

The Impact of Generative AI on Creative Performance: Experimental Evidence from AI- Assisted Ideation

Master of Engineering Management

2024 – 2026

Submitted in partial fulfillment of the requirements for the degree of
Master of Engineering Management



School of Engineering and Digital Sciences

In partnership with

Graduate School of Business

Ayaulym Ayazhan

Dias Nurmukhambetov

Marzhan Kabzhassarova

Supervisor: Professor Maria Tsakalerou

May, 2026

The Impact of Generative AI on Creative Performance:
Experimental Evidence from AI-Assisted Ideation

by

Ayaulym Ayazhan

Dias Nurmukhambetov

Marzhan Kabzhassarova

Submitted in partial fulfillment of the requirements for the degree of
Master of Engineering Management

School of Engineering and Digital Sciences

Graduate School of Business

Nazarbayev University

Supervised by

Professor Maria Tsakalerou

May, 2026

Declaration

We, Ayaulym Ayaulym Ayazhan, Dias Nurmukhambetov and Marzhan Kabzhassarova, declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the author's original work. The thesis has not been previously submitted to this or any other university for a degree and does not incorporate any material already submitted for a degree.

Signature (s):



Date: 02.05.2026

BLANK

Abstract

In this research paper, the effects of Generative Artificial Intelligence (AI) on human creativity will be studied based on the following four factors: fluency, flexibility, originality, and elaboration. By using the Alternative Uses Task (AUT), it is determined how AI affects creativity when compared to those who work without any assistance from it. According to the results, AI helps increase creativity through originality and elaboration but lowers semantic distance scores (measured by GloVe). It indicates that despite increasing the quality of creative responses (OCSAI), AI limits the process of idea generation and tends to encourage convergent thinking and fixation bias.

From these results, there is an evident mixed effect of AI in creativity. AI can serve as a cognitive facilitator by allowing more cognitive effort to be applied to higher-order thinking processes; however, it may also negatively affect critical thinking skills due to cognitive offloading.

In spite of the many useful information that were gained from the experiment, there are still some issues that were raised regarding the limitations of the study such as the use of non-blinded preprocessing, unequal gender distribution, and the lack of using the pre-test/post-test approach in measuring the participants' creativity before and after being exposed to the AI treatment.

BLANK

Acknowledgments

First and foremost, we would like to thank ourselves for the courage and determination we showed in committing to this thesis. It has been a journey full of learning, growth, and challenges, and we did our best to give it our all.

We are deeply grateful to Professor Mariza Tsakalerou, our supervisor and Director of the Engineering Management Program at Nazarbayev University, for her invaluable knowledge, constant support, and patience throughout this process. Your guidance and attention have been essential to the completion of this project.

Our heartfelt thanks go to our families for their unwavering love and encouragement. Their support gave us the strength to keep moving forward, even in the toughest times.

We would also like to thank our fellow colleagues in the Master of Engineering Management class of 2026 at Nazarbayev University. The collaboration and shared experiences have made this journey all the more enriching.

Finally, we are sincerely thankful to Nazarbayev University, our alma mater, for providing us with the opportunity to pursue this path of higher learning and growth. The knowledge and experiences gained here will shape our futures in ways we cannot yet fully comprehend.

BLANK

Contents

Abstract	4
Acknowledgments	6
Contents	8
List of Tables	10
List of Figures	11
List of Abbreviations	12
1. Introduction	1
2. Literature review	2
2.1. Human creativity	2
2.2. Dimensions of Creative Performance and Their Methodologies	4
2.2.1. Fluency	4
2.2.2. Flexibility	5
2.2.3. Originality	5
2.2.4. Elaboration	6
2.3. AI and Human Creativity	8
2.3.1. AI - creativity enabler	8
2.3.2. Blind spots of AI	10
2.4. Mechanisms of AI's Impact on Creativity	11
2.4.1. Cognitive Offloading	11
2.4.2. Fixation Bias	12
3. Research Gap, Aims, and Scope	13
4. Methodology	15
4.1. Design of the Experiment	15
4.1.1. No-AI Group Procedure	16
4.1.2. AI Group Procedure	16
4.2. Creativity Measures Assessed	17
4.2.1. Alternative Uses Task (AUT)	17
4.2.2. Creativity scoring	17
5. Data Analysis	18
5.1. Data Preparation	18
5.2. Analysis	20
6. Results	21
6.1. Fluency	21
6.2. Elaboration	23
6.3. Originality (GloVe and OSCAI)	26
7. Discussion	30
8. Conclusion	33
References	34
Appendix A: CITI Training Certificates	37

A.1 Ayaulym Ayazhan.....	37
A.2 Dias Nurmukhambetov.....	37
A.3 Marzhan Kabzhassarova.....	38
Appendix B: Experimental Protocol	39
Setting Up	39
Greeting and Consent	39
Instructions and Task - Non-AI Group.....	40
Oral Instructions (read aloud to participants)	40
Procedure.....	40
Closing the Session	40
Instructions and Task - AI Group	41
Pre-Session Calculation.....	41
Participant and Room Setup	41
Procedure.....	41
Appendix C: Consent Form.....	42
Appendix D: Code for GloVe	44
D.1 Code parts	44
D.2 Code for t-test	46
D.3 Code for visuals	46
Appendix E: Code Output for Fluency	48
Notes: 49	

List of Tables

Table 6.1.1. Correlation Matrix	22
Table 6.3.1. The mean value of GloVe and OSCAI scores for groups	26
Table 6.3.2. The mean value of GloVe and OSCAI scores based on genders	29
Table 6.3.3. The t-test results.....	29
Table 6.3.4 The modification level results for AI assisted group	30
Table 6.3.5. The t-test results for modification.....	29

List of Figures

Figure 1.1. Creative Performance Measurement	3
Figure 2.1. 4 Pillars of Divergent Creative Thinking.....	8
Figure 3.1. The conceptual framework	14
Figure 6.1.1. The distribution of responses and spread for “pencil”	22
Figure 6.1.2. The distribution of responses and spread for “fork”	22
Table 6.2.1. Participant Elaboration Scores.....	24
Figure 6.2.1. Elaboration Score Means by AI and no AI groups.....	24
Figure 6.2.2. Mean Elaboration Scores for AI groups based on the degree of transformation....	25
Figure 6.3.1. GloVe semantic distance by Group and Prompt	27
Figure 6.3.2. OSCAI by Group and Prompt	28

List of Abbreviations

MEM		Master of Engineering Management
SEDS		School of Engineering and Digital Sciences
NU		Nazarbayev University
		

BLANK

1. Introduction

The concept of quality, historically attributed exclusively to humans, creativity, is an integral part of human nature. It is present in humans as both a capacity (the ability to create) and a process (the act of creation itself) (Habib et al., 2023). Creativity is applied both in everyday life and in situations where solutions to specific problems are centralized in creative activities and artistic endeavors. (Kaufman & Beghetto, 2009; Richards, 2007). However, the rapid development of generative artificial intelligence (GenAI) technologies is expanding the capabilities in creating original content, at the same time challenging the creativity that was once considered an ability exclusive to human beings. (Arriagada, 2020). GenAI reported to excel in many areas that require reasoning and decision making involving what humans typically perceive as creativity: chess, art (Jones, 2018), and poetry (Köbis & Mossink, 2021). Additionally, Doshi & Hauser (2024) reported an increase in the quality and efficiency of production of storytelling, white-collar work, and programming code, enabling better customer relationships.

But to dive deeper and see how Generative AI influences creativity and if GenAI, indeed, has a detrimental impact on the creative performance of humans, we first need to define it.

2. Literature review

2.1. Human creativity

A widely accepted definition of creativity is the ability to create ideas that are unique and useful (Koivisto & Grassini, 2023). Uniqueness is the extent of departure of the idea from the status quo. Usefulness, whereas, addresses the practicality of the idea, its relevance (Doshi & Hauser, 2024). Based on the associative theory of creativity, ideas are created by establishing a connection between unrelated or weakly related concepts. Notion, which is often met in psychological research as a substitute for creativity, is creative divergent thinking.

Guilford (1967) claims that divergent thinking is about creating many different ideas and solutions. Additionally, it includes exploring multiple possibilities and embracing unconventional approaches (Silvia, 2008). In contrast, convergent thinking is about narrowing down generated ideas, finding and sticking with the best, most appropriate one, and refining it into a coherent solution (Kumar et al., 2025).

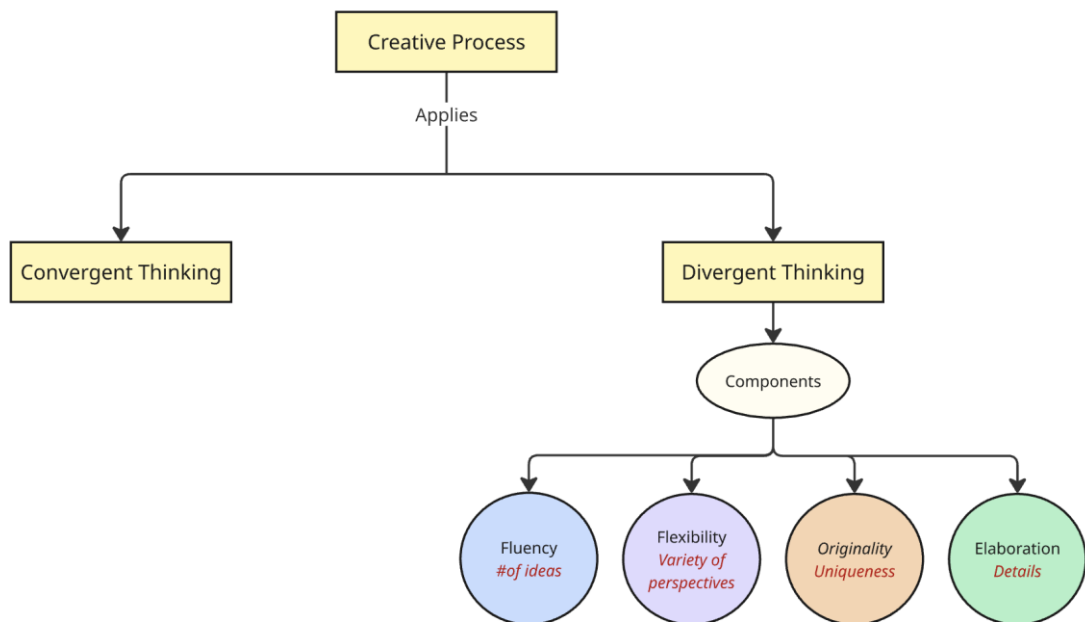


Figure 1.1. Creative Performance Measurement

Divergent creative thinking (DT) typically involves 4 pillars: fluency, flexibility, originality, and elaboration (Koivisto & Grassini, 2023; Hubert et al., 2024).

To assess Divergent Thinking, the common framework called the Alternative Uses Task (AUT) has been implemented. This method instructs participants to generate applications for common day-to-day objects like a pencil or a brick. It is widely accepted, according to Hubert et al., (2024). Participants generate uncommon, creative uses for everyday objects, such as a brick or a paperclip.

Potential approaches investigated but found to be of less significance: Divergent Associations Task and Consequences Task (Koivisto & Grassini, 2023). The Divergent Associations Task offers a relatively new, standardized approach to measuring creative thinking. In this task, individuals are asked to generate 10 nouns that share as little semantic overlap as possible with one another. A creativity score is then derived by calculating the mean semantic distance across all submitted words, as outlined by Doshi & Hauser (2024).

The Consequences Task takes a different approach by prompting participants to think through the potential outcomes of imagined, real-world-defying scenarios. For example, consider what life might look like if sleep were no longer a biological necessity for humans. Respondents are encouraged to produce as many plausible consequences as they can, making it a measure of both ideational fluency and imaginative reach (Doshi & Hauser, 2024).

2.2. Dimensions of Creative Performance and Their Methodologies

As the creative potential is accessed through four dimensions, there are methods used by researchers to assess those dimensions in AUT.

2.2.1. Fluency

Fluency, defined as the total number of interpretable, meaningful, and relevant ideas generated in response to a prompt (Koivisto & Grassini, 2023), is the first dimension assessed in the AUT. Across the literature, instructions for both AI and non-AI groups were broadly comparable. For instance, Habib et al. (2023) simply directed participants to "generate as many uses for [an object] as you can." Haase and Hanel (2023) similarly used identical instructions for both groups, though their prompts were more structured, incorporating the SCAMPER framework.

However, a number of studies deliberately differentiated the prompts between groups, capping the AI's output at a fixed number of responses calculated to match the average

produced by human participants. For example: *"Generate [X] original and creative uses for [item]. Clever, unusual, humorous, or innovative ideas. Practicality not required."*

This approach was designed to neutralize two key confounds. The first is raw computational advantage: unconstrained, AI can generate far more ideas than a human in the same timeframe. The second is the fluency-originality link: the statistical tendency for higher idea volume to yield more original responses by chance alone. Without controlling for fluency, any observed advantage in AI originality could simply reflect quantity rather than quality.

Supporting this, Koivisto and Grassini (2023) noted that without explicit quantity instructions, ChatGPT-3 defaults to exactly 10 ideas, ChatGPT-4 to 7 or 8, and copy.ai to a variable number. These differences reflect technological variation rather than creative capacity. To address this, many studies first administered the task to the human group, established a baseline fluency count, and then embedded that figure into the AI's prompt. Hubert et al. (2024) took a stricter approach, excluding any prompts that contained quantity parameters and instead matching fluency 1:1 across groups. Both methods converge on the same goal: ensuring that differences in originality, elaboration, and flexibility reflect genuine creative quality rather than an artifact of output volume.

2.2.2. Flexibility

Flexibility - the number of different categories of relevant responses. In terms of the creative performance and in the concept of AUT, this metric captures the ability to think about a topic from different perspectives (Koivisto & Grassini, 2023). The experiments in the literature differ; however, the scoring method remains the same by exploring the different categories of relevant responses from AUT. For example, the given ideas can be categorized by use directions, and if an item "brick" is used as a "paperweight" or a "doorstop" would be categorized in one as the "holding things in place or down. However, if the list of uses includes "weapon" would represent a new shift and be different from the initial category (Desdevises, 2025). This is often done by assigned judges. They are instructed to gather and agree on relevant categories so they can later see how different ideas are and code them by category. In the case of Habib et al. (2023), this method was supported by counting the number of categories in which the answer appeared. The more flexible the respondent's answer, the higher the number of categories. Higher flexibility indicated greater creativity, as the answer encompassed various categories of uses for a given object.

2.2.3. Originality

Originality - the statistical rarity of the responses, in other words, the uniqueness of the ideas. According to the literature, it is measured through several methodologies. First one is:

Subjective Human Rating, which is considered the traditional method, involves the judges simply evaluating the quality of ideas. It happened only based on their perceived cleverness or innovation. Judges typically use Likert scales, such as a 1-5 point scale (1 = not at all creative, 5 = very creative) or a 1-7 point scale. In Koivisto & Grassini (2023), the judges were not told which responses were from the AI or the non-AI group and rated each response. The instructions emphasized that participants should prioritize novelty over usefulness and use the instructions given to them as a starting point for rating responses. They also were clearly given instructions that common usage, such as "cut" in the response to the item "scissors," should be rated low= 0, while unclear or illogical responses, as well as incomplete responses, should be rated 1. Each rater chose the order in which to rate the four objects.

The order of presentation of responses within object categories was presented in a random way and separately for each rater. Each rater's ratings were averaged across all responses the participant (or the chatbot in the session) gave to an object, and the final subjective ratings for each object were formed by averaging the ratings of the six raters. Inter-rater reliability was assessed by calculating intraclass correlation coefficients. Another study included rating subjectively, not by pre-trained judges, but by the participants themselves. At the end of the creative task, an additional task was offered to participants to assess ChatGPT's ability to evaluate the creativity of its own results. Participants were asked to rate each idea on a scale of 1 to 7 (where 1 represents "not at all creative" and 7 represents "very creative").

Another available approach is more quantitative: through semantic distance between words. One of the available models is GloVe (Global Vectors for Word Representation), a word-embedding model that represents each word as a numerical vector in a high-dimensional space (Carlson, Bauer, & Manning, 2025). There, the semantic distance is calculated using (Pennington, Socher, Manning, 2014).

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

$$\text{Distance} = 1 - \cos(v_1, v_2)$$

Where v_1 - the baseline word ("fork")

v_2 - the idea of respondents

This model has been developed by researchers from Stanford. Further details:

<https://nlp.stanford.edu/projects/glove/>

2.2.4. Elaboration

Elaboration refers to the amount of detail embedded within each response. In other words, how richly and thoroughly an idea is expressed rather than merely stated. In Dumas et al. (2021), elaboration is typically operationalized through word count, under the assumption that more descriptive responses reflect deeper cognitive engagement with the idea. Several variants of this approach have been developed, each with different levels of linguistic sophistication. The most straightforward measure is unweighted word count, which simply counts the number of words in a response using whitespace tokenization or a natural language processing tool such as SpaCy. While easy to compute and replicate, this measure treats all words as equally meaningful, which is a significant limitation. Filler phrases and content-rich descriptions contribute equally to the score. To address this, part-of-speech (POS) inclusion refines the count by retaining only words that carry conceptual weight: nouns, verbs, adjectives, and adverbs. Grammatical function words, articles, prepositions, conjunctions are excluded on the grounds that they contribute little to the semantic richness of a response. This approach is theoretically motivated by the idea that elaboration should capture what is being described, not how the sentence is structured. A related but distinct refinement is stoplist inclusion, which filters out a predefined list of high-frequency, low-information words, commonly referred to as stopwords. Those are transition or filler terms (e.g., *also*, *however*, *basically*). Unlike POS filtering, which is grammar-driven, stoplisting is corpus-driven, removing words identified empirically as uninformative regardless of their grammatical category. The most nuanced approach is inverse frequency weighting, which weights each word by how rarely it appears in a reference corpus. The underlying logic is that uncommon words signal more specific, elaborated thinking. A response describing a brick as useful for "thermal mass retention in passive solar construction" conveys far greater elaboration than one saying it could "build a wall," even if both contain a similar word count. By down-weighting frequent, generic vocabulary and

amplifying rare, precise terminology, this method captures qualitative richness in a way raw counting cannot.

Taken together, these four operationalizations form a progression from surface-level measurement to semantically sensitive scoring, reflecting ongoing methodological refinement summarized by Dumas et al. (2021). Due to study time constraints, whitespace tokenization or the use of natural language processing tools such as SpaCy could not be applied and tested. The second method of creating part-of-speech (POS) inclusion, which requires a comprehensive knowledge of vocabulary and terminology, will also be excluded from the measurement methods. In this study, information processing will be coded by a corpus-driven stop-listing method, by removing the common words (such as "the," "and," "in," etc.) that don't contribute significantly to the meaning of the response.

The dimensions of DT and methods of measuring those dimensions in AUT are summarized in Figure 2.1.

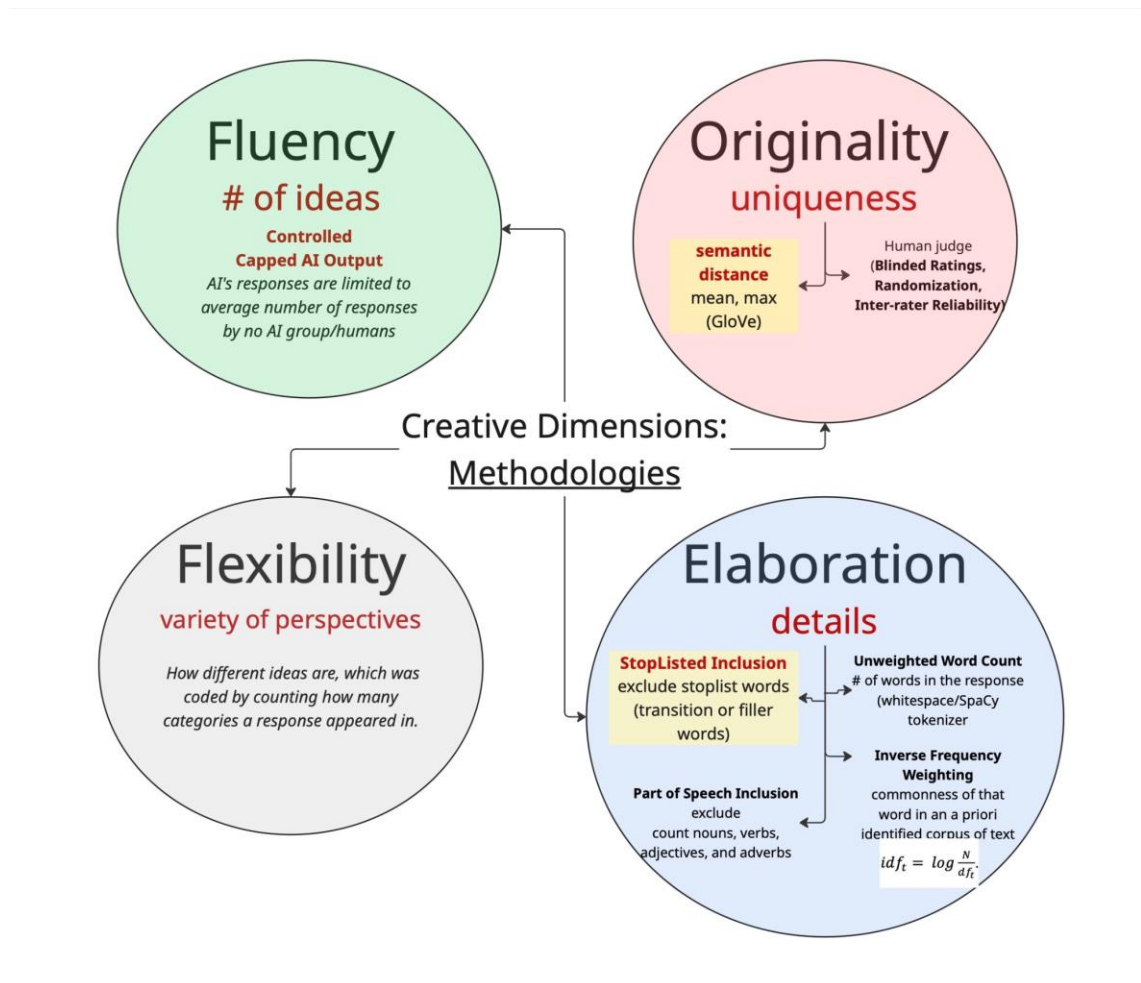


Figure 2.1. 4 Pillars of Divergent Creative Thinking

2.3. AI and Human Creativity

2.3.1. AI - creativity enabler

A growing body of research converges on the finding that AI meaningfully enhances creative performance, though the nature and extent of this benefit vary across contexts and individuals. In the short term, LLM assistance has been shown to boost creative output during assigned tasks, with participants demonstrating stronger performance in the early stages of AI-assisted work (Kumar et al., 2025). This early-stage advantage is particularly important, as it suggests that AI may be most effective during the ideation phase of creative processes, where breadth and exploration are prioritized over refinement. This improvement is not limited to fluency, meaning quantity alone. Dimensions such as flexibility and elaboration also appear to benefit. Users have reported that tools like ChatGPT helped them expand the range and variety of their ideas, pushing them to consider alternative uses and interpretations they might not have arrived at independently (Habib et al., 2023). In this sense, AI appears to function as a catalyst for divergent thinking, enabling users to move beyond habitual cognitive pathways and access a wider semantic search space. This pattern is further supported by Urban et al. (2024), who found that students using ChatGPT produced more original, more elaborated, and higher-quality solutions to complex creative problems, while also reporting greater confidence in their ability to complete the task. Notably, this increase in perceived self-efficacy may itself reinforce creative performance, as *confidence has been shown to positively influence willingness to explore unconventional ideas*.

These gains are further supported by evidence that GPT-4 outperforms human participants on measures of originality and elaboration in different environments: structured creative tasks (Hubert et al., 2024), and educational settings (Marrone, 2026). Across studies, several points of consensus emerge. Scholars broadly agree that generative AI is capable of producing large volumes of ideas at a pace unconstrained by human cognitive limitations, such as fatigue or working memory capacity. This scalability allows users to rapidly iterate across multiple idea trajectories, a process that would otherwise be cognitively and temporally demanding. Beyond this, AI also avoids the fixation bias that commonly narrows human thinking, the tendency to become anchored to initial ideas at the expense of more divergent ones (Desdevises, 2025).

Current large language models, particularly GPT-4, have been found to meet or surpass average human performance across standardized creativity assessments (Koivisto & Grassini, 2023). However, this advantage appears to operate primarily at the level of average performance, as the most creative human individuals can still match or exceed AI in originality, suggesting that AI raises the floor of creativity more reliably than it raises its ceiling (Koivisto & Grassini, 2023). This distinction is critical, as it implies that AI's primary contribution may lie in *democratizing access to creative performance rather than redefining its upper limits*.

Relatedly, Marrone et al. (2026) argue that AI's creative value lies less in independent originality and more in its role as a collaborator that augments human creativity by offloading repetitive or convergent cognitive tasks, thereby enabling greater focus on higher-level idea generation. Perhaps most notably, the creative benefit of AI assistance appears greatest for individuals with lower baseline creative ability, suggesting an equalizing effect that narrows the performance gap between less and more creative individuals (Doshi & Hauser, 2024). Taken together, these findings position AI not merely as a tool for efficiency, but as a socio-cognitive amplifier that reshapes how creative work is distributed, performed, and accessed across different user groups.

2.3.2. Blind spots of AI

Not all evidence, however, supports an optimistic view of AI's creative influence. Some researchers have found that while AI exposure increases the diversity of ideas at a collective level, it does not meaningfully improve the creativity of individual ideas themselves, suggesting that AI may broaden the surface of creative output without deepening it (Ashkinaze et al., 2025). This distinction between collective diversity and individual originality highlights a key tension in the literature: AI may expand the pool of ideas without necessarily enhancing their intrinsic novelty or usefulness. Unlike humans, AI tools such as ChatGPT operate without the cognitive constraints that typically limit creative performance, including working memory load, mental fatigue, and attentional fluctuations, enabling them to generate large volumes of ideas rapidly within a single session (Desdevises, 2025). This disparity is evident in preliminary piloting, where ChatGPT-3 consistently produced ten ideas compared to a human median of just three (Koivisto & Grassini, 2023), raising important questions about what AI assistance means for the quality, not just the quantity, of human creative thought.

More critically, over-reliance on AI has been linked to a narrowing of creative thought, particularly in terms of flexibility. When participants based their thinking on AI-generated suggestions, their ability to independently explore alternative ideas became constrained, effectively allowing the tool to define the boundaries of their imagination (Habib et al., 2023). In such cases, AI shifts from being a generative partner to a cognitive anchor, subtly guiding users toward pre-structured solution spaces. Empirical evidence from design research reinforces this concern: Wadinambiarachchi et al. (2024) show that exposure to AI-generated examples increases design fixation, leading participants to produce fewer ideas with lower originality and reduced variety. Hubert et al. (2024) observed parallel outcomes, finding that AI-assisted ideation led participants to fixate on early ideas, ultimately resulting in lower originality, reduced variety, and weaker overall idea production compared to those working without AI support. These findings suggest that while AI can accelerate ideation, it may simultaneously compress the exploratory phase of creativity by prematurely converging on salient or highly probable ideas. These findings open into several broader points of disagreement within the literature. While many studies position AI as a consistently strong creative performer, others argue that the most creative human individuals still match or surpass AI in originality, complicating any straightforward claim of AI superiority (Hubert et al., 2024; Koivisto & Grassini, 2023). The question of fixation itself remains controversial. Some evidence suggests that AI replicates a human-like fixation effect, gravitating toward conventional, predictable patterns embedded in its training data, while others maintain that AI overcomes fixation by drawing on remote semantic associations more efficiently than humans can (Desdevises, 2025). This inconsistency points to the possibility that AI's impact on creativity is highly context-dependent, varying according to task structure, prompt design, and the degree of user reliance. A further limitation concerns evaluative capacity: whereas humans tend to distinguish between their conventional and genuinely original ideas, AI models have been shown to struggle with this kind of differential judgment. They rate predictable outputs as highly as truly novel ones, a blind spot with significant implications for how AI-generated creativity is assessed (Camarda et al., 2024; Desdevises, 2025). Consequently, without critical human oversight, AI-assisted creativity risks amplifying not only idea generation but also the acceptance of mediocre or derivative outputs, thereby diluting overall creative quality.

2.4. Mechanisms of AI's Impact on Creativity

2.4.1. Cognitive Offloading

While AI can influence people's creativity in various ways, as described in existing groundworks, it is very important to understand the fundamental processes and mechanisms behind these influences. One key factor is cognitive offloading, in the context of creativity, this refers to the practice of using external tools, particularly artificial intelligence, to perform mental tasks that would otherwise require significant cognitive effort from a human (Risko & Gilbert, 2016; Marrone, Cropley, & Medeiros, 2026) . Because AI allows people to delegate routine cognitive tasks, it is considered a "cognitive enhancer" or a tool that reduces mental workload and frees up cognitive resources to redirect attention to complex creative processes . The supporting findings were highlighted by (Wang, Tao, Ma, Li, & Wu, 2025), who found that AI assistance allowed cognitive resource reallocation from technical execution toward higher-order creative evaluation. Participants in the study reported that they “could allocate more time to evaluating and refining concepts rather than creating initial drafts”, demonstrating how AI tools helped to reveal more creative potential. The study concluded that the use of AI tools improves creative performance in design students due to cognitive resource optimization (Wang, Tao, Ma, Li, & Wu, 2025). On the other hand, some studies argue that while AI helps humans use cognitive resources more optimally and frees up, at the same time, it may also lead to a decline in cognitive engagement and the development of skills. The widespread availability and use of such AI tools, which provide quick solutions and prepare everything, discourages people from using their critical and creative thinking even in basic cognitive tasks. Gerlich (2025) draws a parallel with the situation where search engines provide available information, which can negatively impact memory retention and the ability to deeply process information. His findings also noted that because people become accustomed to the simplicity and convenience of AI-provided solutions, even in professional and everyday scenarios, cognitive processes are affected, reducing the need for independent critical analysis. This could ultimately lead to a workforce that is highly effective but potentially less capable of critical and creative thinking. Apart from Gerich (2025), other research underscores the potential drawbacks of overreliance on AI. Shukla et al. (2025) claim that AI-driven auto-resolution of problems results in ‘deskilling,’ as workers transition from active participation to passive monitoring roles. The potential shift of cognitive tasks like brainstorming or problem framing towards AI reliance rises questions in the aspect of unintended skill erosion. In conclusion, even if AI can

enhance creativity by optimizing cognitive resources through delegating certain tasks, overreliance on it can reduce cognitive activity and hinder skill development. In other words, overuse of AI can lead to reduced critical and creative thinking.

2.4.2. Fixation Bias

Following the discussion on cognitive offloading, it's also important to consider AI in terms of fixation bias and the creative process in general. Fixation bias, also known as fixation effect or idea fixedness, is a cognitive phenomenon in which people become fixated on initial, generally accepted, or easily accessible ideas (Wadinambiarachchi et al., 2024). This cognitive tendency consequently narrows down the creative thinking scope as it limits the ability of people to explore novel or alternative solutions. For example, in one study of student creativity, he argued that overreliance on AI demonstrated how it hindered creativity in participants. Students who participated in the experiment acknowledged that when working with ChatGPT, it prevented them from coming up with additional ways to solve the problem, ultimately limiting their creativity. According to the authors, this occurred because focusing on the AI's ideas effectively prevented them from thinking independently and limited rather than expanded their capabilities (Habib et al., 2023). Similar results were demonstrated by Hubert (2024) that when a person uses support from AI during the ideation process, it can lead to fixation of ideas in the initial ones. As a result, participants who used showed lower results in idea production compared to the baseline group.

There's also a counter-argument in the literature that AI, precisely because of the inherent human cognitive phenomenon of idea fixedness, can solve this problem and help people become more creative. This is because AI isn't limited by the cognitive factors that typically influence idea generation in humans. While mental fatigue, attentional lapses, or working memory overload can hinder the development of more creative solutions, AI is capable of rapidly generating large volumes of ideas (Desdevises, 2025). To sum up, while AI can overcome human cognitive limitations and help people expand their creative capabilities, in some cases, it can also hinder creativity. It can lead to decreased creative productivity due to fixation bias, causing people to become stuck in easily generated ideas by AI and lose the ability to come up with new ideas.

3. Research Gap, Aims, and Scope

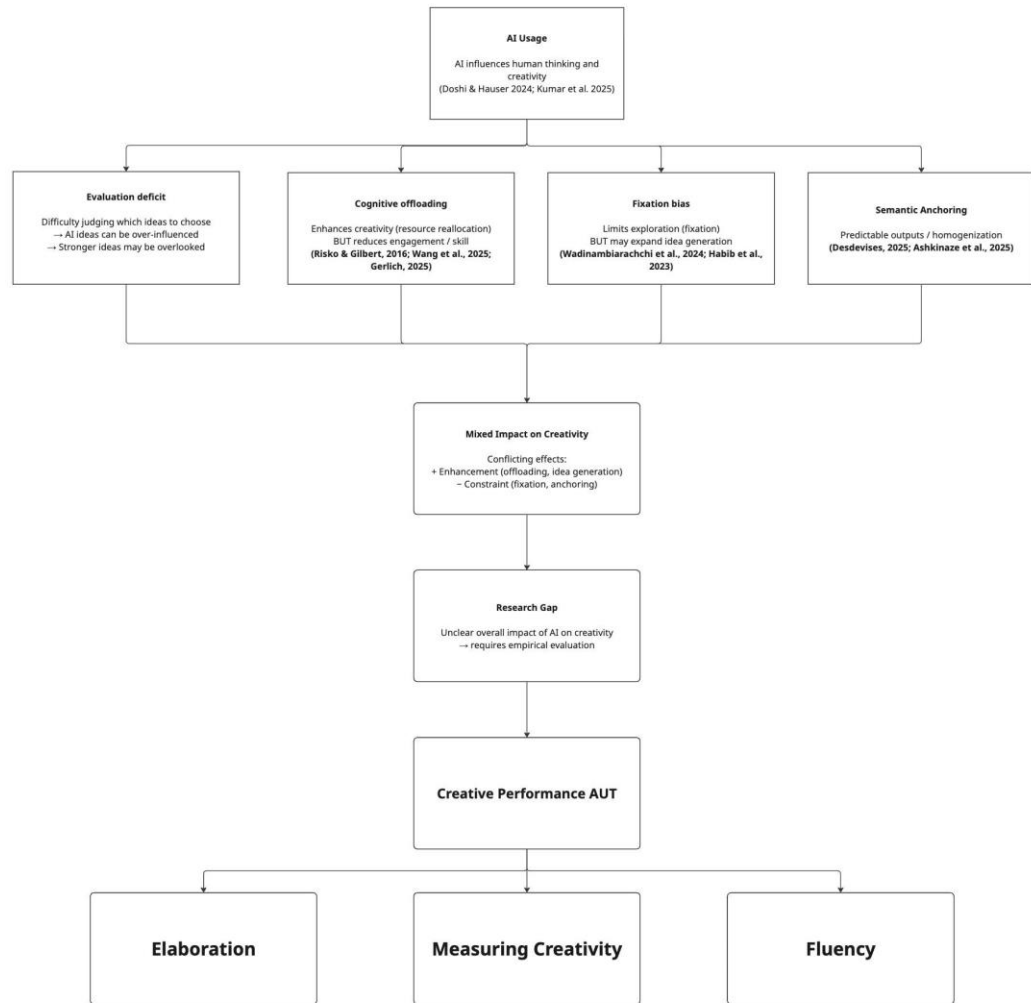


Figure 3.1. The conceptual framework

Despite growing scholarly interest, the relationship between AI and human creativity remains questionable, warranting further empirical investigation. This project addresses that need by examining how AI use affects creative performance across four key dimensions: fluency, flexibility, originality, and elaboration, using the Alternative Uses Task (AUT) as the primary measure of creative output. The study's contributions are operationalized through five key deliverables: a Conceptual Framework Diagram, an Experimental Protocol Document, a Synthesis from the Literature, and a broader Knowledge Contribution to the Thesis. Together, these aim to advance the ongoing debate and offer practical recommendations for the effective integration of AI in creative contexts. A key motivation for this study and the methodology adopted lies in the inherent asymmetry between human and AI ideation capacities. Several gaps in the existing literature further justify this study's design. Most prior research has relied on

earlier model versions such as GPT-3.5 or GPT-4, leaving the capabilities of more advanced iterations underexplored. By focusing on ChatGPT-5, this study documents a more current point in the trajectory of AI development. Additionally, while much existing evidence draws from general online samples, this project's focus on undergraduates aged 18-25 in a controlled laboratory setting provides more targeted evidence relevant to educational policy. The study also addresses underexplored questions around metacognitive conflict. Specifically, whether AI-assisted students generate ideas without adequately evaluating them and the risk of ideational homogenization, whereby AI anchors all users toward statistically similar outputs, reducing collective diversity even as it supports individual performance. It is important to note the boundaries of this inquiry. The study is limited to a single, text-based assessment of creative output and does not track long-term creative development, evaluate collaborative dynamics between humans and AI, or extend to non-textual forms of ideation. Within these parameters, the research aims to determine whether generative AI functions as a supplement to or a substitute for students' creative cognition.

4. Methodology

4.1. Design of the Experiment

For this experiment, students ($N = 50$) were recruited via outreach efforts conducted by the experiment organizers. Inclusion criteria were: adults aged 18 or older, current students, prior experience using AI tools, completion of all survey sections with at least 95% valid responses, proficient English speakers, and born/raised in Kazakhstan.

Demographic characteristics are described in detail. The age range of participants was 18–34. In the AI group ($n = 25$), 44% were female, and 56% were male. In the non-AI group ($n = 25$), 28% were female, and 72% were male.

To control for individual differences in creativity and ability, we focused on participants with similar educational backgrounds, specifically selecting only SEDS graduate students for the experiment. MEM, Master of Engineering Management, was a predominant major for the selected participants, while their Bachelor's degree major differed. This was done to control for and minimize the influence of different educational experiences that could otherwise influence creativity. Creating a baseline group that controlled for all factors such as prior knowledge of AI or personal abilities was not possible due to time constraints. However, selecting a homogeneous group of participants helps isolate the impact of AI on creativity. In other words, by ensuring that

all participants have similar academic backgrounds, we aim to more accurately assess the impact of AI without the confounding influence of different experiences.

The average time to complete the survey was 6 minutes.

Before participating, all examiners completed CITI training on Social & Behavioral Research. Participants provided written informed consent prior to starting the experiment.

4.1.1. No-AI Group Procedure

Participants in the no-AI condition were invited, where each participant was seated at a computer of equal computational capacity and completed tasks via Google Form. They first signed consent forms and provided demographic information.

For divergent thinking, participants completed the Alternative Uses Task (AUT) (Hubert, Awa & Zabelina, 2024). The instructions were as follows:

“For this task, you’ll be asked to come up with as many original and creative uses for [item] as you can. The goal is to generate creative ideas that are clever, unusual, interesting, uncommon, humorous, innovative, or different. Ideas do not need to be practical or realistic; they can be silly or strange, as long as they are CREATIVE. You may enter as many ideas as you like. The task will take 3 minutes. Creative quality is more important than quantity; it is better to have a few truly original ideas than many uncreative ones. List as many ORIGINAL and CREATIVE uses for [item] as possible.”

Responses were entered into a text box, recorded using Google Form, and automatically exported to Excel for analysis. This design prioritized Originality over Fluency.

4.1.2. AI Group Procedure

Each participant in the AI condition was instructed to open ChatGPT5 without signing in. AI responses were structured to match the mean of no AI group on timed tasks. For example, if a group with human participants will produce 6 ideas on average, the AI group will be instructed to produce exactly six responses. Only valid responses were retained.

Because originality and other divergent-thinking quality scores are mathematically and methodologically confounded with fluency, and high correlations with fluency are often artifactual rather than indicative of unidimensional creativity, matching AI and human responses on fluency in timed tasks is a necessary control to isolate idea quality from idea quantity and avoid volume-driven distortions in originality estimates (Forthmann et al., 2020; Puente-Díaz et al., 2025)

Instructions for the AI group were adapted from the AUT prompts, with the number of responses limited to match human fluency:

“For this task, you’ll be asked to come up with [insert fluency number] ORIGINAL and CREATIVE uses for [item]. Ideas should be clever, unusual, interesting, uncommon, humorous, innovative, or different. Practicality is not required.”

4.2. Creativity Measures Assessed

4.2.1. Alternative Uses Task (AUT)

To assess divergent creative thinking, the AUT was implemented using common objects, such as a fork and a pencil. Responses were evaluated on three dimensions:

- Fluency: number of responses. It was controlled in AI sessions to match the average output of the No-AI group in order to eliminate quantity-driven advantages and ensure a fair comparison of quality (described in details in Section 2.2.1).
- Originality: uniqueness of responses.
- Elaboration: number of words per valid response.

This measure was excluded from the analysis:

- Flexibility: variety of perspectives. This measure was excluded due to its reliance on consistent category construction and judge-based coding, as described in Section 2.2.3. This process introduces additional subjectivity and methodological complexity, and was therefore out of the scope of this study.

Given these considerations, the analysis focuses on originality and elaboration as primary indicators of creative performance.

4.2.2. Creativity scoring

The performance on divergent thinking tasks is usually measured with human judges, but this process can be subjective and variable, as well as difficult to scale. For this reason, researchers have developed various scoring algorithms to automate the assessment process (Organisciak et al., 2023; Xie et al., 2024).

In the current study, semantic originality was first assessed with GloVe embeddings. Following existing practices, each response and prompt was mapped into semantic space as vectors, and originality was estimated as their cosine distance. GloVe is a model trained on statistics of word co-occurrences, encoding meaning in semantic relationships between them (Pennington et al., 2014).

While earlier studies used the 840B version of GloVe (Koivisto & Grassini, 2023), this experiment leverages its 2024 update, built on Dolma, which was trained on an even larger dataset with 220 billion tokens, including a vocabulary of 1.2 million unique words and 300-dimensional vectors.

When developing the experiment, only GloVe embeddings were used to estimate originality in the first analysis. However, during the presentation, the jury advised that

word-level similarity cannot capture the conceptual link and creativity of participants' thoughts, necessitating the development of a new methodological direction.

Therefore, the Open Creativity Scoring AI (OCSAI) was employed to measure the perceived originality of divergent thinking tasks. OCSAI consists of fine-tuned large language models evaluating responses as a whole to consider conceptual and contextual features of creativity, proven to be highly correlated with human judgments (exceeding 27,000 responses, $r = .81$) (Organisciak et al., 2023).

As a result, originality scores were calculated with two different techniques: (1) semantic distance obtained from GloVe embeddings and (2) scores estimating perceived creativity using OCSAI.

Before scoring, all responses underwent preprocessing (translations when required), cleaning, and standardization. In addition, the divergent thinking task included two common objects, "pencil" and "fork," to create diversity in functional affordances and possible creative applications.

5. Data Analysis

5.1. Data Preparation

The responses were collected via Google Form for No-AI groups and AI allowed groups. The participants entered their responses to the Google Form fields. Two datasets were developed for the AUT, collecting responses for the prompts "pencil" and "fork." One dataset comprised responses from human participants, and the other contained AI-generated/assisted responses. No-AI group responses were collected, and it was derived that the average responses for both items was around 5. Thus, the prompt for the AI group capped a fluency number, "use cases", to 5.

The collected data underwent several steps of cleaning and standardization. Some responses were originally submitted in Russian and were translated into English for compatibility. Once all responses were translated, duplicate, non-viable (no consent, no "use cases"), and irrelevant submissions (e.g., set of meaningless letters) were filtered out.

Elaboration was operationalized as the amount of meaningful detail in each response. To calculate this, each response was first converted to lowercase and cleaned by removing punctuation. The text was then tokenized into individual words.

Common English stop words were removed using the stop-word list from spaCy's English language module. This list includes frequent function words such as "the,"

“and,” “to,” “of,” “in,” and similar terms, as well as contractions such as “n’t,” “’d,” “’ll,” and related apostrophe variants.

Original code:

https://github.com/explosion/spaCy/blob/a741de7cf658ce9a90d7afe67c88face8fb658ad/spacy/lang/en/stop_words.py

After preprocessing, the elaboration score for each response was calculated as:

Elaboration score = number of remaining non-stopword tokens

For example, a short response such as “bookmark” receives a low score, while a more developed response such as “use a fork to loosen soil in small pots” receives a higher score because it contains more meaningful descriptive content.

Mean elaboration scores were calculated separately for pencil and fork, and then combined into an overall mean elaboration score.

AI-generated responses were often long-form, containing descriptions, jokes, metaphors, or stories.

When originality was assessed, in order to compare them to the human responses, the text was trimmed to short phrases, capturing only the essence of the tool. For example, “Micro-Dough Rolling Pin: The perfect culinary tool for flattening out the dough for exactly one ravioli, a single miniature tortilla, or a bite-sized pie crust” was simplified to a tool for flattening dough. This trimming was done for the participants who reported that the answers were mostly copy pasted and slightly modified, which is around 64% of AI group respondents. Therefore, the trimming was essential to filter out unnecessary characters and invalid information.

In total, 241 responses generated by AI were collected, and all were determined to be valid, but 54.4% of these responses required further preprocessing to trim down to the relevant information. On the other hand, 93.1% of human responses were found to be valid, while 20.6% of responses required trimming. Only 6.9% of the human responses could not be used due to being irrelevant or rude. For example, one of the respondents mentioned that “split in half and get 2 pencils” or that the fork could be used “for chebureks (Russian fried dumplings)”.

All responses, both AI and human, underwent preprocessing using natural language processing methods. Responses were lowercased, punctuations removed, and responses were tokenized into words. Stop words (such as the, and, of) were stripped from

responses. Words not available in GloVe's vocabulary were dropped from consideration when generating sentence embeddings.

5.2. Analysis

After the preprocessing stage, semantic originality was estimated using the GloVe embeddings. Tokenization was performed for all responses, followed by vectorization through calculating the average of the embeddings of the tokens left after removing stopwords and vocabulary filtering. For each prompt, a vector was also obtained in a similar manner. The similarity between the vectors of response and prompt was measured using cosine similarity, while semantic distance was defined as 1-cosine similarity, with larger distances indicating a bigger difference from the original object.

The scoring used the 2024 Dolma-based GloVe embeddings, 300-dimensional and trained on a large corpus. In order to avoid unnecessary computations, only the most frequently used tokens were used to represent the vocabulary, and words that did not have any embeddings were ignored.

Perceived creativity was estimated using the OCSAI scoring approach. A model evaluated every response and assigned it an originality score that indicated the degree of creativity of the response. Originality scores were first calculated at the response level and averaged across participants and prompts.

OCSAI was calculated using ocsai2 – a cross-lingual, multi-task model based on fine-tuned large language models. The originality score was calculated as a weighted probability. The task type was "uses," with short prompt description and English language parameters in order to match the AUT task.

Response-level scores were averaged at the participant level for each prompt (i.e., "pencil" and "fork"). Then group-level averages for human- and AI-produced responses were calculated for further analyses.

6. Results

6.1. Fluency

Figures 5.1.1 and 5.1.2 showcase the spread of the ideas for both words, with a mean of 5.62 for pencil and 5.04 for fork. Therefore, it was decided to set the fluency score of the AI group to 5. Interestingly the word that came second in the survey, fork, had a significantly lower median of 3.5 (compared to 5.5). Yet, the same word gathered 10 (max) use cases from 6 students, while the pencil got 4.

The results reveal that, although the mean fluency scores for the two objects are similar, the underlying distributions differ substantially. The pencil illustrates a more balanced and consistent pattern of idea generation, with the mean and median closely aligned (5.5 and 5.62), suggesting that most participants were able to produce a moderate number of uses. This likely reflects the object’s familiarity and the breadth of its conventional associations (e.g., writing, drawing, marking), which make it easier to extend into related functions. In contrast, the fork demonstrates a more polarized distribution, characterized by a lower median but the presence of high-performing individuals who reached the maximum number of ideas. This indicates that while many participants remained constrained by the object’s primary function, a subset was able to overcome this limitation and generate highly divergent uses.

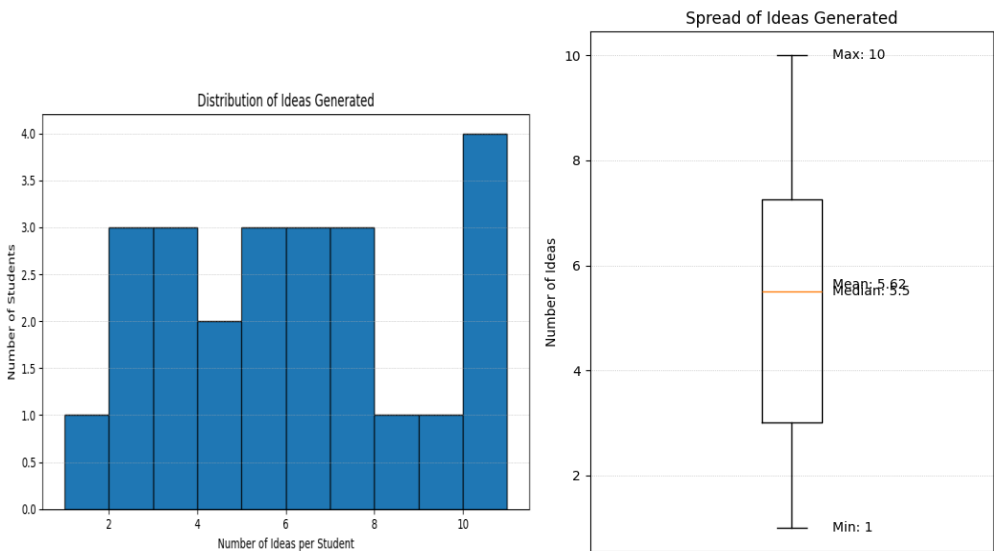
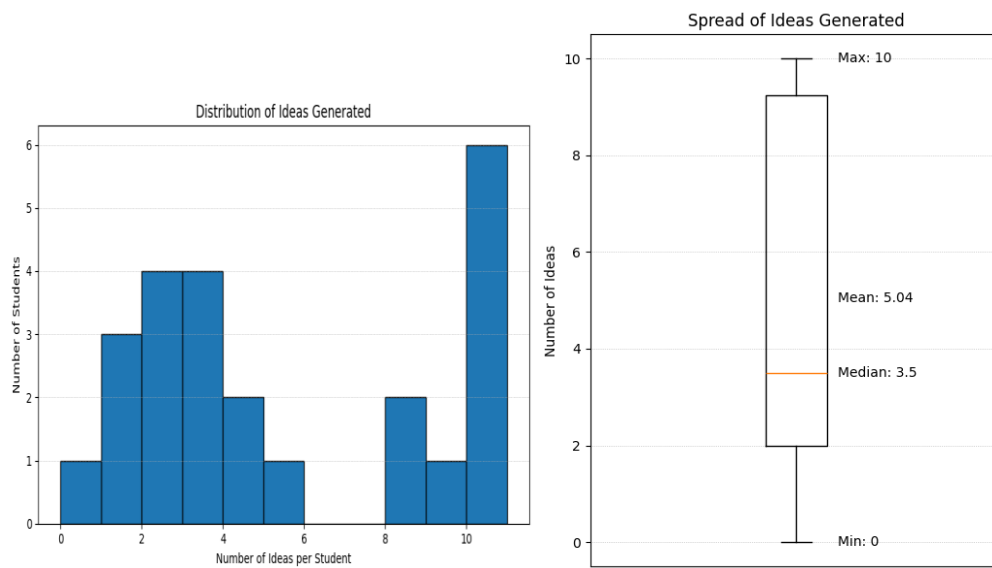


Figure 6.1.1. The distribution of responses and spread for “pencil”**Figure 6.1.2. The distribution of responses and spread for “fork”**

Because fluency score (# of ideas) was experimentally controlled in the AI condition, it cannot be meaningfully compared between groups. In contrast, fluency in the No-AI group was unconstrained and varied across participants; therefore, its relationship with participant characteristics was analyzed using correlation analysis. For participant characteristics, variable parameters like English proficiency, AI usage frequency, and gender has been taken. The results indicate that English proficiency demonstrated the strongest association with idea output as it is shown in Table 6.1.3.

	English_num	AI_freq_num	Gender_num	Ideas
English_num	1.000000	-0.101409	-0.511364	0.401188
AI_freq_num	-0.101409	1.000000	0.084324	-0.056078
Gender_num	-0.511364	0.084324	1.000000	-0.209010
Ideas	0.401188	-0.056078	-0.209010	1.000000

Table 6.1.1. Correlation Matrix

Specifically, the Pearson correlation between English level and number of ideas was $r = 0.40$, while the Spearman rank correlation was $\rho = 0.39$ ($p = 0.059$), indicating a moderate positive relationship that approaches, but does not reach, conventional levels of statistical significance. Descriptive statistics further support this pattern: mean idea scores increased consistently across proficiency levels, from 3.0 (A1) to 3.8 (B2), 6.13 (C1), and 7.0 (C2). This suggests a clear upward trend in idea generation as English proficiency improves.

In contrast, AI usage frequency showed no meaningful association with idea output. The Pearson correlation was $r = -0.06$, and the Spearman correlation was $\rho = -0.12$ ($p = 0.576$),

indicating a negligible and non-significant relationship. Group means also did not reveal a consistent pattern: participants using AI daily generated on average 5.5 ideas, compared to 6.67 for weekly users and 5.0 for monthly users. These variations appear unsystematic and are likely influenced by small and uneven subgroup sizes rather than a true underlying effect.

Additional variables showed limited explanatory power. Gender exhibited a weak negative correlation with idea output ($r = -0.21$), with males averaging 6.0 ideas and females 4.71, though this difference was not statistically significant. Age differences were minimal, with participants aged 18-24 producing 5.57 ideas on average, compared to 6.0 among those aged 25-34.

Overall, these findings suggest that English proficiency may play a role in supporting higher idea generation potentially by enabling broader semantic access or more effective interaction with prompts. AI usage frequency alone does not appear to enhance creative output in this sample. The absence of statistically significant results across most tests, however, highlights the limitations of the small sample size and indicates that these conclusions should be interpreted as exploratory rather than definitive. Refer to Appendix E for extended results.

6.2. Elaboration

Mean elaboration scores were summarized in Table 6.2.1 and Figure 6.2.2.

Participants	AI			no AI		
	mean_elab_pencil	mean_elab_fork	mean_elab_total	mean_elab_pencil	mean_elab_fork	mean_elab_total
1	3,6	3,6	3,6	0,4	0,1	0,25
2	3,4	13,8	8,6	0,5	0,4	0,45
3	1,2	1,4	1,3	2,3	1,8	2,05
4	3,4	3,2	3,3	1,1	0,8	0,95
5	5,4	6,6	6	0,7	0,3	0,5
6	17,4	18,8	18,1	2,3	2,2	2,25
7	2,4	2	2,2	1,3	1,4	1,35
8	2,4	1,4	1,9	0,2	0,3	0,25
9	16,6	15,2	15,9	2,5	1,9	2,2
10	14,6	14,4	14,5	0,4	0,2	0,3
11	4,2	3,8	4	2,3	1,6	1,95
12	16	15,4	15,7	1,8	0	0,9
13	21	31,8	26,4	0,3	0,1	0,2
14	3	3,2	3,1	1	0,3	0,65
15	17,2	18,8	18	0,5	0,3	0,4
16	1	2,8	1,9	0,7	0,5	0,6
17	19	15	17	2	0,7	1,35
18	3,2	3,2	3,2	1,1	2,1	1,6
19	3,2	2,6	2,9	1,7	2,5	2,1
20	4,2	32,6	18,4	2,7	2,2	2,45
21	2,6	1,4	2	1,9	2	1,95
22	37,2	13,2	25,2	0,8	0,6	0,7
23	3,8	3,4	3,6	1,7	1,9	1,8
24	10,8	3	6,9	2,4	2,5	2,45
25	4,2	0	2,1	0,3	0,7	0,5
mean	8,84	9,224	9,032	1,316	1,096	1,206

Table 6.2.1. Participant Elaboration Scores

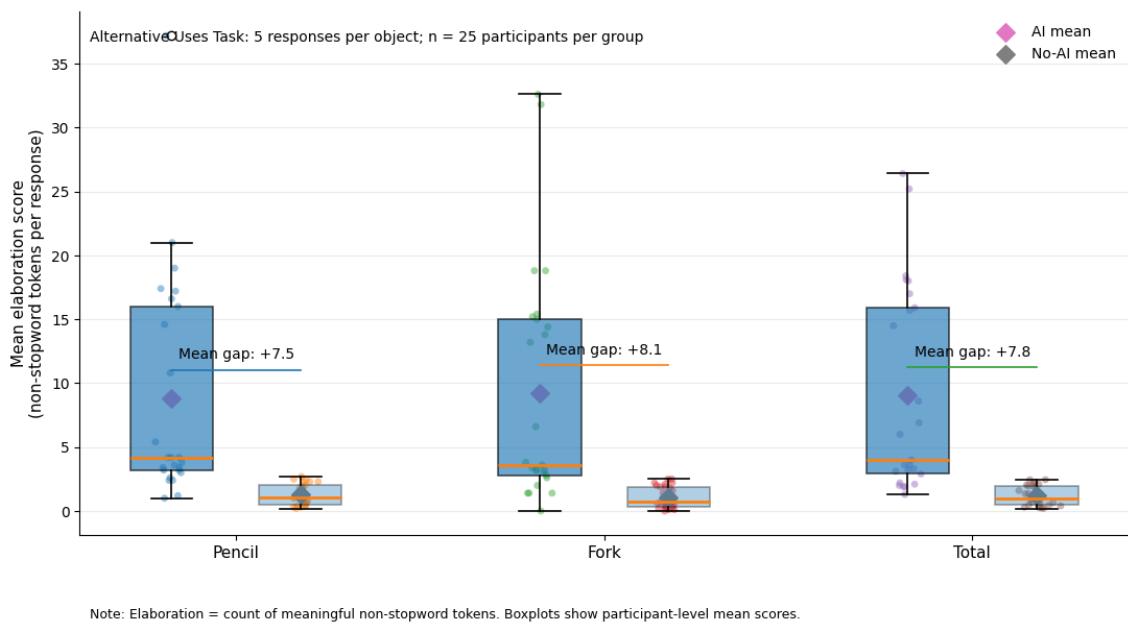


Figure 6.2.1. Elaboration Score Means by AI and no AI groups

The AI-assisted group showed substantially higher elaboration than the No-AI group, in other words, AI-assisted responses contained much more descriptive detail.

Welch’s t-tests revealed that elaboration scores were significantly higher in the AI-assisted condition compared to the no-AI condition for pencil ($t = 4.24, p < .001$), fork ($t = 4.36, p < .001$), and total responses ($t = 4.86, p < .001$).

Qualitatively, the No-AI group mostly produced short labels or brief phrases (like eat, write, scratch, sew), while the AI-assisted group more often produced developed explanations, examples, or scenario-based uses (e.g. The Low-Tech "Air-Quality" Sensor. If you are suspicious of the humidity or "vibes" in a room, simply shave a pencil down until you have a long, thin ribbon of cedar wood. Suspend it from the ceiling. Because cedar is sensitive to moisture, the ribbon will curl or straighten based on the humidity). This suggests that AI supports increased elaboration by expanding responses beyond simple idea labels into more detailed descriptions. Moreover, the divergence between mean and median in the AI condition indicates a right-skewed distribution, suggesting that while most participants produced moderately elaborated responses, a subset generated highly detailed outputs.

Due to gender imbalance in groups, stratified analysis has been conducted to check if it affected the results of the study. Stratified analyses by gender showed that the effect of AI assistance on elaboration is robust across both groups.

For female participants, elaboration was higher in the AI condition than in the no-AI condition across all tasks: pencil (6.78 vs. 1.23, $\Delta = +5.55$, $t = 2.94$, $p = .014$), fork (7.65 vs. 1.00, $\Delta = +6.65$, $t = 3.34$, $p = .007$), and total responses (7.22 vs. 1.11, $\Delta = +6.10$, $t = 3.25$, $p = .008$).

For male participants, the same pattern was observed, with larger differences: pencil (10.46 vs. 1.35, $\Delta = +9.11$, $t = 3.26$, $p = .006$), fork (10.46 vs. 1.13, $\Delta = +9.32$, $t = 3.15$, $p = .008$), and total responses (10.46 vs. 1.24, $\Delta = +9.22$, $t = 3.74$, $p = .002$).

Importantly, baseline elaboration levels in the no-AI condition were similar across genders, indicating that gender alone does not account for differences in performance. However, the magnitude of the AI effect was consistently larger among male participants, suggesting a potential interaction between group and gender, where AI assistance amplifies elaboration more strongly in this subgroup.

Another step was to check how the degree of transformation of the response in the AI group has affected the elaboration score. Figure 6.2.2. showcase the responses of our participants on the question “*To what extent did you modify or build upon ChatGPT’s ideas when creating your final answers?*” mapped with their elaboration score (total for two words: pencil and fork). Important note: among 25 participants only 1 had fully original ideas compared to 12 participants who mostly copied the ideas from ChatGPT.

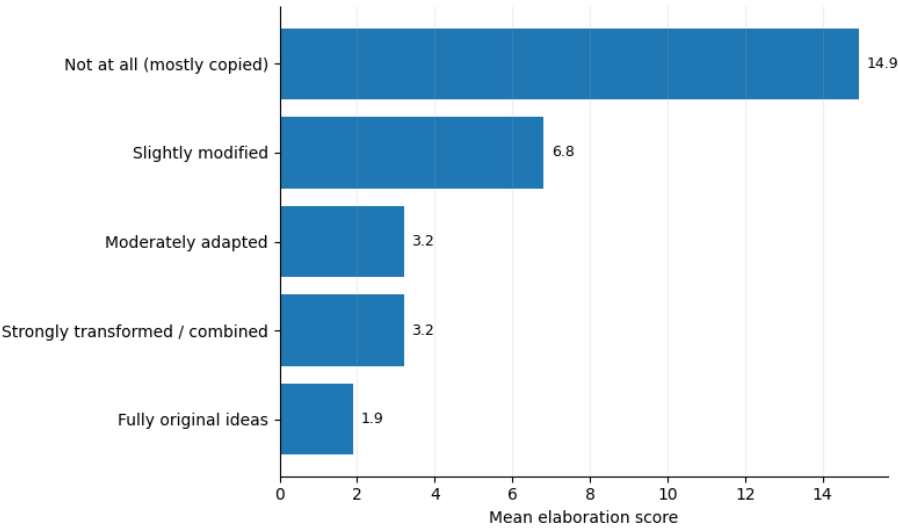


Figure 6.2.2. Mean Elaboration Scores for AI groups based on the degree of transformation

6.3. Originality (GloVe and OSCAI)

This part describes the findings of a comparative analysis of semantic distance and originality scores of human versus AI-generated responses according to both prompts. Table 6.3.1 provides the average score of each group in each prompt according to the measures.

Group	Prompt	GloVe	OSCAI
No-AI group	Fork	0.504	2.991
No-AI group	Pencil	0.470	2.885
AI assisted	Fork	0.466	3.580
AI assisted	Pencil	0.444	3.239

Table 6.3.1. The mean value of GloVe and OSCAI scores for groups

As shown in Table 6.3.1, the results indicate that human-generated responses achieved higher scores in terms of semantic distance according to both prompts when comparing the performance of humans and AI systems. As shown in Table 6.3.1, in the "fork" condition, human responses produced a significantly higher average semantic distance score than AI responses (Mean = 0.504 versus Mean = 0.466; $t = -2.426$, $p = .019$). Similarly, in the case of the pencil prompt, although the average scores differed (Mean = 0.470 versus Mean = 0.444), this difference was not statistically significant ($t = -1.308$, $p = .198$).

Contrastingly, as seen in Table 6.3.1, the results obtained using the OSCAI measure demonstrated the opposite pattern. In all cases, human-generated responses achieved lower scores than the scores of AI responses. This can be exemplified through examining the "fork" condition, where human responses achieved an average score of 2.991 versus the average score of AI responses of 3.580, which was a statistically significant difference ($t = 4.181$, $p < .001$). In the case of the pencil prompt, human responses achieved a score of 2.885 compared to the average score of AI responses of 3.239, representing a difference that was statistically significant ($t = 3.476$, $p = .001$).

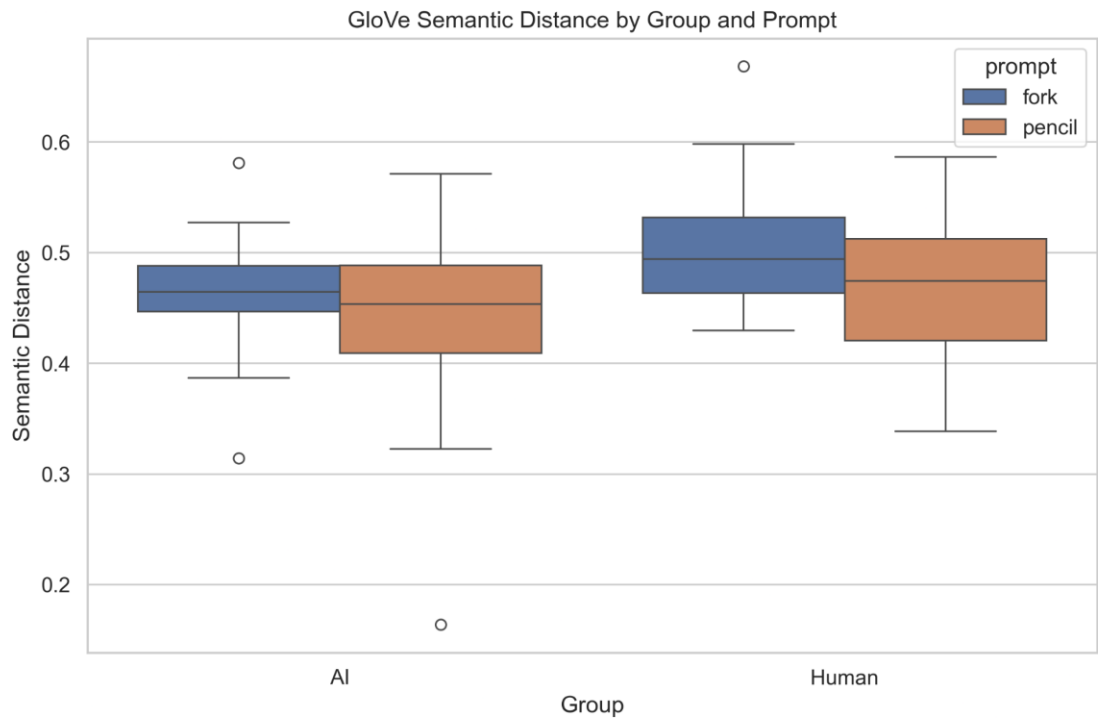


Figure 6.3.1. GloVe semantic distance by Group and Prompt

Figure 6.3.2. shows the distribution of semantic distance scores according to the prompts and groups. It can be observed that the boxplot shows that humans generate more scattered scores with the tendency to produce higher scores, especially in the "fork" prompt. In contrast, AI-generated responses show less variability within a narrow range of scores.

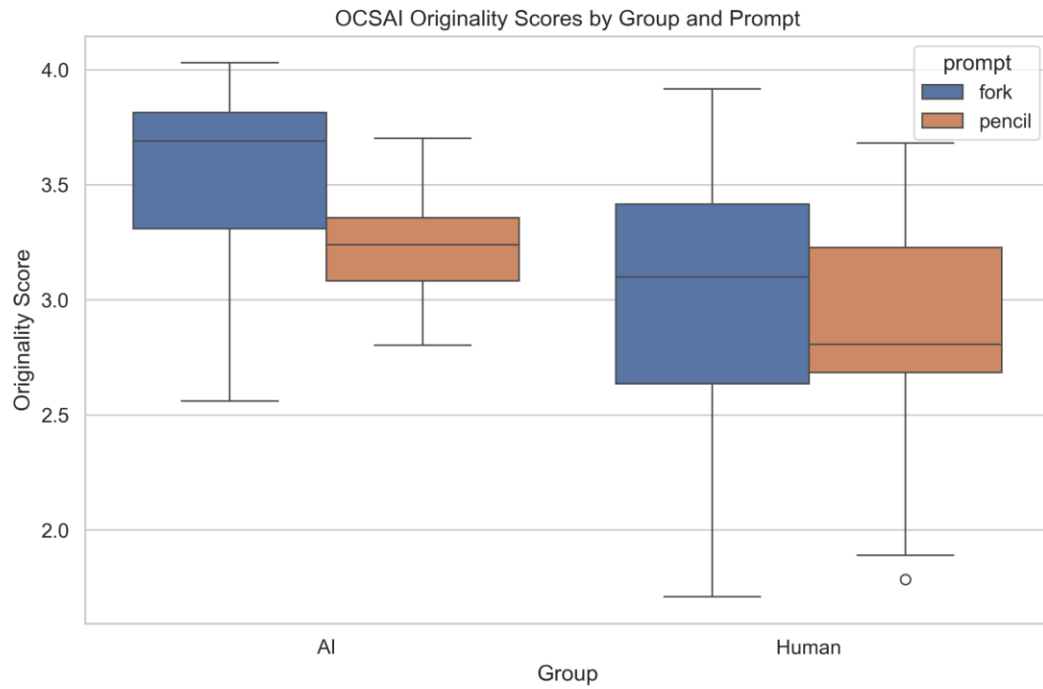


Figure 6.3.2. OCSAI by Group and Prompt

Figure 6.3.3 illustrates the same comparison, but according to the OCSAI measure. It can be seen that AI responses produce higher scores of central tendency than human responses according to both prompts. There is some overlap between distributions of scores; however, the separation of distributions between groups is higher according to the OCSAI metric.

It seems that another pattern could be identified in terms of prompts used. Responses to "fork" achieve higher average scores according to both metrics than responses to "pencil," indicating that objects influence the range of creativity.

Gender analyses were performed using averages at the participant level to ensure that each participant contributed one data point for each prompt. The number of female participants in the AI-assisted condition was 11 compared to 14 males, while the No AI group consisted of 6 females and 18 males.

The means of GloVe and OCSAI scores for all groups and genders are given in Table 6.3.2. Within the AI-assisted group, the average GloVe score among female participants was 0.479, while their OCSAI score was 3.434. Male participants' mean GloVe score in the AI-assisted group was 0.436 and their mean OCSAI score was 3.390. Female participants in the No AI group had an average GloVe score of 0.503 and their average

OCSAI score was 3.016. Male participants in the No AI group had a mean GloVe score of 0.483 and their mean OCSAI score was 2.931.

Group	Gender	n	GloVe avg	OSCAI avg
AI-assisted	Female	11	0.479	3.434
AI-assisted	Male	14	0.436	3.390
No AI	Female	6	0.503	3.016
No AI	Male	18	0.483	2.931

Table 6.3.2. The mean value of GloVe and OSCAI scores based on genders

Independent sample t-tests (Welch) were used on the data of both women and men for AI-assisted and No AI conditions. In the case of women, no statistical significance was found between the two conditions in terms of GloVe scores, $t = -1.039$, $p = .323$. On the other hand, OCSAI scores were found significantly higher in the AI-assisted condition ($M = 3.434$) than in the No AI condition ($M = 3.016$), $t = 2.608$, $p = .037$.

In the case of men, the No AI condition gave rise to significantly higher GloVe scores ($M = 0.483$) than the AI-assisted condition ($M = 0.436$), $t = -3.167$, $p = .005$. In contrast, AI-assisted conditions generated significantly higher OCSAI scores ($M = 3.390$) than the No AI conditions ($M = 2.931$), $t = 3.245$, $p = .003$.

Gender	Metric	t	p
Female	GloVe	-1.039	0.323
Female	OSCAI	2.608	0.037
Male	GloVe	-3.167	0.005
Male	OSCAI	3.245	0.003

Table 6.3.3. The t-test results

In order to determine if the degree of modification of response by users impacted the creativity scores, the participants using AI assistance were divided into two categories. The first category was called "No/slight modification," which includes participants that mostly copied the AI answer or made slight modifications to it. The second category was termed "Moderate/high modification," and it involved participants who slightly or largely adapted the response or came up with something new.

Modification group	n	GloVe average	OCSAI average
No/Slight modification	16	0.454	3.448
Moderate/High modification	9	0.457	3.340

Table 6.3.4 The modification level results for AI assisted group

The mean GloVe scores were relatively equal between the two groups, with the Moderate/high modification group having a slightly higher mean (0.457) than the No/slight modification group (0.454). On the other hand, the mean score was higher among the No/slight modification group (3.448) than the Moderate/high modification group (3.340) when considering the OCSAI.

Metric	t	p
GloVe	-0.174	0.864
OCSAI	0.980	0.343

Table 6.3.5. The t-test results for modification

According to Welch independent-samples t-test results, neither difference reached statistical significance in the case of GloVe, $t = -0.174$, $p = .864$, nor OCSAI, $t = 0.980$, $p = .343$. Therefore, modification level was not found to be a significant predictor of creativity scores.

Overall, the results reveal a divergence between the two evaluation strategies. Human responses showed greater semantic distance from the object, whereas AI-generated responses were rated as more creative according to the OCSAI measure.

7. Discussion

The results of this study show that AI assistance affected the originality dimension of creativity differently depending on the evaluation method. In particular, the AI-assisted responses did not score above average on GloVe-based semantic distance but scored higher on the OCSAI originality index. However, this does not mean that the two indices are equally representative of creativity and can be used interchangeably as such. GloVe calculates the semantic distance between the prompt and response, while OCSAI aims at approximating human evaluations of originality. Hence, while evaluating creativity originality in this case, it is best to take OCSAI as the more credible indicator.

It means that the main conclusions from this analysis need to be somewhat revised in light of the

methodological considerations raised. In particular, lower values for the AI assistance group concerning GloVe do not actually mean reduced originality since they are indicative of the smaller semantic distance from the original object. At the same time, higher values for OCSAI imply increased originality as perceived by this more reliable indicator of creativity. In other words, AI assistance increased originality in terms of the chosen measure.

At the same time, these conclusions appear to be consistent with the characteristics of the two tools for measuring originality. Namely, GloVe is a neural word embedding model that uses co-occurrence statistics to represent a word as a vector. As used in this analysis, response vectors were created using averaging the word embeddings within a response. While it makes possible a quick assessment of semantic distance, it also does not take into account the context of words, their functions, and general meanings. Thus, GloVe can identify semantic remoteness but not originality, usefulness, or reinterpretation of the object. This explains why human responses were generally characterized by greater semantic distance according to GloVe despite being rated by humans as less original using OCSAI.

OCSAI, on the contrary, provides a more comprehensive evaluation of the originality of creative works. Being a model-based measure trained to approximate human evaluations of the target feature, it provides more appropriate estimates in regard to perceived creativity. Higher originality scores associated with AI assistance imply that it enabled participants to generate more coherent, elaborate, and recognized responses that can be regarded as original. This result is aligned with research demonstrating positive effects of generative AI in supporting creative thinking and improving creative outputs (Urban et al., 2024) and showing how AI assistance influences ideation and creative output diversification (Ashkinaze et al., 2025).

It should be added that preprocessing differences are relevant to discuss here as well. More precisely, responses generated under AI assistance tended to be longer, more descriptive, and more narrative than those in the No AI condition. Responses where users copied AI text were cut down into shorter functional phrases in order to make it comparable to human AUT responses. It means that AI-assisted data required more preprocessing before analysis compared to data collected among No AI respondents.

Nevertheless, additional analysis of modification-level preprocessing shows that preprocessing did not lead to a difference in the main result. Namely, in the AI assistance group, participants were separated into the No/slight modification and Moderate/high modification categories based on how extensively they altered initial ChatGPT-generated text. There were no statistically significant differences between these two categories in relation to GloVe and OCSAI measures. Hence, even if there is a preprocessing difference between AI assistance and No AI groups, the results are unlikely to be impacted by it.

Finally, several practical implications should be discussed. First, as seen, AI assistance can be useful to increase perceived originality at least if evaluated with measures based on human

assessments. This means that it may be helpful for assisting in generating more creative ideas that are rated higher in terms of originality. Second, lower GloVe scores imply that AI assistance will not necessarily drive users to come up with more semantically remote ideas but instead help to improve idea development and articulation without moving far away from established concepts.

Thus, it can be concluded that AI assistance should be understood as helping with different aspects of creativity rather than just promoting or decreasing it. In this specific case, AI assistance made responses perform better in OCSAI originality, while the lack of assistance promoted more divergent ideas according to GloVe semantic distance. Hence, AI seems to assist with improving originality and quality, while people on their own are capable of generating more diverse ideas.

In turn, these findings demonstrate the potential benefits of using generative AI to enhance creativity performance as long as participants receive appropriate instructions. In education and work-related settings, AI should be primarily used as a tool for generating ideas and developing responses rather than directly copy-and-pasting texts. Finally, this case emphasizes the importance of using different measures to assess creativity. Had GloVe been used only, one would reach the conclusion about better human performance. However, since OCSAI is more aligned with human evaluations, the conclusion should be drawn about positive effects of AI assistance.

From the results obtained in this study, it is evident that the AI-assisted responses performed better on OCSAI originality index while performing worse in GloVe semantic distance. However, an interesting theory question can be posed from this regarding whether or not this is in line with fixation bias or cognitive offloading.

With the lower GloVe index scores, it means that the AI-generated response was closer to the original object. Therefore, the role of the AI was crucial in ensuring that the participant focused on elaborating and evaluating the concept. As explained, cognitive offloading refers to a situation whereby AI assumes the routine process of task completion and leaves high-level thinking to humans. With this, it can be assumed that the participants' minds had been cleared and therefore able to evaluate and elaborate.

At the same time, the low performance of the model in the GloVe metric could also be explained by the fixation bias. The AI is converging toward common solutions, offering concepts that are closer to the initial idea. It means that AI-driven generation might have caused participants to fixate on more conventional concepts, hindering divergent thinking processes. Therefore, AI-driven generation might lead to fixation instead of promoting divergent thinking.

In summary, the findings indicate a dual effect of artificial intelligence on creativity. On the one hand, AI contributes to the optimization of cognitive resources (supports elaboration processes) and enhances originality. On the other hand, AI could reduce the scope of ideation, which coincides with the notion of fixation bias.

8. Conclusion

The current research aimed to analyze how GenAI impacted creativity. Particularly, it sought to evaluate whether GenAI could influence creative behavior concerning creativity fluency, flexibility, originality, and elaboration. It appeared that the use of GenAI resulted in better originality and elaboration as assessed via OCSAI. Nevertheless, AI-assisted answers had lower scores on GloVe semantic distance, which means that GenAI did not foster ideas that were more semantically distant from each other. Such findings suggest that while AI can improve the quality of ideation, it can have adverse implications for creativity due to the narrowing of ideation scope and the creation of the fixation bias. In such a way, AI can enhance creativity only partly, while its use poses risks of offloading some creative functions to AI and thus creating fixation biases among people.

However, several limitations regarding the analysis of the results obtained in this study should be discussed. First, the non-blind nature of the pre-processing of AI-generated ideas may introduce some bias, as participants may modify or adapt the AI-generated ideas in some way. Furthermore, a gender disparity was observed in the analyzed samples, favoring men in both the experimental and control groups. The study was conducted using a single-session approach, and a baseline level of creativity was not determined, making it difficult to track the evolution of creativity over time or in response to the use of AI. Furthermore, no validation procedure was used to ensure that participants in the experimental sample were using ChatGPT to create something genuinely new.

In terms of future research, efforts need to be made to implement blinded preprocessing in order to avoid bias and get accurate results. A representative sample of study participants needs to be considered for the purpose of avoiding biased samples. Implementation of pre and post-tests is important since it will help to understand the impact of AI on creativity in the long run. It is also critical to make sure that participants actually engage with AI (as opposed to copying its output).

References

- [1] Arriagada, L. (2020). CG-Art: Demystifying the anthropocentric bias of artistic creativity. *Connection Science*, 32(4), 398–405.
<https://doi.org/10.1080/09540091.2020.1741514>
- [2] Ashkinaze, J., Mendelsohn, J., Li, Q., Budak, C., & Gilbert, E. (2025). How AI ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. In *Proceedings of the ACM Collective Intelligence Conference (CI 2025)*. <https://doi.org/10.1145/3715928.3737481> (new)
- [3] Camarda, A., De Neys, W., Ozkalp-Poincloux, B., Hooge, S., Le Masson, P., Weil, B., & Cassotti, M. (2024). Detecting fixation bias in creative idea generation: Evidence from design novices and experts. *Creativity Research Journal*, 38(1), 189–209. <https://doi.org/10.1080/10400419.2024.2424620> (new)
- [4] Carlson, R., Bauer, J., & Manning, C. D. (2025). *A new pair of GloVes*. arXiv. <https://arxiv.org/abs/2507.18103>
- [5] Desdevises, J. (2025). The paradox of creativity in generative AI: High performance, human-like bias, and limited differential evaluation. *Frontiers in Psychology*, 16, 1628486. <https://doi.org/10.3389/fpsyg.2025.1628486>
- [6] Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. <https://doi.org/10.1126/sciadv.adn5290>
- [7] Forthmann, B., Szardenings, C., & Holling, H. (2020). Understanding the confounding effect of fluency in divergent thinking scores: Revisiting average scores to quantify artifactual correlation. *Psychology of Aesthetics, Creativity, and the Arts*, 14(1), 94–112. <https://doi.org/10.1037/aca0000235>
- [8] Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6. <https://doi.org/10.3390/soc15010006>
- [9] Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.
- [10] Habib, S., Vogel, T., Anli, X., & Thorne, E. (2023). How does generative artificial intelligence impact student creativity? *Journal of Creativity*, 34(1), 100072. <https://doi.org/10.1016/j.yjoc.2023.100072>
- [11] Haase, J and Hanel, PHP (2023) Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity* 33(3), 100066. <https://doi.org/10.1016/j.yjoc.2023.100066>

- [12] Hubert, K.F., Awa, K.N. & Zabelina, D.L. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Sci Rep* **14**, 3440(2024).
<https://doi.org/10.1038/s41598-024-53303-w>
- [13] Jones, J. (2018, October 26). A portrait created by AI just sold for \$432,000. *The Guardian*. <https://www.theguardian.com/artanddesign/shortcuts/2018/oct/26/call-that-art-can-a-computer-be-a-painter>
- [14] Kaufman, J. C., & Beghetto, R. A. (2009). *Beyond big and little: The four C model of creativity*. *Review of General Psychology*, 13(1), 1–12.
<https://doi.org/10.1037/a0013688>
- [15] Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553.
- [16] Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13, 13601.
<https://doi.org/10.1038/s41598-023-40858-3>
- [17] Kumar, H., Vincentius, J., Jordan, E., & Anderson, A. (2025). Human creativity in the age of LLMs: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). ACM. <https://doi.org/10.1145/3706598.3714198>
- [18] Marrone, R., Cropley, D., & Medeiros, K. (2026). How does narrow AI impact human creativity? *Creativity Research Journal*, 38(1), 150–160.
<https://doi.org/10.1080/10400419.2024.2378264>
- [19] Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). *Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models*. *Thinking Skills and Creativity*, 49, 101356.
- [20] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>
- [21] Puente-Díaz, R., Cavazos-Arroyo, J., & Puente-Díaz, J. (2025). The role of metacognitive experience on divergent thinking. *Journal of Intelligence*, 13(3), 27.
<https://doi.org/10.3390/jintelligence13030027>

- [22] Richards, R. E. (2007). Everyday creativity and new views of human nature: Psychological, social, and spiritual perspectives (pp. xiii-349). American Psychological Association.
- [23] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- [24] Silvia, P. J. et al. (2008) Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychol. Aesthet. Create Arts* 2(2), 68–85. <https://doi.org/10.1037/1931-3896.2.2.68>
- [25] Shukla, P., Bui, P., Levy, S. S., Kowalski, M., Baigelenov, A., & Parsons, P. (2025). De-skilling, cognitive offloading, and misplaced responsibilities: Potential ironies of ai-assisted design. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7. <https://doi.org/10.1145/3706599.3719931>
- [26] Urban, K., et al. (2024). ChatGPT improves creative problem-solving performance in university students. *Computers & Education*, 213, 104967. <https://doi.org/10.1016/j.compedu.2024.104967>
- [27] Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., & Velloso, E. (2024). The effects of generative AI on design fixation and divergent thinking. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3613904.3642919>
- [28] Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., & Lu, Z. (2024). "It felt like having a second mind": Investigating human-AI co-creativity in prewriting with large language models. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), Article 84. <https://doi.org/10.1145/3637361>
- [29] Wang, S., Tao, X., Ma, H., Li, F., & Wu, C. (2025). EEG assessment of artificial intelligence-generated content impact on student creative performance and neurophysiological states in product design. *Frontiers in Psychology*, 16, 1508383. <https://doi.org/10.3389/fpsyg.2025.1508383>

Appendix A: CITI Training Certificates

A.1 Ayaulym Ayazhan



Completion Date 30-Jan-2024
Expiration Date 30-Jan-2028
Record ID 60910339

This is to certify that:

Ayaulym Ayazhan

Has completed the following CITI Program course:

Responsible Research Training
(Curriculum Group)
Social, Behav, Edu, Etc
(Course Learner Group)
1 - Basic Course
(Stage)

Under requirements set by:

Nazarbayev University

Not valid for renewal of
certification through CME.

CITI
Collaborative Institutional Training Initiative
101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Generated on 26-Jan-2026. Verify at www.citiprogram.org/verify/?w9f905b4c-c4bc-4e3c-8904-3622bf9a6fb2-60910339

A.2 Dias Nurmukhambetov



Completion Date 27-Mar-2022
Expiration Date 26-Mar-2026
Record ID 48165479

This is to certify that:

Dias Nurmukhambetov

Has completed the following CITI Program course:

Responsible Research Training
(Curriculum Group)
Engineering (RCR)
(Course Learner Group)
1 - Basic Course
(Stage)

Under requirements set by:

Nazarbayev University

Not valid for renewal of
certification through CME.

CITI
Collaborative Institutional Training Initiative
101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Generated on 16-Apr-2026. Verify at www.citiprogram.org/verify/?w79c65d61-7e1b-4ca4-a010-55beff76a8ab-48165479

A.3 Marzhan Kabzhassarova



Completion Date 28-Jan-2026
Expiration Date 28-Jan-2029
Record ID 74979389

This is to certify that:

Marzhan Kabzhassarova

Has completed the following CITI Program course:

Social & Behavioral Research - Basic/Refresher
(Curriculum Group)
Social & Behavioral Research - Basic/Refresher
(Course Learner Group)
1 - Basic Course
(Stage)

Not valid for renewal of
certification through CME.

Under requirements set by:

Nazarbayev University



101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Generated on 28-Jan-2026. Verify at www.citiprogram.org/verify/?wb6570dd0-6c84-45da-8aca-16d74bf13103-74979389

Appendix B: Experimental Protocol

Setting Up

Before participants arrive, set up each computer with a single browser tab open - the Google Form survey link. The survey consists of three sequential pages:

- Signed consent form
- Demographic information
- Alternative Uses Task (AUT)

Preparation checklist before the session begins:

- Gather participants and divide into 2 groups equally.
- Take a photo together.
- Make sure all devices are ready for use and verify a stable internet connection.
- Confirm the Google Form is open and loading correctly on each screen.
- Check that screens are not frozen and that no accounts are signed into ChatGPT (for the AI condition).
- Ask for no AI participants to use one tab only - Google Form.
- Arrange tables so that computers are spaced apart. No assessment is being made, so cheating is not a concern, but distance helps maintain an appropriate working environment.

Greeting and Consent

Participants will be informed of the venue and time in advance via email. As they are students familiar with the campus, they should be able to navigate to the lab independently. However, meet participants outside the building to greet them and guide them in.

Once everyone has arrived, welcome participants into the prepared classroom. Say:

"Please sit wherever you like, as long as the seat is not taken - one person per computer. In front of you is a computer with a Google Form open. All instructions you need for the experiment are available on the screen."

Direct participants to the first page of the form - the Consent Form. Ask them to read through it carefully, paying particular attention to the sections on Participation and Voluntary Nature and Confidentiality. At the end of the consent form, participants will find the following acknowledgement:

By clicking "I Agree" below, you acknowledge that:

- *You have read and understood the study information.*
- *You voluntarily agree to participate in this study.*
- *You understand that your participation is confidential and that you can withdraw at any time.*

If you choose option "I Don't Agree", you'll exit the experiment right away so please be careful.

Once participants have agreed, thank them and direct them to the second page - the demographic information form. Ask them to fill in their age, gender, and highest level of education, as well as their academic major, level of English proficiency (ability to read and understand English), and a general question: *How often do you use Generative AI?* (options: Daily, Weekly, Once a month, Never).

Additionally, at the end of the section make sure to note which group are you in AI or no AI.

Once demographics are complete, confirm with participants that they are ready and proceed to the main task.

Instructions and Task - Non-AI Group

The third section of the Google Form contains the Alternative Uses Task (AUT). Two objects will be used: a pencil and a fork. Display images of both objects on the main/presenter screen for participants to refer to throughout the task.

Oral Instructions (read aloud to participants)

"For this task, you will be asked to come up with as many original and creative uses for a pencil and a fork as you can. The goal is to think of creative ideas, ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different. Your ideas do not have to be practical or realistic; they can be silly or strange, as long as they are creative uses rather than ordinary ones. You can enter as many ideas as you like, but time will be limited, so please be aware that creative quality is more important than quantity. It is better to have a few really good ideas than many uncreative ones."

Procedure

Start with the pencil. Start the timer (3 minutes) on the main screen. Participants type their ideas into the available text boxes in the Google Form. Remind them they may use all boxes or only some. When the timer ends, check that participants have finished this section and instruct them to proceed to the next page.

The second object is the fork. Start a new 3-minute timer. Participants complete the same task for the fork.

Closing the Session

Once both objects are completed, address participants:

"Thank you for your participation. Your answers are very important to us. All responses will remain completely anonymous. If you would like to receive the results of our study, please share your email address with us after the session and we will send you the findings once they are available."

You'll have 10 empty spaces to place your answers in. You dont need to fill them all. If you need extra spaces make sure to place you answers in one box separating them using commas.

Instructions and Task - AI Group

The AI group session must be conducted after the Non-AI group session has been completed and the fluency scores (number of ideas generated per person) have been calculated. This is necessary to calibrate the number of ideas the AI group will be asked to generate.

Pre-Session Calculation

Calculate the mean fluency score from the Non-AI group (i.e., the average number of ideas generated per object across all participants). This number will be used as the target number of ideas for the AI group to generate using ChatGPT. For example, if the mean fluency score was 4 ideas per object, instruct AI-group participants to generate exactly 4 ideas per object.

Participant and Room Setup

Ensure the AI group consists of entirely different participants from the Non-AI group - no overlap between groups. Each computer should have two tabs open: the Google Form survey and a ChatGPT tab (no account signed in, browsing history cleared).

Instruct participants not to sign into ChatGPT and assign a researcher to monitor compliance. The ChatGPT tab should be in a fresh, unauthenticated session.

Procedure

As with the Non-AI group, participants first complete the consent form and demographic information. Then move to the main task. The same oral instructions as in the Non-AI condition are given, but participants are additionally told to use ChatGPT with a provided prompt to generate their ideas. The prompt should request the exact target number of ideas (based on the Non-AI mean fluency score) for each object.

The same timed structure applies: 3 minutes per object, starting with the pencil and then the fork. Display images of both objects on the main screen as before.

Close the session with the same debrief statement used in the Non-AI condition.

Appendix C: Consent Form

Online Consent Form for Participation in Research Study

Study Title: The Impact of Generative AI on Creative Performance: Experimental Evidence from AI-Assisted Ideation

Principal Investigator: Marzhan Kabzhassarova, Dias Nurmukhambetov, Ayaulym Ayazhan

Institution: Nazarbayev University

Contact Information:

ayaulym.ayazhan@nu.edu.kz

Study Supervisor: Mariza Tsakalerou, mariza.tsakalerou@nu.edu.kz

1. Purpose of the Study

You are invited to participate in a research study examining the impact of AI on human creativity. This study will involve completing a creative ideation task, where you will generate original and creative uses for everyday objects (fork and pencil.). The aim is to assess how AI assistance influences your creative thinking.

2. Study Procedures

- You will complete a creative ideation task using a list of objects provided to you.
- For each object, you will generate as many creative uses as you can within a 3-minute time limit.
- The task will either be performed with AI assistance or without it, depending on your group.
- You will be asked to generate responses based on prompts such as: “Come up with as many creative uses for a Fork as you can.”

Task Time: 3 minutes for each object.

3. Participation and Voluntary Nature

Your participation is completely voluntary, and you may withdraw from the study at any time without penalty. You can also skip any task or question without consequence. If you choose to withdraw, any data collected up to that point will be deleted from the study.

4. Risks and Benefits

- **Risks:** There are no significant risks associated with this study. You may experience some mental fatigue or frustration during the ideation tasks, but this is minimal.
- **Benefits:** Your participation will contribute to understanding how AI can impact human creativity, which could influence the use of AI in creative industries.

5. Confidentiality

Your responses will remain confidential. No personally identifiable information will be linked to your data, and all responses will be anonymized. The data will be stored securely, and only the research team will have access to it.

6. Compensation

-

7. IREC Guidelines

This study follows the ethical guidelines set by the Institutional Review Ethics Committee (IREC). Your rights as a participant are protected, and if you have any concerns, you can contact the IREC/ Office of Research Compliance and Integrity at researchcompliance@nu.edu.kz

If you have any questions about the study or your participation, please contact Ayaulym Ayazhan at ayaulym.ayazhan@nu.edu.kz.

Thank you for participating in our research!

Appendix D: Code for GloVe

D.1 Code parts

```

df["cleaned_response"] = cleaned_responses
df["all_tokens"] = all_tokens
df["known_tokens"] = known_tokens
df["known_token_count"] = known_token_counts
df["glove_distance"] = distances
df["status"] = statuses

output_all = "/Users/diasnurmkhambetov/Work-dev/capstone/all_responses_glove_scored.xlsx"
df.to_excel(output_all, index=False)

print("\nSaved response-level file:")
print(output_all)

print("\nValid scored responses:", df["glove_distance"].notna().sum())
print("Invalid skipped:", (df["status"] == "invalid_skipped").sum())
print("Trivial skipped:", (df["status"] == "trivial_skipped").sum())
print("Response missing:", (df["status"] == "response_missing").sum())

print("\nPreview:")
print(df[[
    "participant_id",
    "prompt",
    "response",
    "Response valid",
    "known_tokens",
    "known_token_count",
    "glove_distance",
    "status"
]].head(20))

participant_avg = (
    df[df["status"] == "ok"]
    .groupby(["participant_id", "prompt"])["glove_distance"]
    .mean()
    .reset_index()
)

output_avg = "/Users/diasnurmkhambetov/Work-dev/capstone/participant_average_glove_scores.xlsx"
participant_avg.to_excel(output_avg, index=False)

print("\nSaved participant-average file:")
print(output_avg)

print("\nParticipant averages preview:")

```

```

1 import pandas as pd
2 import numpy as np
3 import re
4 from gensim.models import KeyedVectors
5 from numpy.linalg import norm
6
7 glove_path = "/Users/diasnurmkhambetov/Work-dev/capstone/dolma_300_2024_1.2M.100_combined.txt"
8
9 model = KeyedVectors.load_word2vec_format(
10     glove_path,
11     binary=False,
12     no_header=True,
13     limit=800_000
14 )
15
16 print("Loaded", len(model), "words")
17
18 file_path = "/Users/diasnurmkhambetov/Work-dev/capstone/prompt_response_cleaned_with_translation.xlsx"
19 df = pd.read_excel(file_path)
20
21 df.columns = df.columns.str.strip()
22
23 print("\nColumns found:")
24 print(df.columns.tolist())
25
26 required_cols = ["participant_id", "prompt", "response", "Response valid"]
27 for col in required_cols:
28     if col not in df.columns:
29         raise ValueError(f"Missing required column: {col}")
30
31 stop_words = {
32     "a", "an", "the", "and", "or", "but", "if", "while", "with", "to", "of", "in", "on", "for",
33     "at", "by", "from", "up", "down", "out", "over", "under", "again", "then", "once",
34     "is", "are", "was", "were", "be", "been", "being", "do", "does", "did", "have", "has",
35     "had", "it", "its", "as", "into", "about", "after", "before", "between", "this", "that",
36     "these", "those", "use", "using"
37 }
38
39 def clean_text(text):
40     if pd.isna(text):
41         return ""
42     text = str(text).lower()
43     text = re.sub(r"[^a-zA-Z\s]", " ", text)
44     text = re.sub(r"\s+", " ", text).strip()

```

```

def clean_text(text):
    if pd.isna(text):
        return ""
    text = str(text).lower()
    text = re.sub(r"[^a-zA-Z\s]", " ", text)
    text = re.sub(r"\s+", " ", text).strip()
    return text

def tokenize(text):
    return [t for t in clean_text(text).split() if t not in stop_words]

def embedding_info(text):
    cleaned = clean_text(text)
    tokens = tokenize(text)
    known = [w for w in tokens if w in model]
    vectors = [model[w] for w in known]

    if len(vectors) == 0:
        return cleaned, tokens, known, None

    return cleaned, tokens, known, np.mean(vectors, axis=0)

def cosine_similarity(a, b):
    denom = norm(a) * norm(b)
    if denom == 0:
        return None
    return float(np.dot(a, b) / denom)

def semantic_distance(a, b):
    sim = cosine_similarity(a, b)
    if sim is None:
        return None
    return 1 - sim

distances = []
statuses = []
cleaned_responses = []
all_tokens = []
known_tokens = []
known_token_counts = []

for _, row in df.iterrows():
    valid_flag = str(row["Response valid"]).strip().lower()

```

```

    if valid_flag != "yes":
        distances.append(None)
        statuses.append("invalid_skipped")
        cleaned_responses.append(None)
        all_tokens.append(None)
        known_tokens.append(None)
        known_token_counts.append(None)
        continue

    prompt = row["prompt"]
    response = row["response"]

    if clean_text(response) in ["yes", "no"]:
        distances.append(None)
        statuses.append("trivial_skipped")
        cleaned_responses.append(clean_text(response))
        all_tokens.append(", ".join(tokenize(response)))
        known_tokens.append("")
        known_token_counts.append(0)
        continue

    _, prompt_tokens, prompt_known, prompt_vec = embedding_info(prompt)
    cleaned_resp, resp_tokens, resp_known, resp_vec = embedding_info(response)

    cleaned_responses.append(cleaned_resp)
    all_tokens.append(", ".join(resp_tokens))
    known_tokens.append(", ".join(resp_known))
    known_token_counts.append(len(resp_known))

    if prompt_vec is None:
        distances.append(None)
        statuses.append("prompt_missing")
        continue

    if resp_vec is None:
        distances.append(None)
        statuses.append("response_missing")
        continue

    d = semantic_distance(prompt_vec, resp_vec)
    distances.append(d)
    statuses.append("ok")

df["cleaned_response"] = cleaned_responses
df["all_tokens"] = all_tokens
df["known_tokens"] = known_tokens

```

D.2 Code for t-test

```
import pandas as pd
from scipy.stats import ttest_ind

ai = pd.read_excel("AI_Participant_Response_RESULTS.xlsx")
human = pd.read_excel("NO_AI_GROUP_ALL_Results.xlsx")

ai.columns = ai.columns.str.strip()
human.columns = human.columns.str.strip()

ai = ai.rename(columns={"glove_avg": "glove", "ocsai_avg": "ocsai"})
human = human.rename(columns={"glove_distance": "glove", "ocsai_avg": "ocsai"})

def run_ttest(metric, prompt):
    (parameter) metric: Any
    ai_vals = ai[ai["prompt"] == prompt][metric].dropna()
    human_vals = human[human["prompt"] == prompt][metric].dropna()

    t_stat, p_val = ttest_ind(ai_vals, human_vals, equal_var=False)

    print(f"{metric} | {prompt}")
    print("AI mean:", round(ai_vals.mean(), 4))
    print("Human mean:", round(human_vals.mean(), 4))
    print("t =", round(t_stat, 3))
    print("p =", round(p_val, 4))
    print("-----")

run_ttest("glove", "fork")
run_ttest("glove", "pencil")
run_ttest("ocsai", "fork")
run_ttest("ocsai", "pencil")
```

D.3 Code for visuals

```
sns.set_theme(style="whitegrid", font_scale=1.1)

plt.figure(figsize=(9, 6))
sns.boxplot(
    data=combined,
    x="group",
    y="glove",
    hue="prompt"
)
plt.title("GloVe Semantic Distance by Group and Prompt")
plt.ylabel("Semantic Distance")
plt.xlabel("Group")
plt.tight_layout()
plt.savefig("/Users/diasnurmukhambetov/Work-dev/capstone/glove_boxplot.png", dpi=300)
plt.show()

plt.figure(figsize=(9, 6))
sns.boxplot(
    data=combined,
    x="group",
    y="ocsai",
    hue="prompt"
)
plt.title("OCSAI Originality Scores by Group and Prompt")
plt.ylabel("Originality Score")
plt.xlabel("Group")
plt.tight_layout()
plt.savefig("/Users/diasnurmukhambetov/Work-dev/capstone/ocsai_boxplot.png", dpi=300)
plt.show()

combined.to_excel("/Users/diasnurmukhambetov/Work-dev/capstone/combined_ai_human_results.xlsx", index=False)

print("\nSaved:")
print("- combined_ai_human_results.xlsx")
print("- glove_boxplot.png")
print("- ocsai_boxplot.png")
print("- glove_barplot.png")
print("- ocsai_barplot.png")
print("- glove_vs_ocsai_scatter.png")
```

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import ttest_ind

ai = pd.read_excel("/Users/diasnurmukhambetov/Work-dev/capstone/AI_Participant_Response_RESULTS.xlsx")
human = pd.read_excel("/Users/diasnurmukhambetov/Work-dev/capstone/NO_AI_GROUP_ALL_Results.xlsx")

ai.columns = ai.columns.str.strip()
human.columns = human.columns.str.strip()

ai = ai.rename(columns={
    "glove_avg": "glove",
    "ocsai_avg": "ocsai"
})

human = human.rename(columns={
    "glove_distance": "glove",
    "ocsai_avg": "ocsai"
})

ai["group"] = "AI"
human["group"] = "Human"

combined = pd.concat([ai, human], ignore_index=True)

print("Combined columns:", combined.columns.tolist())
print("\nPreview:")
print(combined.head())

def run_ttest(metric, prompt):
    ai_vals = combined[(combined["group"] == "AI") & (combined["prompt"] == prompt)][metric].dropna()
    human_vals = combined[(combined["group"] == "Human") & (combined["prompt"] == prompt)][metric].dropna()

    t_stat, p_val = ttest_ind(ai_vals, human_vals, equal_var=False)

    print(f"\n{metric} | {prompt}")
    print("AI mean:", round(ai_vals.mean(), 4))
    print("Human mean:", round(human_vals.mean(), 4))
    print("t =", round(t_stat, 3))
    print("p =", round(p_val, 4))

run_ttest("glove", "fork")
run_ttest("glove", "pencil")
run_ttest("ocsai", "fork")
run_ttest("ocsai", "pencil")
```

Appendix E: Code Output for Fluency

```
Correlation matrix:
      English_num  AI_freq_num  Gender_num  Ideas
English_num    1.000000    -0.101409   -0.511364    0.401188
AI_freq_num   -0.101409    1.000000    0.084324   -0.056078
Gender_num    -0.511364    0.084324    1.000000   -0.209010
Ideas         0.401188   -0.056078   -0.209010    1.000000

Spearman: English vs Ideas = SignificanceResult(statistic=0.3912383427776727, pvalue=0.058690831860060334)
Spearman: AI frequency vs Ideas = SignificanceResult(statistic=-0.12023037208732472, pvalue=0.5757527611602362)

Mean ideas by English:
Your level of English
A1    3.000000
B2    3.800000
C1    6.133333
C2    7.000000
Name: Ideas, dtype: float64

Mean ideas by AI usage:
How often do you use AI tools (e.g., ChatGPT, Claude, Perplexity, Gemini)?
Daily          5.500000
Once in a Month 5.000000
Weekly         6.666667
Name: Ideas, dtype: float64

Mean ideas by Gender:
Gender
Female    4.714286
Male      6.000000
Name: Ideas, dtype: float64

Mean ideas by Age:
Age
18-24     5.571429
25-34     6.000000
Name: Ideas, dtype: float64
```

Notes: