even the embedding-based metric. We therefore do not replicate our results with this metric, since the analyses will be nearly identical.

D Additional Results

D.1 User Attitudes About AI Use

We now describe users' self-reported strategies in using AI to complete their tasks.

Participants felt agency and used AI only as a tool. Most participants treated the AI as a tool to generate an initial draft or structure, while viewing themselves as the agent who refines, edits, and enforces personal voice. Participants frequently described workflows such as “*I prompted the AI and used the text as a framework,*” “*AI laid it out nicely. . . I took some material from each of two drafts, then edited and altered the text to my taste,*” or “*Get a very strong baseline and then edit that manually.*” These accounts correspond to our finding that there exists substantial anchoring on AI-generated content.

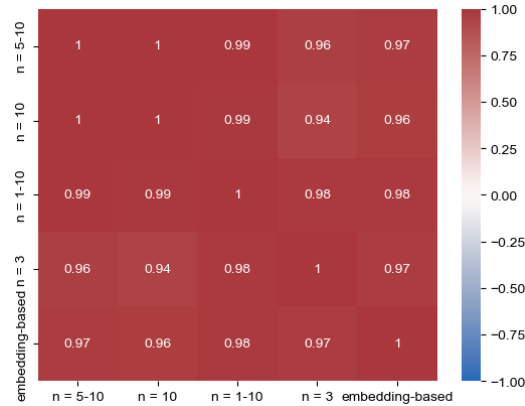


Figure 13: Correlation matrix of various adoption metrics. For example, $n = 1 - 10$ means that the adoption metric aggregates from $n = 1$ to 10.

At the same time, users consistently emphasized their retained control and authorship. Many explicitly referenced expressing their own writing style, describing efforts to “*make it more like my voice*” or “*adapt it to my taste*.” Notably, this language appears even among the high-adoption group. Despite relying heavily on AI drafts and often making only incremental edits, these participants report leaving the task with a sense of ownership and satisfaction. This suggests that no matter the strategies participants ended up employing, they generally perceived a sense of control, perhaps due to the full freedom given to them in interacting with the AI tool.

Participants were almost always unsatisfied with the AI drafts. We also found that initial AI drafts were frequently perceived as unsatisfactory (e.g., “*it gave me something that wasn’t very satisfying or quite like what I had hoped for*,” or “*[the draft] seemed quite poor*”). However, most users responded by iterating rather than abandoning the text: “*When that wasn’t very good [I would] ask it to redraft... then adjust where necessary*.” Iterative refinement appears to be the modal response to imperfect outputs.

In contrast, participants in the low-adoption group were more likely to disengage after evaluating the AI’s suggestions. Their responses often reflected disappointment with quality (“*It did not stay on point... I decided I could be better creative on my own*,” “*I tried to use the AI to assist, but it wasn’t very useful*,” “*It made it feel too computer generated*.”). Others expressed principled objections independent of quality concerns: “*I severely object to using AI in creative writing, as a novelist myself*,” or “*To ignore it almost completely. I felt it was too easy*.” These participants either abandoned the AI draft early or limited its role substantially.

Even so, low-adoption participants did not necessarily avoid the tool altogether. Many still used it for limited purposes such as inspiration (“*I used it to give me some inspiration to how to word certain settings*”) or consulting (“*I used it to assess my paragraphs and writing*”). These auxiliary uses—consulting, grammar checking, phrasing suggestions—also appear across all adoption levels.

Incentivizing originality changed how users perceived their role in co-creation. Finally, we also observe subtle but meaningful shifts in self-reported strategies between the **AI-O** and **AI-T** conditions. Participants in **AI-O** tended to self-report lower adoption on AI, frequently emphasizing that they were prioritizing creativity and personal voice (“*Add my own touches*,” “*translate it to my own voice*”). This shift in strategy suggests that participants are generally aware that AI creates a homogeneous style, and that the treatment altered how users conceptualized their role in the co-creation process.

D.2 Diversity of Stories

Table 5 shows the effect sizes from difference-in-means tests over the final submissions and the first valid draft the AI suggested. Red stars indicate significance after a Benjamini-Hochberg adjustment, but the results remain largely consistent.

D.3 Characterizing Writing Trajectories

In Figure 2, we plotted results for all-MiniLM-L12-v2. Here, we display the same plots for all pre-registered similarity (and diversity in the case of NGDS) metrics. Figure 14 shows the average similarity across randomized groups for both the final submission and the first full draft of AI. Figure 15 plots the trajectories of the similarity metric across groups through time. Finally, Figure 16 shows scatter plots of the first and second principal components (PC) from the story’s text embeddings. All the results as discussed in the main text generally hold.

D.4 Differential Strategies in Using AI

Participants exert more effort in using AI when incentivized to be original, with mixed results.

We find that participants in **AI-O** exert more effort in prompting AI: across adoption levels, they prompt the AI more frequently and spend up to 50% more time interacting with it than participants in **AI-T** (Figure 3(b, d)). This evidence shows that participants in **AI-O** engage substantially with AI to pursue their objectives, but this effort does not yield significant changes in diversity between **AI-O** and **AI-T**. This may be because participants cannot meaningfully differentiate their outputs via AI use, or because their writing objectives do not exactly align with our diversity measure.

Users with more AI experience tend to anchor more on AI suggestions. Table 4 compares participant characteristics between the high- and low-adoption groups, where most covariates are well balanced across the two groups. The most pronounced difference is prior experience with AI, both in general and for writing tasks: participants in the high-adoption group report substantially greater AI experience than those in the low-adoption group (e.g., 2.52 vs. 1.81 for general AI experience). Unsurprisingly, participants with more AI experience tend to prompt more, spend more time interacting with the tool, and adopt similar prompting strategies to those in the high-adoption group. In particular, they are more likely to ask AI for partial edits throughout the session. This pattern suggests that greater familiarity with AI is associated with increased trust in, and reliance on, AI tools. In the context of our task, however, this tendency is counterproductive because heavier anchoring on AI drafts results in lower diversity. Note that we cannot rule out the possibility that these more experienced users are using AI for some other goal that our metrics do not capture, such as improving quality.

Adoption rate	High ($n = 27$)	Low ($n = 37$)
<i>Continuous variables</i>		
Total Approvals	1221.48 (96.18)	1263.19 (60.43)
Age	39.19 (2.18)	42.16 (2.39)
AI Exp. (General)	2.52 (0.21)	1.81 (0.19)
AI Exp. (Writing)	2.00 (0.26)	1.14 (0.19)
<i>Binary variables</i>		
Student	0.07 (0.05)	0.16 (0.06)
Sex (Male)	0.30 (0.09)	0.41 (0.08)
Residence in US	0.52 (0.10)	0.43 (0.08)
Primary Lang (Eng.)	0.93 (0.05)	0.95 (0.04)
<i>Ethnicity</i>		
Asian	14.8%	8.1%
Black	14.8%	8.1%
Mixed	7.4%	5.4%
Other	3.7%	5.4%
White	59.3%	73.0%
<i>Employment status</i>		
NIPW	11.1%	16.2%
Full time	44.4%	29.7%
Other/Expired	18.5%	18.9%
Part time	11.1%	29.7%
Unemployed	14.8%	5.4%

Table 4: Participant demographics for high and low adoption groups. For continuous and binary variables, we display the mean and standard error (in parentheses). “NIPW” stands for Not in Paid Work. AI Experience is on a Likert scale from 0 to 4, with 0 being not at all familiar and 4 being extremely familiar with AI tools.

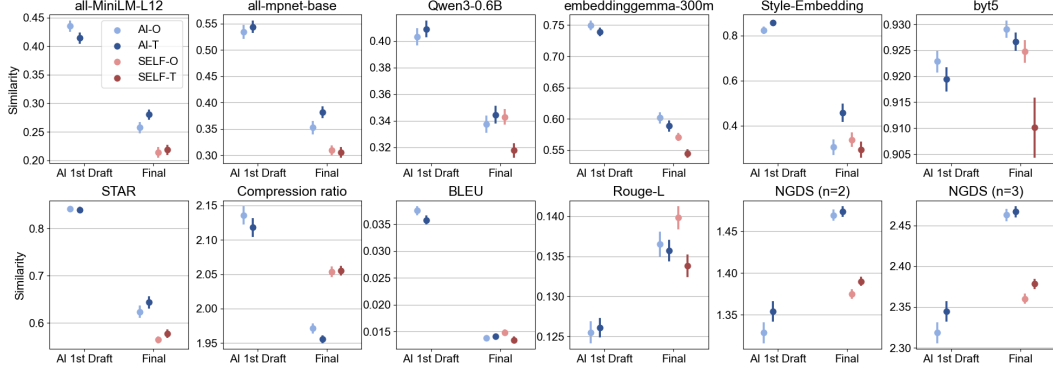


Figure 14: Average similarity (diversity for NGDS) across randomized groups, for both the final submission and the first full draft the AI produces (when available). NGDS is the diversity metric, in that a higher value means lower similarity.

Table 6 displays the effect sizes from difference-in-means tests, where the **AI** group is stratified into high and low adoption ($A > 0.8$ and $A < 0.2$, respectively).

Comparison	AI-O – SELF-O		AI-T – SELF-T		AI (T – O)		SELF (T – O)
Version	Final	First (AI)	Final	First (AI)	Final	First (AI)	Final
Embeddings							
all-MiniLM-L12	0.043***	0.212***	0.062***	0.191***	0.023**	−0.017	0.004
all-mpnet-base	0.043***	0.227***	0.076***	0.232***	0.029**	0.001	−0.003
Qwen3-Embedding	−0.005	0.060***	0.027***	0.087***	0.007	0.001	−0.025
embeddinggemma	0.031***	0.178***	0.044***	0.189***	−0.013	−0.015	−0.026
Style-Embedding	−0.033	0.485***	0.164***	0.551***	0.153***	0.023**	−0.044
byt5	0.004**	−0.003	0.017***	0.010**	−0.002	−0.001	−0.015
star	0.059***	0.273***	0.067***	0.253***	0.020*	−0.008	0.013*
N-gram/token based							
compression ratio	−0.083	0.087***	−0.100	0.056***	−0.016	−0.029	0.002
self bleu	−0.001	0.022***	0.001	0.022***	0.000	−0.001	−0.001
rougel	−0.003	−0.016	0.002	−0.008	−0.001	0.002	−0.006
n gram diversity score 2	0.095	−0.051***	0.083	−0.030***	0.004	0.036	0.015
n gram diversity score 3	0.103	−0.048***	0.088	−0.029***	0.004	0.037	0.018

Table 5: Effect sizes from difference in means tests (Welch t-test, one-sided) over the final submissions (“Final”) and the first valid draft the AI suggested when available (“First (AI)”). Black significance stars are standard p-values, while red stars use Benjamini–Hochberg (BH) FDR-adjusted p-values: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

E Discussion (Continued)

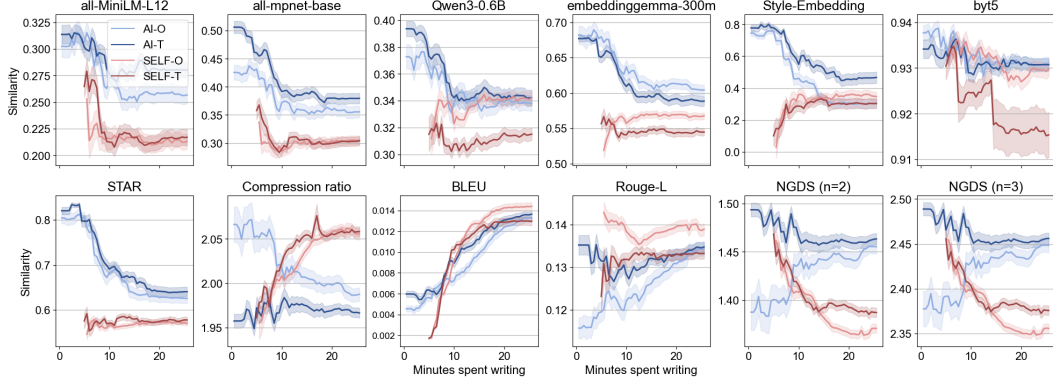


Figure 15: Average similarity (diversity for NGDS) randomized groups through time spent in the session. NGDS is the diversity metric, in that a higher value means lower similarity. We only include text that has at least 200 words for every timestep to ensure that the embeddings capture enough information about the story.

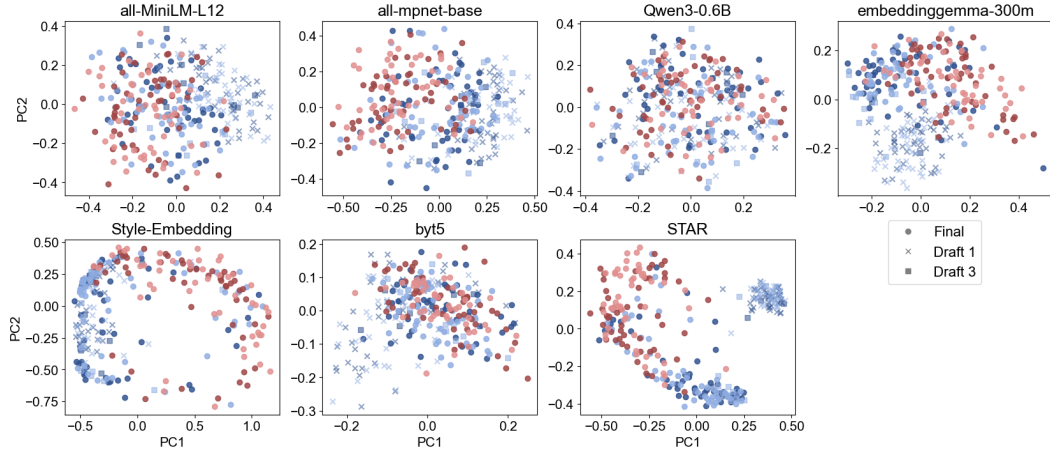


Figure 16: First and second principal component of the embeddings across groups, for both the drafts that AI produces (when available) and the final submissions.

Comparison	AI-O- SELF-O		AI-T- SELF-T		AI-T- AI-O	
AI Adoption level	Low	High	Low	High	Low	High
Embeddings						
all-MiniLM-L12-v2 ↓	0.015	0.069***	0.051***	0.072***	0.041**	0.008
all-mpnet-base-v2 ↓	0.016	0.061***	0.021	0.115***	0.002	0.050***
Qwen3-Embedding-0.6B ↓	-0.016	0.009	0.004	0.043***	-0.005	0.009
embeddinggemma-300m ↓	0.007	0.036***	0.011	0.072***	-0.023	0.010
Style-Embedding ↓	-0.157	0.042	-0.045	0.261***	0.068	0.175***
byt5 ↓	0.004*	0.004	0.016***	0.016***	-0.002	-0.003
star ↓	-0.027	0.116***	-0.025	0.115***	0.015	0.011*
N-gram/token based						
Compression Ratio ↓	-0.053	-0.111	-0.079	-0.117	-0.024	-0.004
BLEU ↓	-0.002	-0.000	-0.000	0.000	0.001	-0.001
ROUGE-L ↓	-0.006	-0.003	0.002	0.000	0.002	-0.003
NGDS ($n = 2$) ↑	0.067	0.119	0.053	0.108	0.001	0.004
NGDS ($n = 3$) ↑	0.074	0.129	0.056	0.115	0.000	0.004

Table 6: Effect sizes from difference in means tests (Welch t-test, one-sided) with significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, over the final submissions stratified by low and high adoption on AI among the AI-assisted group.