Ca' Foscari University of Venice

Department of Economics

Working Paper

Federico Atzori
Luca Corazzini
Valeria Maggian
Filippo Pavesi
Massimo Scotti

Human-AI Interaction in Creative Tasks: an Experimental Investigation

ISSN: 1827-3580
No. 16/WP/2026

Working Papers
Department of Economics
Ca' Foscari University of Venice
No. 16/WP/2026
ISSN 1827-3580

Human-AI Interaction in Creative Tasks: an Experimental Investigation

Federico Atzori
Sapienza University

Luca Corazzini
University of Milan - Bicocca

Valeria Maggian
Ca' Foscari University of Venice

Filippo Pavesi, Massimo Scotti
LIUC University

Abstract
We investigate how generative AI shapes creative performance and human-AI interaction in an open-ended writing task that employs a laboratory experiment in which participants are randomly assigned to either receive access to a large language model (ChatGPT-4.2) or not. Creative performance is measured by the average score assigned by independent evaluators recruited through the Prolific platform, and detailed logs of human-AI interaction are analyzed to measure AI use, prompting intensity, ideation requests, and the textual overlap between AI outputs and participants' final writings. Three main results emerge. First, AI access increases performance, but the gain is entirely driven by active use: participants with access who do not submit queries perform no better than those without AI. Second, the relationship between interaction intensity and performance is concave, peaking at roughly eight queries, consistent with iterative exploration rather than mechanical copying. Third, structural mediation analyses show that ideation requests affect performance primarily indirectly, by increasing downstream incorporation of AI-generated language; the direct effect of requesting an idea from the AI is negligible once execution-stage reliance is accounted for. We further document heterogeneity in AI reliance: cultural capital (proxied by books owned) predicts lower AI use, while prior AI exposure predicts higher use. By contrast, incentive schemes have limited effects on both outcomes and AI-related behaviors.

Keywords
Human-AI Interaction; Creativity; Generative AI; Laboratory Experiment

JEL Codes
C91, D83, J24, O33

Address for correspondence:
Luca Corazzini
Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta S.Giobbe
30121 Venezia - Italy
e-mail: luca.corazzini@unimib.it

The Working Paper Series
is available only on line
(https://www.unive.it/dec-workingpapers)
For editorial correspondence, please contact:
wp.dse@unive.it

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta San Giobbe
30121 Venice Italy
Fax: ++39 041 2349210

HUMAN-AI INTERACTION IN CREATIVE TASKS: AN EXPERIMENTAL INVESTIGATION

FEDERICO ATZORI, LUCA CORAZZINI, VALERIA MAGGIAN, FILIPPO PAVESI, AND MASSIMO SCOTTI

ABSTRACT. We investigate how generative AI shapes creative performance and human-AI interaction in an open-ended writing task that employs a laboratory experiment in which participants are randomly assigned to either receive access to a large language model (*ChatGPT-4.2*) or not. Creative performance is measured by the average score assigned by independent evaluators recruited through the Prolific platform, and detailed logs of human-AI interaction are analyzed to measure AI use, prompting intensity, ideation requests, and the textual overlap between AI outputs and participants' final writings. Three main results emerge. First, AI access increases performance, but the gain is entirely driven by active use: participants with access who do not submit queries perform no better than those without AI. Second, the relationship between interaction intensity and performance is concave, peaking at roughly eight queries, consistent with iterative exploration rather than mechanical copying. Third, structural mediation analyses show that ideation requests affect performance primarily indirectly, by increasing downstream incorporation of AI-generated language; the direct effect of requesting an idea from the AI is negligible once execution-stage reliance is accounted for. We further document heterogeneity in AI reliance: cultural capital (proxied by books owned) predicts lower AI use, while prior AI exposure predicts higher use. By contrast, incentive schemes have limited effects on both outcomes and AI-related behaviors.

Key words and phrases. Human-AI Interaction; Creativity; Generative AI; Laboratory Experiment.
JEL: C91; D83; J24; O33.

Federico Atzori: CIMEO, Sapienza University, Rome, Italy fatzori19@gmail.com; **Luca Corazzini, corresponding author**: University of Milan - Bicocca, Italy - luca.corazzini@unimib.it; **Valeria Maggian**: Ca' Foscari University of Venice, Italy - valeria.maggian@unive.it; **Filippo Pavesi**: LIUC University, Castellanza, Italy - fpavesi@liuc.it; **Massimo Scotti**: LIUC University, Castellanza, Italy - mscotti@liuc.it.

Funding sources, Ethical approval and Data linking: This work was financially supported by the project "Gender Stereotypes and Women's Achievements (GENS)," PRIN - 2022 N. 20229RB8HY.
The experimental protocol received ethics approval from the Ethical Committee of Ca' Foscari University of Venice (n. 2/2025).
All data, *STATA* do files, and other relevant codes for reproducing the results reported in the analysis can be found online at https://doi.org/10.6084/m9.figshare.32132980.

1. Introduction

The rapid emergence of generative artificial intelligence (AI) is transforming how individuals perform cognitive and creative tasks (Krakowski, 2025). Systems such as large language models can generate text, audio, images, and ideas, and are increasingly integrated into professional and everyday activities, enabling new forms of human-AI collaboration in domains including writing, design, and marketing (Bervar et al., 2026; Tsao et al., 2025; Heigl, 2025; Sun et al., 2024; Epstein et al., 2023; Gero et al., 2023; Yuan et al., 2022). As a result, creative production is shifting from a predominantly human activity to a hybrid process in which humans and artificial agents jointly generate, refine, and finalize ideas. Understanding how individuals interact with these systems, and how such interactions affect creative performance is therefore a central question at the intersection of economics and psychology.

Reflecting this growing relevance, experimental research on AI and creativity has expanded rapidly in recent years (Holzner et al., 2025). This literature has mainly developed along three dimensions: comparing the creative performance of generative AI and humans, assessing whether AI support enhances human creativity relative to unaided effort, and examining how AI assistance affects the diversity of ideas produced by humans. Our paper contributes to the second strand by investigating how access to generative AI affects human creative output.

Despite growing evidence on the effects of generative AI on creative outcomes, the behavioral mechanisms underlying these effects remain poorly understood (Holzner et al., 2025). In particular, little is known about how individuals interact with AI systems during the creative process and how different interaction patterns shape final outcomes. In this paper, we address this gap by examining the modalities of human-AI interaction in a controlled experimental setting and linking them to observed creative performance.

To this end, we conduct a laboratory experiment in which participants perform an open-ended creative task under different incentive schemes, either with or without access to a generative AI system (ChatGPT-4.2). The task consists of writing a short composition based on the question: *"If you had the talent to create something simply with your thoughts, what would you create?"* (Charness and Grieco, 2019). The resulting compositions are subsequently evaluated by independent external raters via a survey.

The selected task belongs to the category of "open" creative tasks, as participants are free to provide any response without constraints (Charness and Grieco, 2019). Such tasks are especially suitable for studying human-AI collaboration because they admit multiple possible solutions and leave individuals substantial discretion in how to approach the task and whether to rely on AI assistance. Consistent with this objective, we impose no restrictions on how participants in the AI condition may use ChatGPT. Importantly, our design allows us to observe not only final outputs but also the complete interaction history between participants and the AI system, including all prompts and responses. Combined with the controlled laboratory environment, this provides rich and reliable data for analyzing how individuals use AI during the creative process and how these behaviors translate into creative outcomes.

We structure the analysis as follows. First, we consider a simple measure of interaction intensity, defined as the number of prompts an individual submits while completing the creative task. We then examine how individuals use and interact with AI along the creative process. To this end, we conceptualize creativity as unfolding in two sequential stages: an ideation stage, in which individuals generate and select ideas, and an execution stage, in which these ideas are developed into a final output. This distinction, rooted in the dual-process view of creativity (Guilford, 1957; Mednick, 1962), allows us to move beyond a simple measure of AI use and instead characterize how individuals rely on AI at different stages of the creative process. Exploiting the sequential nature of this process, we implement a mediation framework to disentangle the direct effect of AI-supported ideation from the indirect effects operating through subsequent reliance on AI during content development. This approach allows us to gather a better understanding of the overall impact of AI-supported ideation on creative performance.

Our results provide several insights into the nature of human-AI collaboration in creative tasks. First, we find that access to generative AI increases performance, but the gain is entirely driven by active use: participants with access who do not submit queries perform no better than those without AI. Second, the relationship between AI use and performance is non-linear: performance increases with the number of queries up to approximately eight and declines thereafter, suggesting diminishing returns to AI interaction. Third, mediation analysis shows that AI-supported ideation affects performance primarily indirectly, by increasing reliance on AI-generated content in the development of the final output. The direct effect of requesting ideas from the AI is negligible once this channel is accounted for. This suggests that requesting ideas from AI does not improve performance per se; rather, its benefit arises from shaping subsequent interaction with the AI and the extent to which AI-generated content is incorporated into the final composition. In particular, individuals who rely on AI during the ideation stage tend to engage more intensively with the system and to incorporate more AI-generated content, and it is this combination that enhances creative performance.

Additional findings shed further light on the mechanisms underlying the Human-AI interaction. We show that individual creative ability, measured using the Divergent Association Task (DAT), has a significant concave relationship with evaluations when participants work without AI assistance. However, this relationship disappears when AI is available, suggesting that AI tools may partially compensate for differences in baseline creative ability. Moreover, when examining the role of monetary incentives, we find that their structure has limited effects on both creative performance and patterns of interaction with the AI system.

We also document heterogeneity in AI usage across individual characteristics. In particular, more intensive AI usage is associated with lower levels of cultural capital and greater prior exposure to AI. This pattern points to both substitution and learning dynamics in human-AI collaboration: individuals with greater cultural capital or lower familiarity with AI are less likely to exploit the opportunities offered by these technologies. As a result, AI-assisted creativity may contribute to a convergence in creative performance, allowing individuals with fewer initial resources to catch up, provided they are willing to engage with and experiment with AI systems.

Taken together, our findings contribute to the literature on human-AI collaboration in three main ways. First, the study provides new experimental evidence on how humans interact with generative AI systems in open-ended creative tasks. Second, it advances understanding of the behavioral mechanisms underlying AI-assisted creativity by analyzing how individuals integrate AI assistance across different stages of creative production. Third, it introduces a methodological framework that combines behavioral data with mediation analysis to identify the roles of ideation and output finalization in creative performance.

The paper proceeds as follows. In Section 2 we present the experimental design. Section 3 discusses our predictions and conceptual framework. Section 4 presents our results. Finally, in Section 5, we discuss how our results relate to the literature and conclude.

2. Experimental Design

This section describes the experimental design used to study how access to generative AI affects creative performance under alternative incentive schemes.

The study consists of two complementary components. First, a laboratory experiment in which participants complete a creative writing task under different payment schemes and with or without AI assistance. Second, a survey experiment in which independent external evaluators assess the creativity of the texts produced in the laboratory. Because compensation in the laboratory depends on these evaluations, final payments are determined only after the survey experiment. Figure 1 illustrates the timeline of the experimental study.

2.1. **The Laboratory Experiment.**

Participants in the laboratory experiment completed a creative writing task based on the question: *"If you had the talent to create something simply with your thoughts, what would you create?"* (Charness and Grieco, 2019). Participants were encouraged to spend no more than 15 minutes on the task and to produce a text of approximately 150 words.

The experiment follows a 2×3 between-subjects design that varies: (i) the availability of generative AI assistance (AI vs. No AI) and (ii) the incentive scheme. We build our experiment on the framework

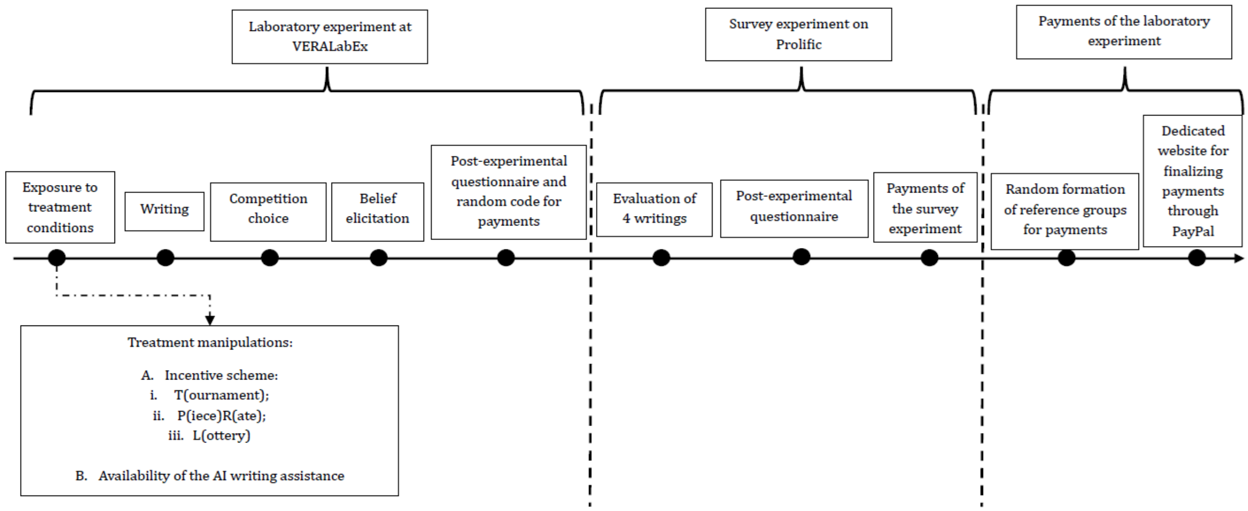

FIGURE 1. Time flow of the experimental study

of Niederle and Vesterlund (2007) and Erat and Gneezy (2016), adopting a between-subjects design as, given the open-ended and creative nature of the task, a within-subjects design would be prone to learning, carryover, and idea-reuse effects, thereby threatening the validity of our results.

In AI treatments, participants had access to a generative AI assistant implemented via the *ChatGPT-4.2* API and embedded in the experimental interface. They could submit written queries and receive real-time responses, with no restrictions on interaction. All exchanges were recorded, allowing us to observe both the content and intensity of AI use. Participants were informed that their inputs would not be stored or used for model training.

Participants were assigned to one of three incentive schemes:

- *Piece-Rate (PR)*: Earnings equal 0.6 times the average creativity score assigned by four external evaluators;
- *Tournament (T)*: Participants are randomly grouped in fours (two males and two females).[1] The group member with the highest average score receives 2.4 euros per point (i.e., 4 times higher than the Piece rate payment scheme) while others receive zero.
- *Lottery (L)*: Within each group of four, one participant is randomly selected to receive 15 euros, independently of performance.[2] This treatment removed any direct link between performance and earnings.

The experiment consists of three stages. Instructions about each stage are revealed sequentially only after

[1]Each experimental session was exactly balanced with respect to gender. Participants were informed about the random assignment into groups and the gender composition of each group before the implementation of stage 1.

[2]The fixed prize of 15 euros was chosen to match, in expected value, the average payments participants could receive under the Piece-Rate and Tournament schemes, assuming an average creativity rating of 6.

the previous stage is completed. At the end of the session, either Stage 1 or Stage 2 is randomly selected with equal probability to determine, together with Stage 3, participants' earnings.

- *Stage 1 (Creative Writing Task)*: Participants complete the writing task under their assigned incentive scheme, with or without AI access.
- *Stage 2 (Competition Choice)*: Participants choose between a tournament and a piece-rate scheme. If this stage is selected for payment, earnings depend on this choice and on the Stage 1 evaluation. Before making any choice, participants are informed that tournament entrants are matched with participants from sessions conducted under the tournament condition, with all groups being gender-balanced.
- *Stage 3 (Belief Elicitation)*: Participants report their expected score in stage 1 and their expected rank when matched in gender balanced groups, independently of their choice of incentive scheme in Stage 2. Each correct guess is rewarded with 2 euros.

The aim of the present paper is to analyze human-AI interaction and its impact on creative performance. Hence, the analysis in this paper focuses on Stage 1, which isolates the effect of AI on creative performance. The impact of AI on competition choices (i.e. Stage 2) is examined in a separate study. Importantly, instructions for later stages were not disclosed during Stage 1.[3] This design ensures that behavior in the writing task is not influenced by anticipation of later decisions.

Experimental Procedures. The experiment was conducted at VERALabEx (Ca' Foscari University of Venice) with 360 participants, primarily students in economics-related programs. Eligibility required Italian as a first language. The sample size was chosen to align with similar experiments on human-AI interaction (e.g., Zhu and Zou, 2026; Chen and Chan, 2024; Noy and Zhang, 2023). Sessions included up to 20 participants and lasted about 50 minutes. Average earnings were 8.01 Euro, including a show-up fee of 3 Euro.

The experiment was implemented in *oTree* (Chen et al., 2016). Participants were randomly assigned to treatments and completed comprehension questions before the task.

After the experiment, participants completed a questionnaire including socio-demographics, AI exposure, cultural capital, and the Divergent Association Test (DAT) (Olson et al., 2021). Payments were determined after external evaluation, as explained in the following section, and transferred via *PayPal*.

2.2. **The Survey Experiment and the Evaluation of Creative Writings.** Creative texts produced in the laboratory experiment were evaluated by independent external raters recruited via the *Prolific* platform. Participation was restricted to individuals whose first language is Italian to ensure linguistic comparability.

[3]The English translation of the instructions for the *T&AI*-treatment is reported in Appendix A of the Online Appendix.

Each evaluator was asked to assess the creativity of four randomly assigned texts and received a payment of 3 Euro for a task lasting approximately five minutes.[4] Creativity was measured on a 1 to 10 scale. The main outcome variable is the average of the four evaluations assigned to each text.

Evaluators were blind to all treatment conditions. In particular, they were not informed about the incentive schemes or the availability of AI assistance in the laboratory experiment. They received only the original writing question and were informed that participants were encouraged to spend no more than 15 minutes on the task and to produce a text of approximately 150 words.

To control for order effects, the evaluation interface was designed so that each text appeared equally often in each presentation position (first through fourth). Each composition received four independent evaluations, one in each position.

After completing the evaluation task, raters filled out a short questionnaire collecting socio-demographic information. The survey experiment was implemented using the *Qualtrics* platform.

2.3. **Payment Procedures for Participants in the Laboratory Experiment.** After collecting the evaluations, participants in the laboratory experiment were grouped according to the rules of the incentive schemes and payments were computed accordingly. Participants were then invited via email to access a dedicated website where they could view their earnings.

To retrieve their payment, participants entered the anonymous identification code provided at the end of the laboratory session and supplied the email address associated with their *PayPal* account. Payments were subsequently processed by the administrative office.

3. Predictions and Conceptual Framework

We derive our predictions about human-AI collaboration and its effects on creative performance by grounding them in the relevant literature. A recent meta-analysis by Holzner et al. (2025) suggests that human-AI collaboration enhances creative performance relative to unaided individuals, although effects vary substantially across tasks. Moreover, studies focusing on open-ended writing tasks similar to ours find that AI augmentation improves creative output compared to individuals working without assistance (Bohren et al., 2024; Doshi and Hauser, 2024; Noy and Zhang, 2023). Taken together, this evidence motivates our first prediction:

> **Prediction 1 (AI enhances creativity):** Individuals with access to AI exhibit greater creative performance than those without access to AI.

We next examine the nature of human-AI collaboration and its relationship to performance along two

[4]The English translation of the instructions of the survey experiment is reported in Appendix A of the Online Appendix.

dimensions. First, the intensity of interaction with AI, measured by the number of queries. Second, the extent to which individuals rely on AI-generated content in crafting their final writing, captured through text-based similarity between AI responses and the final output, and thus reflecting the degree of AI involvement in shaping the final writing (Zhu and Zou, 2026).

Drawing on two online experiments in which participants were asked to develop a business plan, Huang (2025) documents an inverted U-shaped relationship between creative performance and AI usage, measured by the number of queries participants are allowed to make. Motivated by this evidence, we formulate the following prediction:

Prediction 2 (intensity of interaction with AI): the number of queries submitted to AI has a concave effect on creative performance such that, additional interactions increase creative output only up to a certain threshold.

In a recent study, (Zhu and Zou, 2026) show that more involvement with ChatGPT, measured using textual similarity between the final creative output submitted and the AI-generated responses, positively affects creative performance. Building on this evidence we formulate the following prediction:

Prediction 3 (Extent of AI involvement): greater similarity between AI-generated output during the human-AI interaction and the final creative output should have a positive effect on creative performance.

3.1. **The role of AI-assisted ideation: A mediation perspective.** To shed light on the mechanisms underlying the effect of AI on creative performance, we investigate how individuals interact with AI along the creative process. We conceptualize creativity as unfolding in two sequential stages: an ideation stage, in which individuals generate and select ideas, and an execution stage, in which they develop these ideas into a final output. We describe the conceptual framework here, and provide greater detail on how the variables were constructed through textual analysis in subsection 3.2.

Exploiting the sequential nature of the creative process, we estimate a mediation framework to assess how AI-assisted ideation contributes to the overall effect of AI collaboration on creative performance, both directly by generating more creative ideas, and indirectly by affecting the extent to which individuals rely on AI-generated content in developing their final output. We measure AI-assisted ideation using a binary indicator equal to one if a participant submits at least one idea request (*Idea request*), and a continuous counterpart capturing the intensity of such requests (*Idea request index*). Reliance on AI during the creative process is measured using a similarity metric between the final composition and AI-generated responses (*Similarity*), following Zhu and Zou (2026). Figure 2 illustrates the mediation model.

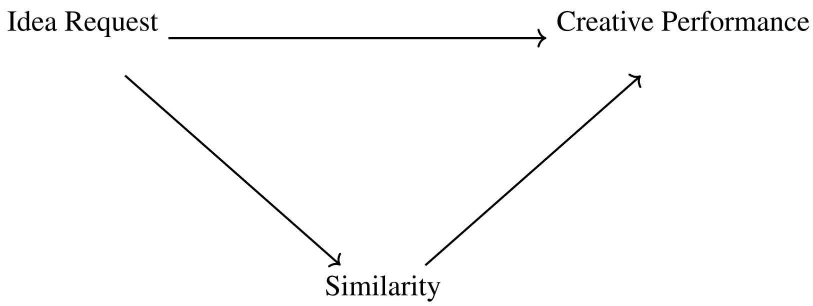

FIGURE 2. Mediation model: Similarity mediates the effect of Idea Request on Creative Performance.

Existing evidence provides only mixed guidance on how AI affects performance across these stages. A stream of literature suggests that idea generation can have a positive impact on some features of divergent thinking (Bangerl et al., 2024; Memmert et al., 2025). However, the dual process model of creativity describes the creative process as the interaction of divergent and convergent thinking (Guilford, 1957; Mednick, 1962), and both are relevant in determining creative performance (Beaty et al., 2015). Those studies that do consider the impact of AI-assisted idea generation on overall creative performance using textual analysis do not find evidence of significant effects (Chen and Chan, 2024; Zhu and Zou, 2026). In the absence of conclusive evidence on the role of AI assistance on creative performance, we do not formulate specific predictions on the impact of AI use across stages. Instead, we treat the analysis of these mechanisms as an empirical question, examining how AI-assisted ideation and reliance on AI during execution each contribute to creative performance. The novelty of our study is that we adopt mediation analysis to exploit the sequential nature of the creative process, while existing papers that employ textual analysis either exclude interaction between idea generation and execution by design (Chen and Chan, 2024), or use measures of AI-based idea generation that cannot be clearly disentangled from output execution (Zhu and Zou, 2026).

Finally, we also consider the role of participants' inherent creative potential, as measured by their scores on the Divergent Association Task (DAT). This allows us to investigate whether AI assistance can reduce the gap in creative performance between individuals of different creative ability, as suggested by Doshi and Hauser (2024). We also explore whether different payments schemes, that may or may not involve explicit rewards based on performance, may play a role.

3.2. **Mediation Analysis: details.** In what follows we provide more details on the construction of the variables used in the analysis, that is based on the recorded interaction logs between participants and the AI.

Idea request and Idea request index (LLM-based classification). We construct a binary indicator, *Idea request*, equal to one if a participant submits at least one query that constitutes a request for idea generation, and zero otherwise. We thus interpret *Idea request* as a measure of AI-supported ideation. Each non-empty query is evaluated using *Claude 4.6 Sonnet* API — namely, a deterministic LLM via an API-based classifier with a fixed prompt and temperature equal to zero. The classifier is instructed to distinguish between queries in which the participant explicitly asks the AI to generate an idea for the task, and queries in which the participant already has an idea and only seeks refinement, clarification, or stylistic improvement.

Importantly, the classification rule is designed using explicit role prompting and constraint setting. In particular, the model is instructed to act as a precise classifier in a scientific experiment and is constrained to return a structured JSON output with a binary label, a confidence score, and a short justification. The prompt further imposes strict decision rules that define what constitutes an ideation request versus a non-ideation query, thereby reducing ambiguity and ensuring consistency across classifications. Classification is based exclusively on the content of participant queries and does not rely on textual similarity between AI output and the final writing.

We extend our analysis by also building a continuous variable for ideation support. For each query, the classifier produces a continuous ideation score between 0 and 1 reflecting the strength of the idea-request signal. Then, for each participant, we compute *Idea request index*, defined as the maximum ideation score across all submitted queries. To convert this continuous measure into a binary indicator, we estimate a data-driven threshold using a two-component Gaussian mixture model applied to the distribution of participant-level maximum scores. Participants are classified as requesting an idea if at least one of their queries exceeds this threshold. The procedure is fully deterministic and reproducible; detailed implementation, together with the complete Python code, is provided in Appendix B of the Online Appendix.

Importantly, the two variables of idea request to the AI, *Idea request* and *Idea request index*, capture the observable decision to solicit idea generation from the AI, independently of whether the suggested idea is ultimately incorporated into the final composition. This distinction preserves the temporal ordering between upstream behavioral choices and downstream writing outcomes.

Similarity. We construct *Similarity*, defined as the cosine similarity between the participant's final writing and the concatenation of all AI-generated content received during the interaction. It is computed using TF-IDF vector representations, which weight informative words more heavily than common terms and yield a normalized measure ranging from zero to one. Higher values indicate greater textual overlap between AI-generated text and the participant's final composition. This measure captures the extent of AI involvement in the creative process: the degree to which AI-generated language contributes to shaping, structuring, or directly informing the participant's final writing. Unlike *Idea request*, which reflects an upstream behavioral decision, *Similarity* captures the overall extent of collaboration with AI throughout the process that leads to the final creative output.

4. RESULTS

4.1. **Descriptive statistics: a first look.** Table 1 reports descriptive statistics on the participants' average score, writing behavior, and measures of interaction with the AI.

A first pattern concerns creative performance, proxied by the average creativity score. A second pattern is the substantial heterogeneity in AI use, both in intensity and purpose. Although many participants in the AI treatments (i.e., between 67.7 and 81.8 percent) choose to interact with the AI during the task, the intensity of use varies widely: the relatively large standard deviations of *Number of queries filled* indicate that while some participants rely on the AI to formulate multiple queries and generate ideas, others make little or no use of the tool. Participants also differ in how they use the AI: some rely on it primarily for idea generation or inspiration, as captured by *Idea request* and *Idea request index*, while higher values of *Similarity* suggest that, in some cases, AI-generated content is incorporated more directly into final submissions. Understanding whether these different interaction patterns translate into differences in creative performance is one of the key questions addressed in the empirical analysis below.

The descriptive statistics also reveal considerable variation in writing behavior. Text length, measured by *Number of characters*, differs markedly across participants, indicating meaningful differences in the effort devoted to developing responses. This heterogeneity provides a useful setting to examine how incentives and AI interaction relate to creative performance.

Taken together, these descriptive statistics highlight three important features of the data that motivate the next subsections: the positive association between AI access and creative performance, substantial heterogeneity in both writing output and AI use, and the variety of ways in which participants combine their own effort with AI assistance.

TABLE 1. Descriptive statistics on creative performance and AI usage, by treatment

	T (1)	PR (2)	L (3)	T&AI (4)	PR&AI (5)	L&AI (6)
Performance and Subjective Variables						
Average score	5.615	5.681	5.816	6.281	6.000	5.972
	(1.317)	(1.089)	(1.311)	(1.363)	(1.363)	(1.329)
Number of characters	890.192	936.638	886.429	1000.227	976.839	982.968
	(259.377)	(386.247)	(298.735)	(272.121)	(280.282)	(245.820)
Male	0.500	0.362	0.367	0.500	0.387	0.339
	(0.505)	(0.486)	(0.487)	(0.503)	(0.491)	(0.477)
DAT score	83.705	82.582	84.233	83.518	83.535	83.323
	(6.323)	(5.231)	(5.198)	(4.834)	(5.070)	(4.832)
Books owned	3.000	2.979	3.082	2.909	3.097	2.855
	(1.066)	(1.093)	(0.975)	(1.002)	(0.936)	(1.143)
AI exposure	3.577	3.277	3.429	3.773	3.484	3.532
	(1.334)	(1.410)	(1.443)	(1.239)	(1.446)	(1.264)
Human-AI Interaction Variables						
AI (use)				0.818	0.677	0.790
				(0.388)	(0.471)	(0.410)
Number of queries filled				2.625	3.113	3.290
				(2.641)	(3.694)	(3.600)
Similarity				0.442	0.375	0.460
				(0.297)	(0.309)	(0.299)
Idea request				0.386	0.435	0.484
				(0.490)	(0.500)	(0.504)
Idea request index				0.415	0.432	0.465
				(0.403)	(0.439)	(0.419)
Obs.	52	47	49	88	62	62

Notes. This table reports descriptive statistics (standard deviations in parentheses) by treatment. Average score is the mean of the four scores assigned to each writing in the survey experiment. Number of characters is the length of the text measured in characters. DAT score, Books owned, and AI exposure are collected through the post-experimental questionnaire. Specifically, the DAT score is the Divergent Association Test (Olson et al., 2021), computed as the average semantic distance (ranging from 0 to 100) between the words provided by the participant, with higher values indicating greater divergent thinking ability. Books owned indicates the number of books owned by the participant at home (either in physical or electronic form) and is measured on a categorical scale: 0 = No books; 1 = Between 1 and 10 books; 2 = Between 11 and 50 books; 3 = Between 51 and 100 books; 4 = More than 100 books. AI exposure indicates how often the participant used AI tools in the last 12 months and is measured on a categorical scale: 0 = Never; 1 = Less than once per month; 2 = Once per month; 3 = Once per week; 4 = More than once per week; 5 = More than once per day. The human-AI interaction variables are constructed only for the AI treatments. AI (use) is a dummy equal to 1 if the participant sent at least one query to the AI, and Number of queries filled is the number of such queries. Similarity is the TF-IDF cosine similarity between the concatenated AI-generated texts and the participant's final writing. Idea request is a LLM-based classification dummy equal to 1 if the participant's queries contain at least one request for ideas to the AI. Idea request index is the maximum cosine similarity score between submitted queries and a reference ideation prompt, ranging from 0 to 1.

4.2. **AI access, AI use, and creative performance.** We begin our parametric analysis with Table 2, by examining whether access to generative AI affects creative performance.

Column (1) shows that allowing participants to access AI improves performance. The coefficient on AI access is positive and statistically significant, indicating that participants assigned to AI treatments receive higher creativity scores than those in the no-AI conditions. This provides initial evidence that generative AI can enhance performance in open creative tasks.

However, Column (2) shows that simple access to AI is not sufficient to generate these gains. Column (2) decomposes AI access into two distinct conditions: participants who had access to AI but did not interact with it, and participants who actively used AI during the task. The results show that the improvement in performance is entirely driven by active use. Participants who used the AI obtain significantly higher scores, whereas participants who had access to AI but did not submit any queries show no performance advantage relative to the no-AI group. This finding suggests that the performance gains associated with AI arise from actual human-AI interaction rather than from mere exposure to the technology. This is consistent with Prediction 1 and leads us to the following result:

Result 1. There is a positive effect of AI on creative performance which is entirely driven by active use rather than mere access.

Column (3) further examines how the intensity of AI interaction relates to performance, using *Number of queries filled* as a measure of usage intensity. The positive coefficient on the linear term and negative coefficient on its squared term, both highly significant, reveal a concave relationship between interaction intensity and creative performance, a result that is consistent with our Prediction 2 based on Huang (2025). According to a nonlinear test based on the coefficients of the two terms, the implied maximum occurs at approximately eight queries. This inverted-U pattern is inconsistent with a simple copy-and-paste mechanism, with participants merely asking the AI to generate an idea and directly copying the output into their final composition. Instead, it suggests that participants use the AI in an iterative manner, engaging in multiple exchanges that likely reflect exploration, refinement, and revision of ideas rather than mechanical copying.

Result 2. Increased intensity of interaction with AI, measured in number of queries, enhances creative performance up to a threshold value, beyond which this relationship becomes negative (inverted U shape).

Columns (4) and (5) estimate the same specification separately for the no-AI and AI samples, yielding two additional insights. First, the positive effect of writing length on performance disappears when AI is available. Only in the no-AI sample, *Number of characters* is strongly associated with higher scores. This suggests that, with AI support, writing length becomes less informative about the quality of the final output.

Second, only in the no-AI sample we do observe a significant inverted-U relationship between performance and baseline creative potential, measured by the DAT score. Both the linear and squared *DAT score*

TABLE 2. AI use, writing characteristics, and creative performance

	(1)	(2)	(3)	$AI = 0$ (4)	$AI = 1$ (5)
Number of characters	9.16×10^{-4}**	9.54×10^{-4}**	9.46×10^{-4}**	1.36×10^{-3}**	4.78×10^{-4}
	(3.05×10^{-4})	(2.97×10^{-4})	(2.94×10^{-4})	(3.98×10^{-4})	(3.82×10^{-4})
AI	0.299*		-0.163		
	(0.131)		(0.181)		
AI (use)		0.411**			
		(0.138)			
AI (no use)		-0.082			
		(0.231)			
Number of queries filled			0.269***		0.267***
			(0.063)		(0.062)
Number of queries filled^2			-0.017***		-0.017***
			(0.004)		(0.004)
Male	0.086	0.069	0.065	-0.008	0.118
	(0.138)	(0.137)	(0.134)	(0.182)	(0.188)
DAT score	0.606**	0.578**	0.597**	0.805***	-0.067
	(0.190)	(0.204)	(0.194)	(0.110)	(0.477)
DAT score^2	-0.004**	-0.003**	-0.004**	-0.005***	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
Constant	-20.049*	-18.933*	-19.710*	-28.237***	7.421
	(7.825)	(8.416)	(8.003)	(4.274)	(19.779)
F	5.10	5.02	6.55	18.10	3.80
$p > F$	0.000	0.000	0.000	0.000	0.001
Obs.	360	360	360	148	212

Notes. This table reports OLS estimates (robust standard errors in parentheses). The dependent variable is the mean of the four scores assigned to the participant's writing by the external evaluators in the survey experiment. Column (4) uses the No-AI sample ($AI = 0$), while Column (5) uses the AI sample ($AI = 1$). Number of queries filled^2 and DAT score^2 denote the squared terms of the corresponding variables. AI (no use) is a dummy that takes the value 1 if the participant belongs to an AI treatment but does not send any query to the AI. The other remarks of Table 1 apply. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

terms are highly significant only in Column (4), suggesting that AI assistance attenuates the relationship between individual baseline creative potential and creative performance.

4.3. **Further Exploring AI use: textual analysis and mediation approach.** WhileTable 2 establishes that active interaction with AI improves creative performance, it does not reveal *how* this improvement occurs. In the next analysis, we aim to shed light on the channels through which AI exerts its impact by estimating the mediation model presented in Section 3.

Table 3 reports the mediation analysis estimated through structural equation models (SEMs) on the sub-sample of participants with AI access. For each of the two ideation measures introduced in Section 3 (i.e., *Idea request* and *Idea request index*), we estimate two SEM specifications. The first is a parsimonious model that focuses on the core mediation pathway linking idea request, similarity, and creative performance. The second augments the model with the same set of controls used in Table 2, including writing length and baseline individual characteristics.

Across both specifications, we also account for AI interaction intensity using a residualized measure of the number of filled queries. This residualized query measure is constructed to address the fact that ideation support and query intensity are mechanically related: submitting an idea request necessarily increases the number of queries. To isolate variation in AI interaction that is not mechanically induced by ideation support, we first regress the number of filled queries on the ideation variable, either Idea request or Idea request index, and then retain the residual component. We include this residualized measure, together with its squared term, in the SEM. By construction, these variables capture variation in query intensity that is orthogonal to ideation support, allowing us to distinguish the effect of ideation-stage reliance from the effect of broader AI engagement during the task (for similar approaches, see: Wodtke and Zhou, 2020; Linden et al., 2021).

The mediator equation is therefore specified as a function of the ideation measure and the orthogonalized query controls, while the outcome equation includes the mediator, the ideation measure, the same orthogonalized query controls, and, in the richer specification, the additional covariates. This strategy allows us to net out general AI interaction intensity without conditioning directly on post-treatment variables that are mechanically generated by the ideation choice itself.

All models are estimated using SEM techniques by maximum likelihood. Because indirect effects are products of coefficients and may not follow a normal sampling distribution, statistical inference is based on nonparametric bootstrap standard errors computed from 2,000 replications. The mediation analysis is restricted to participants with AI access, since both the ideation measures and similarity outcomes are defined only for this group.

Panel A reports the mediator equation, where the dependent variable is the textual similarity between the participant's final writing and the AI responses. Focusing on the extended specification with the binary *idea request* measure, AI-supported ideation significantly increases reliance on AI-generated content. Participants who submit at least one ideation query exhibit substantially higher similarity between their final text and the AI responses, indicating that relying on AI in the ideation-stage translates into greater incorporation of AI-generated content in the development of the final output.

The mediator equation also reveals a systematic relationship between interaction intensity and Similarity. The residualized number of queries, which captures AI interaction intensity orthogonal to the ideation decision, has a positive and statistically significant effect on *Similarity*, while its squared term enters negatively. This pattern indicates diminishing marginal returns to additional interaction with the AI: increasing the number of queries leads participants to rely more heavily on AI-generated content, but the incremental

TABLE 3. SEMs: ideation and execution (restricted to participants in the AI treatments)

	Parsimonious			Extended			Parsimonious			Extended		
	Direct (1)	Indirect (2)	Total (3)	Direct (4)	Indirect (5)	Total (6)	Direct (7)	Indirect (8)	Total (9)	Direct (10)	Indirect (11)	Total (12)
Panel A: Mediator equation (Similarity)												
Idea request	0.308***			0.303***								
	(0.027)			(0.028)								
Idea request index							0.451***			0.446***		
							(0.033)			(0.034)		
Residualized number of queries	0.089***			0.089***			0.078***			0.078***		
	(0.007)			(0.007)			(0.008)			(0.007)		
Residualized number of queries^2	-0.007***			-0.007***			-0.006***			-0.006***		
	(0.001)			(0.001)			(0.001)			(0.001)		
Number of characters				5.88×10^{-5}						5.59×10^{-5}		
				(6.44×10^{-5})						(6.50×10^{-5})		
Male				0.018						0.013		
				(0.029)						(0.029)		
DAT score				0.045						0.031		
				(0.072)						(0.073)		
DAT score^2				-2.91×10^{-4}						-2.05×10^{-4}		
				(4.40×10^{-4})						(4.48×10^{-4})		
Constant	0.354***			-1.430			0.280***			-0.962		
	(0.022)			(2.954)			(0.023)			(3.002)		
Panel B: Outcome equation (Average score)												
Similarity	0.870*		0.870*	0.848*		0.848*	0.955*		0.955*	0.933*		0.933*
	(0.423)		(0.423)	(0.431)		(0.431)	(0.428)		(0.428)	(0.437)		(0.437)
Idea request	0.163	0.268*	0.431*	0.188	0.257	0.444*						
	(0.236)	(0.134)	(0.192)	(0.237)	(0.134)	(0.193)						
Idea request index							0.278	0.431*	0.709**	0.301	0.416*	0.718**
							(0.310)	(0.197)	(0.238)	(0.310)	(0.199)	(0.240)
Residualized number of queries	0.123*	0.078*	0.201***	0.122*	0.076	0.198***	0.109*	0.074*	0.183***	0.107*	0.073*	0.180***
	(0.057)	(0.038)	(0.045)	(0.056)	(0.039)	(0.045)	(0.054)	(0.034)	(0.044)	(0.054)	(0.034)	(0.044)
Residualized number of queries^2	-0.020**	-0.006	-0.025***	-0.020**	-0.006	-0.025***	-0.020**	-0.006*	-0.026***	-0.020**	-0.006*	-0.026***
	(0.006)	(0.003)	(0.006)	(0.006)	(0.003)	(0.006)	(0.006)	(0.003)	(0.006)	(0.007)	(0.003)	(0.006)
Number of characters				3.55×10^{-4}	4.99×10^{-5}	4.05×10^{-4}				3.42×10^{-4}	5.21×10^{-5}	3.94×10^{-4}
				(3.69×10^{-4})	(6.04×10^{-5})	(3.80×10^{-4})				(3.70×10^{-4})	(6.48×10^{-5})	(3.82×10^{-4})
Male				0.147	0.015	0.162				0.153	0.012	0.165
				(0.189)	(0.026)	(0.190)				(0.190)	(0.028)	(0.193)
DAT score				-0.108	0.038	-0.070				-0.124	0.029	-0.095
				(0.499)	(0.063)	(0.499)				(0.499)	(0.069)	(0.499)
DAT score^2				7.24×10^{-4}	-2.47×10^{-4}	4.77×10^{-4}				8.23×10^{-4}	-1.92×10^{-4}	6.32×10^{-4}
				(0.003)	(3.84×10^{-4})	(0.003)				(0.003)	(4.19×10^{-4})	(0.003)
Constant	5.842***		5.842***	9.390		9.390	5.742***		5.742***	9.943		9.943
	(0.193)		(0.193)	(20.730)		(20.730)	(0.181)		(0.181)	(20.690)		(20.690)
Obs.		212			212			212			212	

Notes. This table reports results of structural equation models restricted to treatments in which participants have access to AI ($AI = 1$). SEMs are estimated via maximum likelihood with bootstrap standard errors based on 2,000 replications. The mediator is Similarity. Columns report direct, indirect (through the mediator), and total effects. For mediator equations, indirect effects are not defined; total effects therefore coincide with direct effects. To account for the fact that ideation support mechanically increases the number of queries, the number of filled queries is first residualized with respect to the relevant ideation measure. The residualized component and its square are then included in the SEM, so as to capture variation in query intensity orthogonal to ideation support. The other remarks of Table 1 apply. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

effect declines as interaction intensity increases.

Panel B reports the outcome equation, where the dependent variable is the average creativity score. In the extended specification, Similarity has a positive and statistically significant effect on performance, and confirms our Prediction 3.

Result 3. The extent to which participants incorporate AI-generated language into their writing is positively associated with the creativity of the final output.

Once *Similarity* is taken into account, the direct effect of requesting ideas from AI on performance vanishes. By contrast, the indirect effect operating through *Similarity* is positive and statistically significant, while the total effect remains positive and statistically significant at conventional levels. Taken together, these results indicate that requesting ideas from the AI does not directly improve creative performance. Instead, its influence operates primarily through increased reliance on AI-generated content during the composition of the final writing. Moreover, AI use during the writing process (i.e., beyond the initial ideation decision) plays an independent role in enhancing creative performance, as reflected in the positive coefficient on Similarity.

Interaction intensity, measured with the number of queries submitted, also plays an independent role in explaining performance differences. The residualized number of queries has a positive effect on creativity scores, while its squared term is negative and statistically significant. This concave relationship suggests that moderate interaction with the AI improves creative outcomes, but additional queries yield progressively smaller benefits and may eventually generate diminishing returns.

When AI-supported ideation is measured using the continuous variable *Idea Request Index*, the results are consistent with those reported above. In the extended specification, higher ideation intensity significantly increases *Similarity*, and the indirect effect on performance through AI-generated content remains positive and statistically significant. As in the binary specification, there is no direct effect of ideation intensity on performance.

The extended specifications also include controls for writing length, interaction intensity, gender, and baseline ability measures. Consistent with the patterns observed in the earlier performance regressions, once reliance on AI-generated content and interaction intensity are accounted for, observable differences in writing effort or baseline characteristics play only a limited role in explaining performance differences among participants with AI access.

For completeness, the parsimonious specifications without controls yield very similar results. The positive indirect effect through AI-assisted writing, together with no direct effect of ideation support, is confirmed both when ideation is measured using the binary idea-request indicator and when the continuous ideation-intensity measure is used.

The mediation results provide clear evidence on the mechanism through which AI-supported ideation affects creative performance. While we detect no direct effect of *Idea Request* on performance, the total effect remains positive and significant, indicating that the impact of ideation support operates entirely through its influence on the extent to which participant choose to involve AI in the development of their final output. In particular, requesting ideas significantly increases the extent to which participants incorporate AI-generated content into their final composition, and this increased reliance is strongly associated with higher performance. The coefficient on *Similarity* captures this relationship conditional on the ideation decision, reflecting the contribution of AI-assisted textual incorporation rather than the act of requesting ideas per se. Importantly, this effect is identified by also considering participants who do not request ideas but still rely on AI during writing, implying that AI enhances performance not only by supporting ideation but also by facilitating the structuring, articulation, and refinement of self-generated ideas. Taken together, these findings suggest that the primary contribution of generative AI lies in supporting the development and expression of ideas, rather than directly improving idea generation itself.

From a behavioral perspective, the results suggest that generative AI enhances creative output primarily by supporting the articulation, structuring, and refinement of ideas, rather than by directly improving idea generation alone. In this sense, AI appears to function as a tool that facilitates the translation of partially formed thoughts into coherent and effective written content. This interpretation is consistent with the broader evidence in the paper, which shows that, in the presence of AI support, performance becomes less strongly associated with writing length and baseline ability. Taken together, these patterns indicate that AI tools can partially substitute for individual differences in writing-related skills, while leaving the underlying ideation process largely unchanged.

4.4. **Individual determinants of AI usage.** Having established that the performance gains associated with AI primarily depend on its effective use rather than mere access, we next investigate the individual determinants of AI usage. Table 4 examines how participants' characteristics predict several dimensions of AI interaction during the task.

Across all specifications, two variables emerge as the most consistent predictors of AI reliance: cultural capital and prior exposure to AI technologies. Cultural capital is proxied by the number of books owned by participants' families, a commonly used indicator of the intellectual environment in which individuals

TABLE 4. Individual characteristics, idea ownership, and AI-related behaviors (restricted to participants in the AI treatments)

	AI (use) (Probit AME) (1)	Number of queries filled (Oprobit) (2)	Similarity (OLS) (3)	Idea request (Probit AME) (4)	Idea request index (OLS) (5)
Male	0.028	0.084	0.006	0.037	0.054
	(0.060)	(0.152)	(0.040)	(0.068)	(0.059)
DAT score	0.218	0.250	0.068	0.335*	0.252
	(0.127)	(0.470)	(0.120)	(0.154)	(0.128)
DAT score^2	-0.001	-0.002	-0.000	-0.002*	-0.002*
	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
Number of books owned	-0.069*	-0.151*	-0.052**	-0.066*	-0.060*
	(0.028)	(0.074)	(0.019)	(0.033)	(0.030)
AI exposure	0.045*	0.202***	0.063***	0.045	0.044*
	(0.020)	(0.054)	(0.014)	(0.025)	(0.020)
Constant			-2.421		-9.639
			(5.000)		(5.294)
χ^2, F	16.31	21.60	7.59	16.70	4.77
$p > \chi^2$, $p > F$	0.006	0.001	0.000	0.005	0.000
Obs.	212	212	212	212	212

Notes: Sample restricted to participants with access to AI ($AI = 1$). Columns (1) and (4) report average marginal effects from Probit models (robust standard errors). Column (2) reports coefficients from an Ordered Probit model (robust standard errors). Columns (3) and (5) report OLS estimates with robust standard errors. The other remarks of Table 1 apply. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

develop their cognitive and cultural resources. We find that this variable is negatively associated with all measures of AI use. Participants with greater cultural capital tend to rely less on AI during the writing process. One interpretation is that individuals with richer intellectual resources may feel more confident generating and structuring ideas independently, reducing their need to seek external assistance from AI.

By contrast, prior exposure to AI technologies - measured by how frequently participants reported using AI tools in the previous twelve months - is positively associated with all dimensions of AI use. Participants who had interacted more frequently with AI before the experiment tend to rely more intensively on AI during the task. This pattern is consistent with a learning or familiarity mechanism: individuals who are more accustomed to interacting with AI may have developed more effective prompting strategies and a better understanding of how to integrate AI-generated suggestions into their workflow.

Taken together, these results suggest that the use of generative AI in creative tasks reflects both substitution and learning dynamics. On the one hand, individuals with stronger pre-existing intellectual resources appear less inclined to delegate parts of the creative process to AI, consistent with a substitution mechanism between human cultural capital and AI support. On the other hand, prior experience with AI increases reliance on the technology, highlighting the importance of familiarity and technological literacy in shaping adoption behavior.

More broadly, these patterns connect to the literature on the adoption of new technologies and skill-biased technological change. While many studies emphasize complementarities between digital technologies and human skills, our results suggest a more nuanced picture: AI use appears to be facilitated by technological familiarity but partially substitutes for certain forms of individual intellectual resources during the creative process. In this sense, generative AI may simultaneously lower barriers to creative production for some individuals while reducing the marginal value of pre-existing cultural capital in the execution stage of creative tasks.

4.5. **Incentive schemes, creative performance, and human–AI interaction.** Table 5 examines whether the incentive structure under which participants perform the creative task affects creative outcomes and patterns of human–AI interaction. The table reports regressions where the baseline category is the lottery scheme, so that the coefficients associated with the tournament and piece-rate schemes capture deviations from a condition in which remuneration is independent of performance.

Overall, the evidence suggests that the incentive structure has limited effects on both creative performance and patterns of interaction with the AI, a finding that is in line with Noy and Zhang (2023). In particular, neither the tournament nor the piece-rate scheme is associated with changes in the final score assigned to the text (column 1) or the DAT creativity score (column 3). Similarly, the amount of text produced, measured by the number of characters (column 2), does not vary across incentive schemes. From an economic perspective, these findings suggest that introducing stronger performance-based incentives does not substantially increase effort in this creative task. This result is consistent with the idea that creative performance may be relatively insensitive to marginal monetary incentives when tasks require open-ended ideation rather than routine effort.

We also find little evidence that incentive schemes systematically affect how participants interact with AI. Columns (4)–(8) show that the probability of using AI, the number of queries filled, the similarity between AI-generated content and the final text, and the likelihood of requesting ideas from AI do not substantially vary across the incentive schemes introduced in our experiment. This pattern suggests that the decision to rely on AI support is largely driven by individual strategies or task-related considerations rather than by the external incentive environment. In other words, participants do not appear to substitute AI support for effort more aggressively when facing stronger incentives, nor do they increase AI use to gain a competitive advantage.

TABLE 5. Effects of the incentive scheme: outcomes, writing behavior, and AI-related measures

	Score (OLS) (1)	Number of characters (OLS) (2)	DAT score (OLS) (3)	AI (use) (Probit AME) (4)	Number of queries filled (OLS) (5)	Similarity (OLS) (6)	Idea request (Probit AME) (7)	Idea request index (OLS) (8)
AI	0.398**	84.268**	-0.064					
	(0.139)	(31.587)	(0.577)					
Tournament	0.103	13.106	-0.133	0.028	-0.665	-0.018	-0.098	-0.049
	(0.170)	(33.999)	(0.664)	(0.066)	(0.537)	(0.049)	(0.082)	(0.068)
Piece Rate	-0.045	18.290	-0.600	-0.113	-0.177	-0.085	-0.048	-0.033
	(0.173)	(40.661)	(0.684)	(0.079)	(0.654)	(0.055)	(0.090)	(0.077)
Constant	5.681***	893.283***	83.760***		3.290***	0.460***		0.465***
	(0.147)	(32.489)	(0.583)		(0.457)	(0.038)		(0.053)
χ^2, F	3.26	2.54	0.28	4.13	0.94	1.36	1.42	0.26
$p > \chi^2$, $p > F$	0.022	0.056	0.839	0.127	0.394	0.259	0.493	0.770
Obs.	360	360	360	212	212	212	212	212

Notes: Robust standard errors in parentheses. Columns (1)–(3) use the full sample. Columns (4)–(8) restrict the sample to participants with access to AI ($AI = 1$). Columns (4) and (7) report average marginal effects (discrete changes) from Probit models estimated with robust standard errors. Tournament takes the value 1 in treatments based on the tournament incentive scheme. Piece Rate takes the value 1 in treatments based on the piece rate incentive scheme. The omitted baseline incentive scheme category is Lottery. The other remarks of Table 1 apply. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

5. Discussion and concluding remarks

In this paper, we investigated how individuals interact with generative AI while performing a creative task, and how the interaction affects performance. To address these questions, we designed an incentivized experiment in which participants were asked to write a short creative composition, later evaluated by external judges. Beyond finding evidence that AI acts as a complement to human creativity, we explored the features of the human-AI collaboration in more detail. Our results indicate that even if excessive interaction with AI can be detrimental, greater reliance on AI generated content enhances the creativity of output. In this final section, we briefly comment on how our results relate to those of the relevant literature, and contribute to extending our understanding of human-AI collaboration in creative processes, as well as suggesting avenues for future research.

Recent experimental studies related to our work have examined the impact of access to generative AI on performance in creative tasks, including writing advertisements (Chen and Chan, 2024), business plans (Huang, 2025), and short stories (Bohren et al., 2024; Doshi and Hauser, 2024). These experiments were conducted entirely online, while in our setting participants performed the task in the laboratory, providing a more controlled environment. However, the result that AI assistance enhances creative performance is consistent across different experimental settings, suggesting that this finding is sufficiently general.

Additionally, while Bohren et al. (2024) and Doshi and Hauser (2024) focus on performance and related outcomes, such as the diversity of ideas generated with AI relative to those generated by humans alone, our focus is on how AI is used and where its assistance contributes most along the creative process. This focus on the mechanisms underlying performance enhancement is also shared by Chen and Chan (2024) and Huang (2025). In both studies, participants are allowed to use AI only in specific ways: in Huang (2025), through constraints on the number of queries they can submit; in Chen and Chan (2024), through predefined roles for AI (e.g., ghostwriter or sounding board). These designs allow the authors to cleanly disentangle the effects of different usage modalities on performance and to study the associated behavioral responses. In contrast, our unconstrained setting allows us to examine how individuals interact with AI when they are free to use it as they wish, as is typical in real-world contexts. Considering these fundamental differences, we comment on how our results can be seen as complements to these studies, contributing to represent a more complete picture of the features of human-AI interaction. In particular, Chen and Chan (2024) find that if AI assistance is restricted to ghostwriting, that is, generating a final output that incorporates both an idea and its finalization, this does not have a positive effect on creative performance. Instead, Huang (2025) finds that AI use (if moderate) does produce a positive effect on creativity through greater idea diversity, when ideation is decoupled from finalization. Our results go one step further by shedding light on the relationship between ideation and creativity. Indeed, we show that the request for AI ideas plays a positive role in creativity, only if it is considered as a part of an interactive process. Therefore, idea generation through AI can be seen as a valuable part of a wider creative process, even if requesting ideas from AI does not produce a direct effect on creative performance.

Finally, we comment on how our results compare with those of Zhu and Zou (2026), a closely related paper that also allows subjects to freely interact with AI in a creative task. In a similar laboratory experimental setting they construct a measure of AI-assisted idea generation that considers the number of keywords first introduced by AI that are part of the final creative output, versus the number of keywords first introduced by a human participant, and find that it is associated with lower creative performance. The authors interpret these findings as suggesting that creativity may decline when AI plays a leading role in generating the core ideas.

Our approach differs from this methodology because we distinguish between upstream ideation requests and downstream incorporation of AI-generated content during the writing stage. We do this by measuring whether subjects requested help from AI in generating ideas, regardless of whether they used these suggestions in their final write-up. This part of the creative process clearly precedes the writing of the submitted creative output, and allows us to use a structural mediation framework, since our idea request can either affect creative performance directly or indirectly as part of a finalization process. More specifically, we measure ideation support directly from participants' prompts rather than inferring it from the content of the final text. In contrast, the AI-first measure is mechanically related to the final written output, which introduces a conceptual overlap with *Similarity*, since both are derived from the final text.[5]

Our findings on the heterogeneity in AI reliance suggest several directions for future research. In particular, the result that individuals with higher cultural capital rely less on AI-generated content may reflect greater confidence in one's own abilities or differences in cognitive style, such as a preference for independent rather than assisted problem solving, which reduces reliance on external support. Future research could use experimental design to disentangle whether individuals with higher cultural capital are more likely to critically evaluate and selectively incorporate AI-generated content.

Similarly, our finding that individuals with greater prior exposure to AI rely on it more extensively may reflect simple psychological mechanisms, such as greater trust, increased comfort, or the formation of habits over time. At the same time, repeated use could make individuals less likely to question AI outputs and more prone to accept them at face value. Future research could examine whether familiarity leads to more

[5]In Table 1 of the Online Appendix (Appendix C), we show that when substituting *Idea request* with the AI-first variable introduced by Zhu and Zou (2026) in our mediation analysis, our results on the positive effects of AI-assisted ideation on creative performance continue to hold.

effective use or instead fosters a tendency toward uncritical reliance.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process. During the preparation of this work, the authors used the *ChatGPT-4.2* and *Claude-4.6 Sonnet* APIs. The former was integrated into the *oTree* interface as a writing support tool for participants in the laboratory experiment. The latter was used during the analysis phase to classify participants' queries (see the subsection on API-based classification). Additional minor editorial assistance, limited to text refinement, was provided using *ChatGPT-5.2*. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgements: We thank Thomas Bassetti, Maria Bigoni, and all participants at the workshop "Experimental Approaches to Gender Inequality: Stereotypes, Norms, and Women's Achievements" in Bologna for useful suggestions and comments. We are in debt with Andrea Albarea, Chiara Grigoletti, and all the staff of the VERALabEx for helping run the experiment.

REFERENCES

Bangerl, M. M., K. Stefan, and V. Pammer-Schindler (2024). Explorations in human vs. generative AI creative performances: A study on human-AI creative potential. *Proceedings of the TREW Workshop for Trust and Reliance in Evolving Human-AI Workflows at CHI EA 24*.

Beaty, R., M. Benedek, S. Kaufman, and P. Silva (2015). Executive network coupling supports creative idea production. *Scientific Reports 5*(10964).

Bervar, M., T. Bertoncel, and M. Pejić Bach (2026). Generative artificial intelligence and the creative industries: A bibliometric review and research agenda. *Systems 14*(2), 138.

Bohren, N., R. Hakimov, and R. Lalive (2024). Creative and strategic capabilities of generative AI: Evidence from large-scale experiments. Technical report, IZA Institute of Labor Economics.

Charness, G. and D. Grieco (2019). Creativity and incentives. *Journal of the European Economic Association 17*(2), 454–496.

Chen and Chan (2024). Large language model in creative work: the role of collaboration modality and user expertise. *Management Science 70*(12), 9101–9117.

Chen, D. L., M. Schonger, and C. Wickens (2016). oTree — An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

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

Epstein, Z., A. Hertzmann, the Investigators of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, A. S. Pentland, O. Russakovsky, H. Schroeder, and A. Smith (2023). Art and the science of generative AI. *Science 380*(6650), 1110–1111.

Erat, S. and U. Gneezy (2016). Incentives for creativity. *Experimental Economics 19*(2), 269–280.

Gero, K. I., T. Long, and L. B. Chilton (2023). Social dynamics of AI support in creative writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Guilford, J. (1957). Creative abilities in the arts. *Psychological Review 64*(2), 110–118.

Heigl, R. (2025). Generative artificial intelligence in creative contexts: a systematic review and future research agenda. *Management Review Quarterly 76*(1), 955–992.

Holzner, N., S. Maier, and S. Feuerriegel (2025). Generative AI and creativity: a systematic literature review and meta-analysis. *arXiv preprint arXiv:2505.17241*.

Huang, H.-C. B. (2025). Unlocking creativity with artificial intelligence (AI): Field and experimental evidence on the goldilocks (curvilinear) effect of human–AI collaboration. *Journal of Experimental Psychology: General 154*(12), 3294–3306.

Krakowski, S. (2025). Human-AI agency in the age of generative AI. *Information and Organization 35*(1), 100560.

Linden, A., C. Huber, and G. T. Wodtke (2021). A regression-with-residuals method for analyzing causal mediation: The rwrmed package. *The Stata Journal 21*(3), 559–574.

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review 69*(3), 220–232.

Memmert, L., I. Cvetkovic, and N. Tavanapour (2025). Brainstorming with a generative language model: Effect of exposure to AI ideas on brainstorming performance and cognitive load. *Bus Inf Syst Eng* (3).

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? Do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*, 187–192.

Olson, J. A., J. Nahas, D. Chmoulevitch, S. J. Cropper, and M. E. Webb (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences 118*(25), e2022340118.

Sun, Y., E. Jang, F. Ma, and T. Wang (2024). Generative AI in the wild: Prospects, challenges, and strategies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*.

Tsao, J., C. X. Liang, C. Nogues, and A. Wong (2025). Perceptions and integration of generative artificial intelligence in creative practices and industries: A scoping review and conceptual model. *AI & Society 92*, 7–22.

Wodtke, G. T. and X. Zhou (2020). Effect decomposition in the presence of treatment-induced confounding: a regression-with-residuals approach. *Epidemiology 31*(3), 369–375.

Yuan, A., A. Coenen, E. Reif, and D. Ippolito (2022). Wordcraft: Story writing with large language models. In *Proceedings of the ACM Conference on Intelligent User Interfaces*.

Zhu, F. and W. Zou (2026). Generative AI adoption in human creative tasks: Experimental evidence. *Journal of Economic Behavior & Organization 242*, 107414.

ONLINE APPENDIX. HUMAN-AI INTERACTION IN CREATIVE TASKS: AN EXPERIMENTAL INVESTIGATION

FEDERICO ATZORI, LUCA CORAZZINI, VALERIA MAGGIAN, FILIPPO PAVESI, AND MASSIMO SCOTTI

ABSTRACT. This online appendix contains supplementary material referenced in the main paper. Specifically, it includes: (i) the experimental instructions for both the laboratory and the survey experiments (Appendix A); (ii) detailed explanations, along with the Python code, of how the main AI-related measures used in the analysis, i.e., *Idea request*, *Idea request index*, and *Similarity*, have been constructed (Appendix B); and (iii) additional results obtained by replicating the analysis in Section 4.3 of the paper based on the behavioral indicators of AI reliance introduced by Zhu and Zou (2026), i.e., *AI-first keywords* and *User-first keywords* (Appendix C).

APPENDIX A. EXPERIMENTAL INSTRUCTIONS

The instructions were originally written in Italian. The instructions for the laboratory experiment were distributed at the beginning of each phase and read aloud, while the instructions for the survey experiment were provided on the first page of the Qualtrics platform.

A.1. Instructions for the Laboratory Experiment.

General Instructions. Thank you for participating in this experiment! During the experiment, you are not allowed to speak or communicate in any way with other participants. For any questions, click on the lifebuoy-shaped button at the top right of the screen; we will answer your questions privately. The following rules are the same for all participants.

A total of 20 people are participating in this experiment. The choices you make during the experiment are anonymous: other participants will not know your choices either during or after the experiment. Similarly, you will not be informed of the identity of the other participants, nor will you know the choices they made during the experiment.

The experiment consists of 3 consecutive phases: Phase 1, Phase 2, and Phase 3. You will shortly receive

Date: **February 2026**.
Key words and phrases. Human-AI Interaction; Creativity; Generative AI; Laboratory Experiment.
JEL: C91; D83; J24; O33.

Atzori (fatzori19@gmail.com): CIMEO, Sapienza University, Rome, Italy. **Corazzini** (luca.corazzini@unimib.it): University of Milan - Bicocca, Italy. **Maggian** (valeria.maggian@unive.it): "Ca' Foscari" University of Venice, Italy. **Pavesi** (fpavesi@liuc.it): LIUC University, Castellanza, Italy. **Scotti** (mscotti@liuc.it): LIUC University, Castellanza, Italy.

the instructions for Phase 1. The instructions for the remaining two phases will be provided at the beginning of the corresponding phase.

Although you will participate in three phases, your payment at the end of the experimental session will be determined by adding your earnings from Phase 3 to those obtained in only one of the first two phases. Specifically, at the end of the experiment, we will randomly select which of the first two phases will be used to determine your final earnings by flipping a coin.

In addition to the earnings you obtain during the experiment, you will receive €3 for arriving on time. Your final earnings will be credited to your PayPal account.

Before your earnings are credited to your PayPal account, you will receive an email inviting you to access a webpage where you will find information about your results in the experiment and your earnings.

Instructions for Phase 1.

What do you have to do in Phase 1? In Phase 1, you are asked to write a creative composition based on the following prompt:

"If you had the ability to create something simply with your mind, what would you create?"

We suggest that you do not exceed 150 words and do not spend more than 15 minutes completing your composition. These suggestions are purely indicative. Exceeding 150 words or spending more than 15 minutes will have no effect on the final outcome of Phase 1. To assist you, the computer will display the remaining time (in minutes) as well as a word count.

On the screen where you are asked to write your composition, you will find a box that allows you to access a generative artificial intelligence application (ChatGPT). Using this box, you may send any requests you wish, without restrictions on their form or number. The possibility to use ChatGPT will end when you complete your composition and proceed to the next phase.

If you decide to send requests to ChatGPT, we suggest including indications about the theme or context of your creative composition.

The questions you send to ChatGPT and the corresponding responses will not be saved nor used to train or improve OpenAI models. Data sent through this interface will not be accessible to OpenAI developers, in accordance with the platform's Data Usage Policy. For further information, please contact the researchers involved in this experiment.

When writing your composition and sending requests to ChatGPT, please avoid including personal information or any elements that could reveal your identity.

Your creative composition will receive 4 independent and anonymous evaluations from external evaluators. At the end of this session, your composition will be sent to 4 evaluators randomly selected from approximately 2000 Italian-speaking individuals (male or female) registered on the Prolific platform.

Each evaluator will be asked to carefully read your composition and answer the following question:

> "In general, from 0 to 10, where 0 is the minimum and 10 is the maximum, how do you rate the level of creativity of this composition?"

You will not receive any information about the identity of the evaluators. Similarly, evaluators will not receive any information about your identity or about the rules determining earnings. The only information provided to evaluators will be the prompt of the composition.

How are earnings determined in Phase 1? At the beginning of Phase 1, the computer will randomly assign you, with equal probability, to a group of 4 participants (2 males and 2 females). You will not know the identity or the composition of the other group members.

Your earnings in Phase 1 will be determined by a competitive scheme comparing the average rating of your composition with those of the other 3 group members. Specifically, for each composition, the arithmetic mean of the 4 evaluations will be calculated. Based on these averages, a ranking will be established.

If your composition receives the highest average rating in the group, your earnings (in euros) will be:

$$2.4 \times \text{Average rating}$$

Otherwise, you will receive no earnings. In case of a tie, the winner will be randomly determined among those with the same average rating.

If Phase 1 is selected for payment, the webpage you access before receiving payment will show:

- Your average rating
- Whether your composition ranked first in your group

Instructions for Phase 2.

What do you have to do and how are earnings determined in Phase 2? In Phase 2, you must choose between two payment schemes for your composition written in Phase 1:

- **Non-competitive scheme**: Your earnings will be:

$$0.6 \times \text{Average rating}$$

- **Competitive scheme**: Same as in Phase 1. If your composition has the highest average rating:

$$2.4 \times \text{Average rating}$$

  Otherwise, you receive nothing.

In case of a tie, the winner is randomly selected.

Your earnings in Phase 2 will depend on the scheme you choose.

If Phase 2 is selected for payment, the webpage will show:

- Your average rating
- (If competitive) whether your composition ranked first

Instructions for Phase 3.

What do you have to do in Phase 3? In Phase 3, you must make 2 guesses:

- Guess your rank (1st, 2nd, 3rd, or 4th) within your group based on average ratings
- Guess the integer (from 0 to 10) closest to your average rating

How are earnings determined in Phase 3? For each correct guess, you will earn €2.

The webpage you access before receiving payment will provide information on the correctness of your guesses.

A.2. **Instructions for the Survey Experiment.**

Instructions. Thank you for participating in this experiment! During the experiment, you will be asked to make simple choices and complete a short questionnaire. All the choices you make during the experiment and the information you provide in the questionnaire will remain completely anonymous and will be analyzed by the researchers involved in the study in aggregate form for scientific research purposes.

The expected duration of the experiment, including the time required to complete the questionnaire, is approximately 15 minutes. For participating in this experiment, you will receive a payment of £3 (British pounds). Payment will be made through the Prolific platform. For this reason, we ask you to keep your Prolific ID code at hand. Payment will only be granted if you complete the experiment and the final questionnaire in full.

For further information about the experimental study, please contact the researchers involved in the research project:

Luca Corazzini, University of Milano-Bicocca, luca.corazzini@unimib.it;
Valeria Maggian, Ca' Foscari University of Venice, valeria.maggian@unive.it;
Filippo Pavesi, LIUC - Carlo Cattaneo University, fpavesi@liuc.it;
Massimo Scotti, LIUC - Carlo Cattaneo University, mscotti@liuc.it;
Federico Atzori, University of Milano-Bicocca, federico.atzori@unimib.it

Experimental Instructions.

What do you have to do? You are asked to evaluate the level of creativity of 4 creative compositions produced by participants in a laboratory experiment conducted at the VeraLabEx experimental economics laboratory of Ca' Foscari University of Venice.

In particular, participants in the laboratory experiment were asked to write a creative composition based on the following prompt:

> "If you had the ability to create something simply with your mind, what would you create?"

For each creative composition presented to you, you are asked to provide your evaluation based on the following question:

> "In general, from 0 to 10, where 0 is the minimum and 10 is the maximum, how do you rate the level of creativity of this composition?"

There are no right or wrong answers to the evaluation task. We ask you to carefully read each composition and provide the evaluation that best reflects your opinion of its level of creativity.

In addition to providing your evaluation, for each of the 4 compositions you are asked to indicate 3 keywords that, in your opinion, best describe the content of the composition.

Additional Information on the Creative Compositions and the Study. Participants in the laboratory experiment were advised not to exceed 150 words and not to spend more than 15 minutes completing their composition. However, they were informed that exceeding 150 words or spending more than 15 minutes would have no effect on the outcome of their participation in the experiment.

As indicated in the previous instructions, participants in the laboratory experiment will not receive any information about your identity. Similarly, you will not receive any information about the identity of the 4 participants whose compositions you will evaluate.

Appendix B. Construction of AI Involvement Measures

This appendix describes the code used to construct the measures of *Idea request*, *Idea request index*, and *Similarity* employed in the empirical analysis. The code is implemented in Python and operates directly on the original experimental dataset containing participant writing and complete dialogue logs.

B.1. Overview of the Procedure.

The code reconstructs each participant's interaction with the AI using the ordered sequence of participant queries (`query_1` through `query_19`) and corresponding AI responses (`chat_response_1` through `chat_response_19`). Text fields may be empty. In the code, text is minimally normalized by trimming whitespace and converting missing values to empty strings.

Based on these data, the procedure constructs measures corresponding to two sequential stages of creative production: upstream idea request and downstream AI-assisted composition. Idea request is measured using API-based classification of participant queries, while AI-assisted composition is measured using semantic similarity between participant writing and AI-generated content.

All constructed variables are deterministic functions of the observed dialogue logs, the final writing, and a fixed API prompt and model specification. Given the same input dataset and API environment, the code reproduces the AI involvement measures.

B.2. API-Based Classification of Idea Request.

To identify idea requests, the code evaluates each non-empty participant query using a deterministic large language model classifier accessed via API. The classifier is implemented with a fixed model version, a fixed system prompt, and temperature equal to zero. For each query, the classifier returns:

- a continuous ideation score between 0 and 1, reflecting the strength of the idea-request signal, and
- a binary label indicating whether the query constitutes a request for idea generation.

The classifier is instructed to distinguish between queries in which the participant explicitly asks the AI to generate or suggest an idea for the task, and queries in which the participant already has an idea and requests refinement, clarification, or stylistic assistance. Classification relies exclusively on participant query text and does not use information from the final written composition.

For each participant, the code computes *Idea request index* (`ideation_api_score_max` in the code), defined as the maximum ideation score observed across all submitted queries.

B.3. **Threshold Estimation via Gaussian Mixture Model.** To convert the continuous ideation score into a binary measure, the code estimates a data-driven threshold using a two-component Gaussian mixture model (GMM) applied to the distribution of participant-level maximum scores.

The threshold is defined as the intersection point of the two estimated Gaussian components. Participants are classified as having submitted an idea request if at least one of their queries exceeds this estimated threshold. If the mixture model cannot be reliably estimated (e.g., due to insufficient observations or lack of bimodality), a conservative fallback rule is applied. The GMM is estimated with a fixed random seed to ensure deterministic results.

The resulting binary indicator is *Idea request* (`idea_request` in the code). This variable captures the participant's observable decision to solicit idea generation from the AI and is constructed independently of any similarity-based measures.

In addition, the code constructs the number of non-empty queries sent by the participant (`n_query_filled` in the code), providing a transparent measure of interaction intensity.

B.4. **Similarity.** The code measures downstream AI-assisted composition using TF–IDF vector representations combined with cosine similarity.

For each participant, TF–IDF vectors are constructed pairwise on (i) the participant's final writing and (ii) the concatenation of all non-empty AI responses observed during the interaction. Cosine similarity between these two vectors yields a normalized measure ranging from zero to one.

The resulting variable, *Similarity* (`sim_writing_ai_concat` in the code), captures the semantic dependence of the participant's final composition on AI-generated content during the execution stage.

Importantly, this similarity measure is constructed independently of the API-based idea-request classification. As a result, *Idea request* and *Similarity* capture distinct and sequential stages of the creative process.

B.5. **Output Files.** The code generates the following variables and files:

- API-based variables concerning human-AIinteraction:
  - `idea_request`
  - `ideation_api_score_max`

- `sim_writing_ai_concat`
- `n_query_filled`

- `ideation_audit_api.csv`, which records, for each AI-access participant, the number of queries, maximum ideation score, per-query classification outcomes, and a preview of the final writing. This file provides full transparency and allows independent verification of the API-based classification.

The output dataset is saved in STATA format and can be opened directly without additional processing.

B.6. Code to build behavioral measures of AI involvement in creative writing tasks.

```
"""
This code builds the following behavioral measures:

  - idea_request
  - ideation_api_score_max
  - n_query_filled
  - sim_writing_ai_concat

Ideation classification is performed via the Claude API using:
  - a fixed model
  - temperature = 0 (deterministic)
  - a fixed prompt
  - robust response parsing

The threshold for idea_request is inferred using:
  - a two-component Gaussian Mixture Model (GMM)
  - a fallback rule when the GMM is not reliable:
      - the 75th percentile of observed scores if the sample is small,
      - 0.50 if no valid scores are available.

The execution measure is defined as:
  - TF-IDF cosine similarity between the participant's final writing
    and the concatenation of all AI responses received.

Output:
  1) A .dta dataset with the four variables added to the original dataset
  2) ideation_audit_api.csv (complete log of API classifications)
```

```
The code used to construct the variables has NOT been modified
relative to the original version; only sections related to
other non-essential variables have been removed for this
replication package.

Due to LaTeX text compatibility issues, Italian accented characters are
replaced with "A'", "a'", "E'", "e'", "I'", "i'", "O'", "o'", "U'", and "u'".
"""

# ==============================================================================
# IMPORTS
# ==============================================================================
import json
import math
import os
import time
import unicodedata
import warnings
from pathlib import Path

import anthropic
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.mixture import GaussianMixture

warnings.filterwarnings("ignore", category=UserWarning)


# ==============================================================================
# CONFIGURATION
# ==============================================================================
# Replication setup:
# Place this script and the input dataset in the same folder.
# Paths below are defined relative to the script location for portability.
```

```
REPLICATION_DIR = Path(__file__).resolve().parent

INPUT_PATH = REPLICATION_DIR / "Final Dataset Human AI creativity, ACMPS 2026.dta"
OUTPUT_PATH = REPLICATION_DIR / "Final Dataset Human AI creativity, ACMPS 2026 -
    Textual.dta"
AUDIT_CSV = REPLICATION_DIR / "ideation_audit_api.csv"

# Maximum number of turns/queries in the dataset
MAX_TURNS = 19

# Treatments with AI access (API ideation classification is run only for these
    subjects)
AI_ACCESS_TREATMENTS = {"T_AI", "PR_AI", "LOT_AI"}

# ========= API Model =====================================================
# claude-sonnet-4-6 : ~$1.00 for the entire dataset, high quality
# claude-opus-4-6 : ~$4.70, max quality
API_MODEL = "claude-sonnet-4-6"
API_TEMPERATURE = 0 # deterministic
API_MAX_TOKENS = 150 # sufficien for JSON
API_RETRY_MAX = 3 # attempts in case of API error
API_RETRY_WAIT = 5 # waiting seconds between attempts

# ========= Experimental context ==========================================
# Prepended to each query: provides the model with the experimental frame
# needed to correctly interpret each participant request.
# The prompts are in Italian to match the experimental language.
CONTEXT = (
   "Contesto: un partecipante a un esperimento deve scrivere un "
   "componimento creativo di circa 150 parole rispondendo alla traccia: "
   "\"Se avessi il talento di creare qualcosa semplicemente con il "
   "pensiero, cosa creeresti?\". Il componimento verra’ valutato in base "
   "al livello di creativita’ da 4 giudici esterni. Il partecipante ha "
   "accesso a ChatGPT durante il compito. "
   "La seguente e’ una query che il partecipante ha inviato a ChatGPT: "
)
```

```
# ========= System prompt =========================================
# Key discriminant: who has the idea - the participant or ChatGPT?
# Validated on a sample of 26 real queries: 26/26 correctly classified (100%).
# The prompts are in Italian to match the experimental language.
SYSTEM_PROMPT = """Sei un classificatore preciso per un esperimento scientifico.
Il tuo compito e' classificare query inviate a ChatGPT da partecipanti a un
    esperimento di scrittura creativa.

Devi rispondere SOLO con un oggetto JSON con questa struttura:
{"score": <float tra 0.0 e 1.0>, "label": <0 o 1>, "reason": "<spiegazione breve in
    italiano>"}

Dove:
  label=1 (ideation request): il soggetto NON ha ancora un'idea per il componimento e
      chiede a ChatGPT di suggerirgliene una
  label=0 (not ideation): il soggetto HA GIA' un'idea o bozza propria e chiede aiuto
      per svilupparla, correggerla o migliorarla; OPPURE fa una domanda generica non
      legata al compito creativo

Lo score deve riflettere la tua certezza: 1.0 = certamente ideation request, 0.0 =
    certamente not ideation.
Non aggiungere nulla oltre al JSON."""

# ========= Fallback threshold ==========================================
# Used only if the GMM cannot be estimated or is not reliable.
API_THRESHOLD_FALLBACK = 0.50



# ===============================================================================
# UTILITY
# ===============================================================================
def normalize(x) -> str:
   """Normalize string values from the dataset (NaN/None/.) --> ''."""
   if x is None:
      return ""
   if isinstance(x, float) and np.isnan(x):
      return ""
```

```
    s = unicodedata.normalize("NFC", str(x)).replace("\u00a0", " ").strip()
    return "" if s.lower() in {"nan", "none", "."} else s



# =================================================================================
# CLAUDE API - QUERY CLASSIFICATION
# =================================================================================
def parse_api_response(raw: str) -> dict | None:
    """
    Robust parsing of the API response.

    Attempts, in order:
    1) json.loads on the cleaned text
    2) Field-by-field regex extraction (robust to apostrophes/quotes inside the reason
        field)

    Returns a dict {score, label, reason} or None if all attempts fail.
    """
    import re

    # Attempt 1 - standard json.loads
    try:
        clean = raw.replace("```json", "").replace("```", "").strip()
        result = json.loads(clean)
        return {
            "score": float(result.get("score", 0.5)),
            "label": int(result.get("label", 0)),
            "reason": str(result.get("reason", "")),
        }
    except (json.JSONDecodeError, ValueError):
        pass

    # Attempt 2 - field-by-field regex parsing (robust to problematic characters in the
        reason field)
    score_m = re.search(r'"score"\s*:\s*([0-9]*\.?[0-9]+)', raw)
    label_m = re.search(r'"label"\s*:\s*([01])', raw)
    reason_m = re.search(r'"reason"\s*:\s*"(.*?)"(?:\s*[,}])', raw, re.DOTALL)
```

```
    if score_m and label_m:
        return {
            "score": float(score_m.group(1)),
            "label": int(label_m.group(1)),
            "reason": reason_m.group(1).strip() if reason_m else "",
        }

    return None


def classify_query(client: anthropic.Anthropic, query: str) -> dict:
    """
    Classifies a query via the Claude API.

    Returns a dict: {score: float, label: int, reason: str}

    Error handling:
    - Automatic retries up to API_RETRY_MAX times
    - Two-step robust parsing (json.loads --> regex)
    - Fallback to score=0.5, label=0 if all attempts fail
    """
    user_message = CONTEXT + query

    for attempt in range(1, API_RETRY_MAX + 1):
        try:
            response = client.messages.create(
                model=API_MODEL,
                max_tokens=API_MAX_TOKENS,
                temperature=API_TEMPERATURE,
                system=SYSTEM_PROMPT,
                messages=[{"role": "user", "content": user_message}],
            )
            raw = response.content[0].text.strip()

            result = parse_api_response(raw)
            if result is not None:
```

```
            return result

        # Parsing failed - log-in and try again
        print(f" ! Parsing failed (attempt {attempt}): {raw[:80]}")
        if attempt < API_RETRY_MAX:
            time.sleep(API_RETRY_WAIT)

    except anthropic.RateLimitError:
        print(f" ! Rate limit (attempt {attempt}) - waiting {API_RETRY_WAIT}s...")
        if attempt < API_RETRY_MAX:
            time.sleep(API_RETRY_WAIT * attempt)

    except anthropic.APIError as e:
        print(f" ! API error (attempt {attempt}): {e}")
        if attempt < API_RETRY_MAX:
            time.sleep(API_RETRY_WAIT)

    # Fallback after all attempts
    print("Classification failed - use fallback score=0.5 label=0")
    return {"score": 0.5, "label": 0, "reason": "classification_failed"}


# ================================================================================
# GMM THRESHOLD
# ================================================================================
def _gaussian_intersection(m1, s1, w1, m2, s2, w2) -> float:
    """Intersection between two weighted gaussians (used for GMM threshold)."""
    s1, s2 = max(float(s1), 1e-6), max(float(s2), 1e-6)
    A = (1 / (2 * s2**2)) - (1 / (2 * s1**2))
    B = (m1 / s1**2) - (m2 / s2**2)
    C = (
        (m2**2 / (2 * s2**2))
        - (m1**2 / (2 * s1**2))
        + math.log((w2 / s2) / (w1 / s1))
    )
    if abs(A) < 1e-12:
        return float(-C / B) if abs(B) >= 1e-12 else np.nan
```

```
    disc = B**2 - 4 * A * C
    if disc < 0:
        return np.nan
    r1 = (-B + math.sqrt(disc)) / (2 * A)
    r2 = (-B - math.sqrt(disc)) / (2 * A)
    lo, hi = min(m1, m2), max(m1, m2)
    candidates = [r for r in (r1, r2) if lo <= r <= hi]
    if not candidates:
        return np.nan
    return float(min(candidates, key=lambda r: abs(r - 0.5 * (m1 + m2))))


def infer_gmm_threshold(
    values: np.ndarray,
    fallback: float,
    label: str,
) -> tuple[float, float, float]:
    """
    Infers the threshold using a two-component Gaussian Mixture Model (GMM).

    Returns (threshold, mean_low, mean_high).

    If the GMM cannot be estimated or is not reliable:
    - the 75th percentile of the observed scores is used (if available),
    - otherwise the fallback value (0.50) is used.

    mean_low/mean_high = NaN when a fallback rule is applied.
    """
    values = np.asarray(values, dtype=float)
    values = values[np.isfinite(values) & (values >= 0) & (values <= 1)]
    nan_pair = (np.nan, np.nan)

    # If the sample is too small: use a conservative fallback (75th percentile if
        available)
    if values.size < 40:
        fb = float(np.quantile(values, 0.75)) if values.size > 0 else fallback
        print(f" ! [{label}] Solo {values.size} valori - fallback {fb:.4f}")
```

```
        return fb, *nan_pair

    gmm = GaussianMixture(n_components=2, random_state=123, n_init=10)
    gmm.fit(values.reshape(-1, 1))

    order = np.argsort(gmm.means_.ravel())
    means = gmm.means_.ravel()[order]
    sds = np.sqrt(gmm.covariances_.ravel())[order]
    weights = gmm.weights_.ravel()[order]

    # Minimum bimodality check: if the clusters are too close --> use fallback
    delta = abs(means[1] - means[0])
    if delta < 0.10:
        print(f" ! [{label}] Non bimodale (Delta={delta:.3f}) - fallback {fallback:.4f}"
            )
        return fallback, *nan_pair

    thr = _gaussian_intersection(*means, *sds, *weights)
    if not np.isfinite(thr) or not (0.05 <= thr <= 0.95):
        print(f" ! [{label}] Intersezione fuori range ({thr:.3f}) - fallback {fallback
            :.4f}")
        return fallback, *nan_pair

    print(
        f" [{label}] soglia={thr:.4f} "
        f"cluster: {means[0]:.3f} / {means[1]:.3f} (Delta={delta:.3f})"
    )
    return float(thr), float(means[0]), float(means[1])

# ================================================================================
# TF-IDF COSINE (execution)
# ================================================================================
def tfidf_cosine(a: str, b: str) -> float:
    """TF-IDF Cosine similarity between two texts."""
    a, b = normalize(a), normalize(b)
    if not a or not b:
        return 0.0
```

```
    vect = TfidfVectorizer(lowercase=True, token_pattern=r"(?u)\b\w+\b", min_df=1)
    X = vect.fit_transform([a, b])
    return float(cosine_similarity(X[0], X[1])[0, 0])


# ==============================================================================
# MAIN
# ==============================================================================
def main() -> None:

    # ====== [1/5] Initialize Claude API client ================================
    print(f"\n[1/5] Initialize Claude API client")
    print(f" Model : {API_MODEL}")
    print(f" Temp : {API_TEMPERATURE} (deterministic)")

    # The API key is read from an environment variable to avoid
    # including sensitive credentials in the replication package code.
    api_key = os.getenv("ANTHROPIC_API_KEY")
    if not api_key:
        raise RuntimeError(
            "ANTHROPIC_API_KEY not found. Set it as an environment variable"
            "before running the script."
        )

    client = anthropic.Anthropic(api_key=api_key)
    print(" Client ready")

    # ====== [2/5] Data upload ================================
    print(f"\n[2/5] Caricamento dati: {INPUT_PATH}")
    df = pd.read_stata(str(INPUT_PATH), convert_categoricals=False)
    print(f" {len(df)} righe x {len(df.columns)} colonne")

    # Check that the necessary columns exist
    required = (
        ["writing", "treatment"]
        + [f"query_{k}" for k in range(1, MAX_TURNS + 1)]
        + [f"chat_response_{k}" for k in range(1, MAX_TURNS + 1)]
```

```
)
missing = [c for c in required if c not in df.columns]
if missing:
    raise ValueError(f"Colonne mancanti: {missing}")

# Identify AI subject (API calls for ideation classification are invoked only for
    these subjects)
ai_mask = df["treatment"].astype(str).isin(AI_ACCESS_TREATMENTS)
print(f" Soggetti AI : {ai_mask.sum()}")
print(f" Soggetti NoAI: {(~ai_mask).sum()}")

# ====== [3/5] API classification (AI subjects only) ======================
print(f"\n[3/5] Classificazione ideation via Claude API...")
print(f" (una chiamata per query - ~{ai_mask.sum()} soggetti AI)\n")

# query_results stores, for each AI subject:
# (turn number, query text, API score, API label, explanation)
# It is used to:
# - compute the maximum score per subject
# - determine idea_request
# - generate the audit CSV
query_results: dict[int, list[tuple]] = {}

ai_locs = [i for i, v in enumerate(ai_mask) if v]
n_ai = len(ai_locs)
n_calls = 0

for progress, row_loc in enumerate(ai_locs, start=1):
    if progress % 10 == 0:
        print(f" AI subject {progress}/{n_ai} (API calls: {n_calls})...")

    row = df.iloc[row_loc]
    results = []

    # One API call for each non-empty query (up to MAX_TURNS)
    for k in range(1, MAX_TURNS + 1):
        q = normalize(row.get(f"query_{k}"))
```

```
        if not q:
            continue

        out = classify_query(client, q)
        results.append((k, q, out["score"], out["label"], out["reason"]))
        n_calls += 1

    query_results[row_loc] = results

print(f"\n Classification completed ({n_calls} total API calls)")

# ====== [4/5] GMM threshold --> idea_request + ideation_api_score_max +
    n_query_filled ====
print("\n[4/5] GMM threshold inference and construction of ideation variables...")

# Maximum score per subject (AI subjects only). Subjects with no queries have
    default=0.0,
# but are excluded from the GMM fit via has_query_mask.
score_max_per_subject = np.array(
    [
        max((s for _, _, s, _, _ in query_results[loc]), default=0.0)
        for loc in ai_locs
    ],
    dtype=float,
)

has_query_mask = np.array([len(query_results[loc]) > 0 for loc in ai_locs])
score_max_for_gmm = score_max_per_subject[has_query_mask]
n_no_query = int((~has_query_mask).sum())

print(
    f" AI subjects with query: {has_query_mask.sum()} | "
    f"with no query (excluded by GMM): {n_no_query}"
)

# The ideation threshold is estimated using a two-component GMM on the maximum
    scores.
```

```
# If the GMM is not reliable, fallback=0.50 is used.
api_threshold, _, _ = infer_gmm_threshold(
   score_max_for_gmm,
   fallback=API_THRESHOLD_FALLBACK,
   label="ideation API",
)

# == Construct output columns (only the required variables)
    ===========================
idea_request_col = np.zeros(len(df), dtype=float)
ideation_api_score_max_col = np.full(len(df), np.nan, dtype=float)
n_query_col = np.zeros(len(df), dtype=float)

# Audit: one row per AI subject with a summary of classifications and a preview of
    the final writing
audit_rows = []

for row_loc in ai_locs:
   results = query_results[row_loc]
   n_query_col[row_loc] = float(len(results))

   if not results:
      audit_rows.append({
         "id_code" : normalize(df.iloc[row_loc].get("id_code")),
         "treatment" : normalize(df.iloc[row_loc].get("treatment")),
         "n_queries" : 0,
         "idea_request" : 0,
         "score_max" : np.nan,
         "scores_per_query" : "",
         "reasons_per_query" : "",
         "writing_preview" : normalize(df.iloc[row_loc].get("writing"))[:200],
      })
      continue

   score_max = max(s for _, _, s, _, _ in results)
   ideation_api_score_max_col[row_loc] = score_max
```

```
    # idea_request = 1 if at least one of the subject's queries exceeds the
        estimated threshold.
    # The first occurrence in chronological order is considered.
    first_rank = None
    for rank, (_, _, score, _, _) in enumerate(results, start=1):
        if score >= api_threshold:
            first_rank = rank
            break

    idea_request_col[row_loc] = 1.0 if first_rank is not None else 0.0

    audit_rows.append({
        "id_code" : normalize(df.iloc[row_loc].get("id_code")),
        "treatment" : normalize(df.iloc[row_loc].get("treatment")),
        "n_queries" : len(results),
        "idea_request" : int(first_rank is not None),
        "score_max" : round(score_max, 4),
        "scores_per_query" : "; ".join(
            f"Q{k}={s:.3f}[{l}]" for k, _, s, l, _ in results
        ),
        "reasons_per_query" : " | ".join(
            f"Q{k}: {r[:60]}" for k, _, _, _, r in results
        ),
        "writing_preview" : normalize(df.iloc[row_loc].get("writing"))[:200],
    })

print(f" idea_request assigned (threshold = {api_threshold:.4f})")

# == [5/5] Execution measure (TF-IDF) - ONLY sim_writing_ai_concat + n_query_filled
    ==
# sim_writing_ai_concat measures how similar the participant's final writing is
# (TF-IDF cosine similarity) to the concatenation of all AI responses received.
print("\n[5/5] Calcolo sim_writing_ai_concat (TF-IDF)...")

sim_concat_col = np.zeros(len(df), dtype=float)

for row_loc in range(len(df)):
```

```
    row = df.iloc[row_loc]
    writing = normalize(row.get("writing"))

    responses = []
    q_count = 0
    for k in range(1, MAX_TURNS + 1):
        if normalize(row.get(f"query_{k}")):
            q_count += 1
        r = normalize(row.get(f"chat_response_{k}"))
        if r:
            responses.append(r)

    # For NoAI subjects, n_query_filled is counted as the number of non-empty
        queries.
    if not ai_mask.iloc[row_loc]:
        n_query_col[row_loc] = float(q_count)

    # If writing is missing or there are no AI responses, similarity remains 0 (
        default).
    if not writing or not responses:
        continue

    sim_concat_col[row_loc] = tfidf_cosine(writing, " ".join(responses))

print(" sim_writing_ai_concat calcolata")

# === Attach ONLY the 4 required variables ====================================
# The four variables are added to the original dataset.
# No other variables are created or saved.
df["idea_request"] = idea_request_col
df["ideation_api_score_max"] = ideation_api_score_max_col
df["n_query_filled"] = n_query_col
df["sim_writing_ai_concat"] = sim_concat_col

# ====== Save output ===========================================================
print(f"\nSaving output...")
df.to_stata(str(OUTPUT_PATH), write_index=False, version=118)
```

```
    print(f" Dataset --> {OUTPUT_PATH}")

    pd.DataFrame(audit_rows).to_csv(str(AUDIT_CSV), index=False, encoding="utf-8-sig")
    print(f" Audit CSV --> {AUDIT_CSV}")

    # ====== Final statistics ====================================================
    # Summary statistics for replication checks. They do not affect the output.
    n_idea = int((df.loc[ai_mask, "idea_request"] == 1).sum())
    denom = max(int(ai_mask.sum()), 1)

    print("\n" + "=" * 66)
    print("Final Statistics (CODE + AUDIT)")
    print("=" * 66)
    print(f"Total subjects : {len(df)}")
    print(f"AI subjects (T_AI, PR_AI, LOT_AI) : {int(ai_mask.sum())}")
    print(f"NoAI subjects : {int((~ai_mask).sum())}")
    print(f"Total API calls : {n_calls}")
    print()
    print(f"idea_request = 1 : {n_idea:3d} ({100*n_idea/denom:5.1f}% AI)")
    print("=" * 66)
    print("\nCompletato.\n")


# ==============================================================================
# ENTRY POINT
# ==============================================================================
if __name__ == "__main__":
    main()
```

APPENDIX C. PARAMETRIC ANALYSIS BASED ON *AI-first keywords* AND *User-first keywords*

TABLE 1. SEMs: ideation and execution (restricted to participants in the AI treatments) – AI-first and User-first keywords

	Parsimonious			Extended		
	Direct (1)	Indirect (2)	Total (3)	Direct (4)	Indirect (5)	Total (6)
Panel A: Mediator equation (Similarity)						
AI-first keywords	0.133***			0.134***		
	(0.005)			(0.005)		
User-first keywords	0.116***			0.116***		
	(0.008)			(0.008)		
Residualized number of queries	0.012			0.012		
	(0.007)			(0.007)		
Residualized number of queries^2	-0.001			-0.001		
	(0.001)			(0.001)		
Number of characters				1.84×10^{-5}		
				(3.31×10^{-5})		
Male				-0.003		
				(0.019)		
DAT score				0.020		
				(0.039)		
DAT score^2				-1.18×10^{-4}		
				(2.40×10^{-4})		
Constant	0.105***			-0.755		
	(0.014)			(1.603)		
Panel B: Outcome equation (Average score)						
Similarity	1.513*		1.513*	1.468*		1.468*
	(0.605)		(0.605)	(0.606)		(0.606)
AI-first keywords	0.065	0.202*	0.267***	0.074	0.196*	0.270***
	(0.098)	(0.081)	(0.067)	(0.097)	(0.082)	(0.066)
User-first keywords	-0.118	0.176*	0.058	-0.111	0.170*	0.059
	(0.095)	(0.073)	(0.067)	(0.094)	(0.073)	(0.066)
Residualized number of queries	0.028	0.018	0.046	0.028	0.018	0.046
	(0.051)	(0.011)	(0.056)	(0.053)	(0.011)	(0.056)
Residualized number of queries^2	-0.017*	-0.001	-0.018*	-0.017*	-0.001	-0.019*
	(0.008)	(0.001)	(0.009)	(0.008)	(0.001)	(0.009)
Number of characters				3.14×10^{-4}	2.70×10^{-5}	3.41×10^{-4}
				(3.65×10^{-4})	(4.85×10^{-5})	(3.77×10^{-4})
Male				0.058	-0.004	0.055
				(0.193)	(0.029)	(0.195)
DAT score				-0.051	0.029	-0.022
				(0.557)	(0.057)	(0.570)
DAT score^2				4.34×10^{-4}	-1.73×10^{-4}	2.61×10^{-4}
				(0.003)	(3.50×10^{-4})	(0.003)
Constant	5.578***		5.578***	6.476		6.476
	(0.166)		(0.166)	(23.184)		(23.184)
Obs.		212			212	

Notes. This table replicates the econometric approach of Table 3 in the main paper using the behavioral indicators of AI reliance *AI-first keywords* and *User-first keywords* introduced by Zhu and Zou (2026). SEMs are estimated via maximum likelihood with bootstrap standard errors based on 2,000 replications. The mediator is Similarity. Columns report direct, indirect (through the mediator), and total effects. For mediator equations, indirect effects are not defined; total effects therefore coincide with direct effects. The number of queries is residualized with respect to AI-first and User-first keywords and included in the model together with its squared term in order to capture interaction intensity orthogonal to the keyword-based measures of AI reliance. The other remarks of Table 3 apply. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

REFERENCES

Zhu, F. and W. Zou (2026). Generative AI adoption in human creative tasks: Experimental evidence. *Journal of Economic Behavior & Organization 242*, 107414.