UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

EFFECTS OF USING GENERATIVE ARTIFICIAL INTELLIGENCE COLLABORATION DURING IDEA GENERATION AND EVALUATION

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

JOSIE RUTH KOUBEK

Norman, Oklahoma

2026

EFFECTS OF USING GENERATIVE ARTIFICIAL INTELLIGENCE COLLABORATION DURING IDEA GENERATION AND EVALUATION

A THESIS APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Shane Connelly, Chair
Dr. Lori Anderson Snyder
Dr. Eric Day

© Copyright by JOSIE RUTH KOUBEK 2026

All Rights Reserved.

Abstract

Generative artificial intelligence (GenAI) is increasingly integrated into creative work, yet its effects across the creative process remain unclear. Creativity involves generating ideas that are both novel and useful (Runco & Jaeger, 2012), and emerging research suggests that GenAI may augment aspects of creative cognition, particularly idea generation (Koivisto & Grassini, 2023; Medeiros et al., 2025). The present study examined how GenAI use influences idea generation and idea evaluation in a sample of 407 undergraduate participants. Participants completed a restaurant conceptualization task under varying GenAI conditions; due to heterogeneous engagement, participants were reclassified into behaviorally derived groups (no AI, AI-generated, augmented, and prompt engineering). Results indicated that GenAI use significantly improved idea generation outcomes, with AI-generated use increasing fluency while prompt engineering and augmented use were associated with higher idea quality. Prompt engineering and augmented use also predicted greater originality. Similarly, AI use significantly enhances idea evaluation quality, and originality. These findings suggest that GenAI enhances early-stage creative processes, but its effectiveness likely depends on how individuals engage with the tool, with structured, metacognitive interaction (i.e., prompt engineering) yielding the strongest benefits.

Table of Contents

Introduction **1**
Defining Creativity **2**
Creative Processes **3**
Idea Generation **4**
Idea Evaluation **5**
Creative Performance and Generative Artificial Intelligence **6**
GenAI and Idea Generation **7**
GenAI and Idea Evaluation **8**
GenAI Prompt Engineering **9**
Generative Artificial Intelligence and Self-efficacy **10**
Method **12**
Study Design and Participants **12**
Procedures **12**
Manipulations **13**
Artificial Intelligence Chatbot 13
Prompt-engineering 14
Experimental Task **15**
Criterion Measures **15**
Creative Process and Product Evaluations 15
Chatbot Satisfaction 17
Task-Specific Self-efficacy 17
Artificial Intelligence Self-Efficacy 17
Covariate Measurement **18**
Self-esteem 18
Intelligence 18
Trait Fluency **18**
Trait Affect 19
Results **19**
Manipulation Checks **20**
Creating New GenAI Collaboration Groups **21**
Data Analysis Approach **22**
Generation Outcomes (RQ1, RQ2, RQ5) **22**

Fluency 22
Quality 22
Originality 23
Evaluation Outcomes (RQ3, RQ4, RQ6) 23
Evaluation Quality 23
Evaluation Originality 23
Implementation Outcomes (RQ2, RQ4, RQ6) 24
Self-Efficacy and Attitudinal Outcomes (RQ7) 24
Discussion 24
Theoretical Contribution 28
Practical Implications 31
Limitations and Future Directions 31
Conclusion 32
References 34
Table 1 39
Table 2 40
Table 3 41
Table 4 42
Table 5 43
Appendix A 44
Appendix B 46
Rating Scales and Benchmarks for Idea Generation 46

Introduction

Creativity has long been celebrated as a fundamental driver of progress, influencing not only individual development but also societal advancement. From the arts to technological innovations, creativity plays a key role in shaping history and improving human life. Studies have demonstrated that creative capacity enhances job satisfaction and workplace productivity while also fostering crucial skills such as communication, collaboration, and problem-solving (Forgeard & Eichner, 2014; Kampylis & Valtanen, 2010). Creativity is valued both intrinsically, as a personal asset that individuals strive to cultivate, and extrinsically, as it fuels economic growth and innovation.

The workplace has always required a significant investment of time and resources to nurture creativity (Amabile et al. 1996; Unsworth & Clegg 2010). However, the returns on this investment are vast. Creativity has been linked to better business performance, greater entrepreneurship, and historical breakthroughs in technology and culture (Amabile, 1988; Simonton, 2000). For this reason, understanding how to foster and optimize creativity remains a central concern across industries. As creative disciplines wrestle with the uses and integration of generative artificial intelligence (GenAI) into creative activities, the potential for human-AI collaboration opens possibilities for further advancement within many spaces. GenAI, defined as systems that leverage advanced machine learning models (e.g., large language models) to generate human-like outputs from learned data patterns (Baidoo-Anu & Owusu Ansah, 2023), can process vast amounts of data and produce diverse possibilities for displaying, combining, and interpreting information, thereby augmenting human creativity. However, the implications of this integration, such as whether and how artificial intelligence helps or harms creative processes and outcomes, are still being explored. This study compares humans working alone and humans working in collaboration with GenAI on the quality of two key creative processes (idea generation and idea evaluation) and the quality and novelty of output for a creative task. Additionally, it seeks to investigate how using GenAI influences feelings of satisfaction and self-efficacy concerning creative tasks.

Defining Creativity

Creativity is a complex and multifaceted concept studied across a variety of disciplines. While definitions vary, most scholars agree that creativity involves producing ideas, products, or solutions that are both novel and useful (Runco & Jaeger, 2012; Kaufman, 2016). Within the context of problem-solving, creativity involves formulating original, high-quality solutions to complex, ill-defined problems (Mumford & Gustafson, 1988). This perspective emphasizes creativity as a set of underlying cognitive and problem-solving processes, suggesting that these skills can be developed and enhanced over time. Novelty refers to the originality or uniqueness of an idea, while utility or value encompasses its relevance, usefulness, or appropriateness within a given context. Creativity is not a fixed trait but rather emerges from a dynamic interaction between an individual's abilities, skills, and characteristics (e.g., cognitive abilities, motivation, and personality) and environmental or situational factors (Amabile, 1983), with creative outcomes requiring both originality and effectiveness, that is, ideas must be not only novel but also appropriate or useful (Runco & Jaeger, 2012). Accordingly, creative performance is shaped by not only personal attributes but also by the specific context in which the problem is presented, the resources available, and the social and collaborative dynamics in the work environment. As a result, creativity has been defined and studied by some scholars as a set of processes that lead to creative outcomes.

Creative Processes

Scholars have examined creativity from different perspectives, focusing on both the cognitive processes that lead to creative production and the tangible products that result from creative efforts. This distinction between process and product is critical, as it highlights that creativity is not just about having an idea, but about navigating the steps necessary to develop, refine, and implement that idea successfully (Amabile,1983; Mumford et al., 1991). The creative process refers to the steps or stages individuals go through when producing novel and adaptive solutions to complex problems. Lubart (2001) defines the creative process as a "sequence of thoughts and actions" that leads to the generation of new ideas. Mumford et al. (1991) developed a widely recognized model outlining the multistep processes underlying creative output, identifying eight key processes: problem construction, information encoding, category search, specification, reorganization, idea evaluation, idea implementation, and monitoring. Each of these processes plays an essential part in creative problem solving.

For instance, problem construction is the initial phase, where individuals define and frame the problem they are attempting to solve. How a person frames or constructs a problem influences the types of solutions that will emerge (Mumford et al., 1991). Subsequent creative processes, such as information encoding and category search, involve gathering, accessing and organizing relevant knowledge and exploring different frameworks or categories that might lead to novel solutions. The reorganization process, which involves combining and rearranging existing knowledge in new ways, is often where the most innovative ideas take shape. The creative process model emphasizes that creativity is not a linear or straightforward path; rather, it involves complex and dynamic interactions between different cognitive processes (Mumford et al., 1991). In this study, we focus particularly on two creative processes, idea generation and idea evaluation, investigating whether and how the use of GenAI influences the number and quality of ideas generated, the quality of idea evaluation for those ideas, and the creative performance or final plan for implementing the ideas, relative to human performance in these domains without the use of GenAI.

Idea Generation

The idea generation process involves appraising the constraints and possibilities of a task, generating new concepts, and extending these ideas through exploration and refinement (Mumford et al., 1991). During this phase, individuals use divergent thinking and other cognitive processes to produce multiple possible solutions to a problem. Divergent thinking emphasizes generating a broad range of ideas and exploring many possible solutions from different perspectives, promoting openness, flexibility, and originality in problem solving (Guilford, 1967; Runco, 1991). It often involves techniques like brainstorming, free association, and other creative methods that foster idea exploration without immediate judgment or constraints, enabling unexpected and innovative solutions to emerge (Benedek et al., 2023). Creative ideas emerge from goal-directed memory processes that build on existing memory elements, enabling individuals to transcend the known through imagination by forging new connections between previously unrelated concepts (Benedek et al., 2023). Divergent thinking is the process of making as many of these connections as possible, thus referring to generation. In addition to generating many ideas, generating novel ideas requires balancing originality with usefulness, ensuring that the ideas not only stand out but also serve a purpose within the given context (Runco & Jaeger, 2012). Many techniques have been suggested for improving idea generation, such as analytical strategies, imagination-based strategies, habit-breaking strategies, and relationship-seeking (analogical) strategies (Smith, 1998). Research has shown that generating ideas is a critical part of innovation, as it provides the foundation for further development (Girotra et al., 2010). Although idea generation is typically associated with divergent thinking, it may also involve elements of convergent thinking, particularly when individuals refine or select among emerging ideas (Puccio et al., 2011).

Idea Evaluation

A second key creative process, idea evaluation, is equally important but often less understood than idea generation. Idea evaluation involves critical engagement with the ideas generated, assessing idea feasibility, effectiveness, and relevance to the task at hand (Licuanan et al. 2007). Instead, successful problem-solving or innovation depends on identifying one or a few effective ideas that outperform previous approaches (e.g., Rietzschel et al., 2006). Evaluating ideas requires convergent thinking, as individuals must compare and contrast different possibilities, identify the most promising solutions, and refine them as necessary. Research suggests that individuals with strong evaluation skills are better equipped to filter out less viable ideas and focus on those with the greatest potential (Basadur et al., 2000). This idea highlights the importance of evaluating ideas in context. Mumford et al. (2002) proposed a model of idea evaluation, in which they emphasized the importance of prediction during idea evaluation. Although prediction is necessary, individuals fall susceptible to two common errors during evaluation. First, individuals often overlook highly creative ideas by deeming them unoriginal, impractical, or both (Simonton, 2018). Second, they give creative ideas a negative assessment based on factors such as general preference, predicted success, or the evaluator's personal biases. Highly original ideas often involve unfamiliar or abstract concepts, making assessing their feasibility or relevance difficult. As a result, individuals may struggle to apply existing knowledge to evaluate new ideas accurately (Mumford et al., 2002). Nevertheless, the ability to critically assess ideas is essential for creativity, as it allows individuals to refine and improve upon initial concepts.

Creative Performance and Generative Artificial Intelligence

The increasing use of artificial intelligence in creative domains has raised important questions about how AI influences human creativity. GenAI offers new possibilities for idea generation and evaluation, as it can process large datasets, identify patterns, and suggest novel combinations of ideas or data that might not otherwise be considered. Although a growing body of research demonstrates that GenAI can influence creative performance, relatively little empirical work has examined the different ways individuals interact with these tools throughout the creative process. Researchers have compared human and GenAI capabilities across creative industries (e.g., music and visual arts), suggesting that GenAI is most successful when paired with human creators and decision-makers rather than when attempting to use inputs to produce "creative work" itself (Anantrasirichai & Bull, 2021). Additional research examining the comparison between AI and human production found that across all three studies, which employed different measures, designs, and evaluation targets, human evaluators perceive a production process as less effortful when performed by GenAI compared to humans (Magni et al., 2023). This perception leads them to judge the resulting product's creativity as lower, thereby extending the effort heuristic to the realm of creative production. Given the debate as to whether GenAI is inherently creative and capable of novel and quality production, there is support and calls for increased exploration of human-AI augmentation within the creativity space (Cropley et al., 2022; Medeiros et al., 2023). Reviews and research suggest AI be used as a "creative support tool" rather than for creative production on its own (Rafner et al., 2023; Haase & Hanel, 2023). Additionally, researchers reviewing artificial intelligence's potential within creative industries suggest GenAI is particularly limited in autonomous creativity, as current GenAI tools are typically developed and designed to augment human creativity rather than replace it (Anantrasirichai & Bull, 2021). Although this work demonstrates that GenAI can influence creative performance, most studies conceptualize AI use as a binary condition (AI versus no AI). Far less is known about how different patterns of AI interaction influence creativity. Emerging evidence suggests that the benefits of GenAI depend not only on whether AI is used, but also on how individuals engage with it, including their ability to strategically direct, evaluate, and refine AI-generated outputs (Doshi & Hauser, 2024; Sun et al., 2025).

GenAI and Idea Generation

AI tools can augment the idea-generation process by offering an array of possibilities based on the tools' underlying data sets. With GenAI's ability to analyze massive datasets and draw various connections, AI can provide users with alternative perspectives, helping humans to generate ideas and explore new directions. In terms of technological processes similar to human divergent thinking, this would mean that a GenAI tool develops a large number of possible solutions/designs and therefore can support creative individuals and teams in idea generation. Given the vast scale of data used to train generative AI systems, these tools can draw from a substantially larger space of potential ideas and solutions than what is typically accessible through human memory alone (Brown et al., 2020; Anderson & Schooler, 1991). One study has compared the performance of AI chatbots and human participants in a typical Guilford-style divergent thinking task (Koivisto & Grassini, 2023). Researchers found that on average, the AI chatbots outperformed the human participants in terms of the quantity of responses, however, they found that the most creative ideas generated by humans were comparable or exceeded the most creative ideas of the AI chatbots. This suggests that the human act of crafting high-quality, deeply original ideas still surpasses AI's capabilities. Additionally, at least one study has examined human performance with experimentally manipulated GenAI assistance on divergent thinking tasks (Medeiros et al., 2025). Participants were primed with examples of low, medium, or high creativity from either GenAI or humans, and then completed divergent association tasks.

Researchers found that experimentally manipulated "low creativity" help from ChatGPT lowered human performance on a divergent thinking task, while experimentally manipulated "high creativity" help from AI was no better than any kind of help from a human. Results from these empirical studies suggest augmented creativity may not be straightforward and that there is a need to know more about when aid from GenAI is helpful within the creative process. Similarly, Doshi and Hauser (2024) found that generative AI improved the creativity of individual writing tasks, particularly for lower-performing participants. However, AI-assisted outputs became more similar across participants, suggesting that although AI may enhance average creative performance, it may simultaneously reduce the diversity of ideas generated.

RQ1: How does using generative GenAI during idea generation affect the fluency, quality, and novelty of ideas compared to generating ideas without AI assistance?

RQ2: How does using GenAI during idea generation impact the quality, originality, and elegance of the resulting output for the creative performance task?

GenAI and Idea Evaluation

Although a few studies have been conducted to compare the divergent thinking abilities of humans and GenAI, little work has investigated other phases of the creative process using AI, such as using AI in the evaluative or refinement process. Idea evaluation requires both divergent and convergent thinking capabilities (Runco, 2007). Comparison of human versus GenAI convergent thinking abilities has received less attention relative to divergent thinking, despite acknowledgment of its importance within creative performance (Brophy, 1998). Cropley (2006) outlined convergent thinking as oriented toward deriving the best answer to a clearly defined problem. This process requires the application of knowledge and accuracy in decision-making.

During the evaluation phase, GenAI can assist by quickly processing and analyzing ideas and identifying potential flaws or areas for improvement. Given the broad data that underlies GenAI, these tools can potentially evaluate ideas using a breadth of knowledge that may not thoroughly considered or considered at all. However, since evaluation often involves subjective judgment and contextual understanding, AI's role in this phase may be limited to supporting human decision-making rather than replacing it. Additionally, creative problem-solving tasks are often ill-defined and there may be more than one good solution (Mumford & Gustafson, 1988). Further research is needed to understand AI-human collaboration in idea evaluation.

RQ3: How does using GenAI during idea evaluation affect the quality and novelty of ideas compared to evaluating ideas without GenAI assistance?

RQ4: How does using GenAI during idea evaluation impact the quality, originality, and elegance of the resulting output for the creative performance task?

GenAI Prompt Engineering

Given the ever-changing landscape of GenAI, suggestions continually emerge about how to better increase the applicability and quality of GenAI output. One proposed solution for increasing the likelihood of utilizing GenAI output by academics and practitioners alike is prompt engineering. Prompt engineering is defined as “the art of crafting effective prompts that guide ChatGPT to generate desired responses; this process involves the role of designing, refining, and optimizing input prompts to communicate the user’s intent to (large) language models effectively” (Ekin, 2023, pg. 4). Emerging research suggests that effective use of GenAI depends not only on access to the technology but also on users' ability to strategically guide, evaluate, and iteratively refine AI outputs. For example, Sun et al. (2025) found that employees possessing stronger metacognitive strategies benefited more from LLM use than those who relied more passively on AI-generated responses. Similarly, conceptual work argues that AI functions most effectively as a collaborative creative partner when users actively direct and evaluate its contributions (Ivcevic & Grandinetti, 2024).

Importantly, although prior research has largely compared the presence versus absence of GenAI assistance (for creative tasks, or just in general?), such an approach may overlook meaningful variability in how individuals actually engage with these tools. In practice, users differ substantially in how they interact with GenAI, ranging from passive reliance on AI-generated outputs to active, iterative prompt construction. These differences in engagement may fundamentally shape creative outcomes. Accordingly, the present study adopts a process-oriented perspective, examining not only whether GenAI is used but how it is used in two key creative processes.

RQ5: How does using GenAI with prompt engineering guidance during a) idea generation and b idea evaluation affect the fluency, quality, and novelty of ideas compared to not using AI and using AI without explicit prompt engineering guidance?

RQ6: How does using GenAI with prompt engineering guidance during idea evaluation affect the creativity of the resulting product (quality, originality, and elegance) compared to not using AI and using AI without explicit prompt engineering guidance?

Generative Artificial Intelligence and Self-efficacy

In addition to concerns about creative performance and creative outcomes, there are some questions as to how artificial intelligence makes individuals feel about their performance and abilities to perform similar tasks in the future. This may be related to feelings of efficacy surrounding the task and/or to artificial intelligence use. Self-efficacy is defined as an individual's belief in their ability to execute necessary behaviors related to a specific task (Bandura, 1997). The explanatory construct, derived from self-efficacy theory, has been empirically linked to psychological processes such as achievement striving, perception of accomplishments, career choice, academic pursuits, motivation, and refractory behaviors (Bandura, 1977; Lent et al., 1994). An effective way to apply self-efficacy theory to employee creativity is through the concept of creative self-efficacy, which reflects an individual's belief in their ability to generate creative outcomes (Tierney & Farmer, 2002; Shalley et al., 2004). If confidence in one's creative abilities is essential for improving or promoting creative performance, as suggested by Bandura (1997), and Bandura and Locke (2003), then fostering creative self-efficacy among employees is a key element for organizations aiming to drive innovation. Additional research suggests that positive changes in creative self-efficacy correspond with improvement in creative performance (Tierney & Farmer, 2011).

GenAI has the potential to significantly impact individuals' self-efficacy, particularly in their perceptions of current and future task performance. By automating complex tasks, providing recommendations, and enhancing productivity, GenAI can bolster users' confidence in their ability to perform and excel in certain activities, leading to increased task-specific self-efficacy. For example, when individuals interact with GenAI tools that support problem-solving or offer creative suggestions, they may perceive themselves as more competent, creative, or effective in completing similar tasks in the future. Conversely, reliance on AI can also present challenges that may erode self-efficacy if users feel overshadowed by the capabilities of the technology or perceive themselves as unable to match its speed, accuracy, or efficiency without GenAI assistance.

RQ7: How does variation in GenAI use during creative tasks influence individuals' task-specific self-efficacy, controlling for GenAI self-efficacy?

Method

Study Design and Participants

This study utilizes a 2X3 within-and-between-subjects design to evaluate the effects of two independent variables, creative process (2 within-person levels - idea generation and idea evaluation) and use of GenAI (3 between-person levels - use of GenAI with the idea generation and idea evaluation, use of GenAI and prompt engineering with idea generation and idea evaluation, and no use of GenAI) on dependent variables (creative performance, self-efficacy, and satisfaction with the tool). Given the design, there are six total experimental conditions. A total of 407 undergraduate students from a large, southwestern university were recruited to participate in this study, in exchange for course credit.

Procedures

Participants were brought into the lab and were randomly assigned to one of three conditions: idea generation with GenAI, idea generation with GenAI and explicit prompt-engineering instruction, or idea generation without GenAI. They were given a generative creative task scenario requesting a brief written plan for a new restaurant. Prior to completing this creative task, participants engaged in an idea generation process followed by an idea evaluation process. Participants in all three conditions spent approximately ten minutes on generating a list of restaurant concepts. The task was self-paced. Participants in the prompt engineering condition were provided with 6 best-practices statements about prompt engineering when using GenAI and were instructed to follow these practices.

Following idea generation, all participants engaged in idea evaluation, which lasted approximately ten minutes. During this phase, participants were asked to reflect on the generated ideas, with or without the help of GenAI based on the condition assignment. Participants were asked to select the best idea to move forward with for developing the written plan of the restaurant concept, Participants in the prompt engineering condition GenAI received the best-practice prompt engineering guide and were instructed to follow these practices during the idea evaluation process.

After completing the idea evaluation phase, all participants were required to write a detailed plan describing their selected restaurant concept. This written plan served as the creative product outcome.

Following the experimental task, participants completed a series of scales measuring various constructs, including self-efficacy beliefs, AI-related self-efficacy, and relevant covariates (e.g., self-esteem, intelligence, divergent thinking, trait affect, and GenAI use), as well as demographic information. Finally, participants responded to several multiple-choice questions assessing task engagement and task satisfaction.

Manipulations

Artificial Intelligence Chatbot

To simulate typical GenAI use, participants interacted with a custom AI chatbot embedded within Qualtrics that leveraged OpenAI's GPT-3.5 Turbo (gpt-3.5-turbo-1106) large language model. The interface was designed to replicate the conversational, general-purpose text experience of widely used systems such as ChatGPT while providing a standardized and controlled research environment. Participants accessed the chatbot through a standard web browser, where the interface presented a welcome message and a clean, intuitive chat environment featuring a countdown timer, message history panel, and text input area with a send button.

The chatbot application was developed using a Python Flask backend to manage server-side logic, web requests, and communication with the OpenAI model. The frontend was built using HTML, CSS, and JavaScript, and the application was deployed through Heroku to ensure reliable operation throughout data collection. Sampling parameters were configured to encourage creative idea generation while maintaining coherent responses (temperature = 1.0, top-p = 1.0). At the time of data collection, GPT-3.5 Turbo was one of OpenAI's most widely used and high-performing large language models.

To ensure consistent interactions across participants, the system was standardized such that participant inputs did not alter the model's underlying behavior across sessions. All user prompts and AI-generated responses were recorded in real time and securely stored in a Firebase Firestore database, including message content and timestamps. These complete interaction transcripts enabled subsequent analyses of participants' prompting behavior and patterns of engagement with the GenAI system.

Prompt-engineering

To instruct participants in the prompt-engineering condition, they were provided with a document defining prompt engineering and outlining a set of best practices (Ekin, 2023). Prompt engineering was described as the skill of crafting clear and precise instructions to guide AI chatbot responses. The document encouraged participants to be mindful of several key factors when interacting with the chatbot, including user intent, clarity and specificity, vocabulary choice, constraints, and the use of alternating high- and low-level questions. Participants were prompted to reflect on the type of task they were asking the AI to perform and to iteratively refine their prompts to better align with their desired outcomes. Additional guidance emphasized the importance of being specific, breaking complex tasks into smaller, manageable steps, and rephrasing prompts when responses were unclear or insufficient. Participants were also encouraged to continue engaging with the chatbot to refine outputs and were provided with examples to support effective prompt construction. Finally, they were instructed to fully utilize the interaction window to explore and improve response quality (see Appendix A).

Experimental Task

A restaurant conceptualization creative task was used as the creative performance task in this study (Johnson and Connelly, 2014). In this task, participants were asked to imagine and design a new restaurant concept. They were prompted to write a restaurant concept plan considering theme, cuisine, audience, setting, and why they chose this concept. None of the participants had access to or used GenAI when completing this task.

Criterion Measures

Creative Process and Product Evaluations

Consistent with prior research on creativity, multiple dimensions of performance were assessed across task phases using trained human raters and consensus-based benchmark rating scales (Mumford et al., 2012). For the idea generation task, responses were evaluated in terms of fluency, quality, and originality. Fluency was operationalized as the total number of non-redundant ideas generated, whereas quality and originality were rated using benchmark scales. For the idea evaluation task, responses were assessed on quality and originality using the same benchmark-based approach. For the restaurant concept plan task, outcomes were conceptualized as plan quality, plan originality, and plan elegance, reflecting established multidimensional definitions of creativity.

Benchmark rating scales were developed based on established definitions of creativity, with particular emphasis on quality and originality as core criteria, and extended to include elegance for the final plan task. These scales included anchored examples representing varying levels of performance to guide consistent evaluation across responses (see Appendix B for an example benchmark rating scale).

Raters underwent structured training in which they were familiarized with the benchmark criteria and practiced applying the scales to sample responses. During training, raters discussed discrepancies in their evaluations to calibrate their interpretations of quality, originality, and elegance and to ensure a shared understanding of the rating standards. To establish consistency, raters initially met to discuss and reach consensus on ratings for approximately the first 50 responses, continuing this process until adequate interrater reliability was achieved. Following this calibration phase, raters completed evaluations independently, with periodic checks to maintain alignment.

Interrater reliability was assessed using two-way mixed-effects intraclass correlation coefficients with an absolute agreement definition for average measures, ICC(3,k). Results indicated strong agreement across all dimensions. For idea generation, reliability was high for quality (ICC(3,k) = .91, 95% CI [.89, .93]) and originality (ICC(3,k) = .92, 95% CI [.90, .93]). For idea evaluation, reliability was similarly strong for quality (ICC(3,k) = .86, 95% CI [.83, .89]) and originality (ICC(3,k) = .85, 95% CI [.82, .88]). For the restaurant concept plans, reliability was also strong for plan quality (ICC(3,k) = .89, 95% CI [.87, .91]), plan originality (ICC(3,k) = .94, 95% CI [.93, .95]), and plan elegance (ICC(3,k) = .91, 95% CI [.89, .92]).

Chatbot Satisfaction

Satisfaction with the chatbot was assessed following task completion using a five-item measure developed for this study. Participants rated their agreement with statements reflecting perceptions of the chatbot's performance and usefulness (e.g., "I am satisfied with the chatbot's expertise," "Overall, I am satisfied with the chatbot," "The chatbot was helpful in improving my performance on the task"). Responses were recorded on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Item responses were averaged to create a composite score, with higher values indicating greater satisfaction. Internal consistency was high (Cronbach's $\alpha = .93$).

Task-Specific Self-efficacy

Task-specific self-efficacy was measured using an adapted version of the Personal Efficacy Beliefs Scale (Riggs & Knight, 1994). Participants rated their confidence in their ability to perform the restaurant conceptualization task (e.g., "I have confidence in my ability to complete restaurant conceptualization tasks," "I have the skills needed to excel in this task"). The scale consisted of 10 items rated on a 6-point Likert scale ranging from *strongly disagree* to *strongly agree*. Item responses were averaged to form a composite score, with higher values indicating greater task-specific self-efficacy. Internal consistency was modest ($\alpha = .65$).

Artificial Intelligence Self-Efficacy

Artificial intelligence self-efficacy was assessed using a 20-item scale adapted from Wang and Chuang (2023), capturing confidence in using and interacting with AI technologies across four dimensions: assistance, anthropomorphic interaction, comfort with AI, and technological skills. Participants responded on a 7-point Likert scale. Example items include "I find that AI technologies/products are helpful for learning" and "I find it easy to get AI technologies/products to do what I want it to do." Responses were averaged to create a composite score, with higher values indicating greater AI self-efficacy. Internal consistency was high (α = .89).

Covariate Measurement

Self-esteem

Self-esteem was measured using the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1979), a 10-item measure assessing global self-worth and self-acceptance. Participants rated their agreement with statements (e.g., "I feel that I have a number of good qualities," "I feel that I am a person of worth") on a 7-point Likert scale. Item responses were averaged to create a composite score, with higher values indicating greater self-esteem. Internal consistency was high (α = .85).

Intelligence

Given the verbal nature of the creative task, verbal reasoning ability was assessed using the verbal reasoning subscale of the Employee Aptitude Survey (EAS; Ruch & Ruch, 1983). This measure consists of logic-based verbal items requiring participants to identify relationships among words, complete analogies, and evaluate the logical consistency of verbal statements. Items were scored as correct or incorrect and summed to form a total score. Internal consistency for the scale was acceptable (α = .74).

Trait Fluency

Trait fluency was assessed using an adapted version of the Alternative Uses Task (Guilford, 1967). Participants were asked to generate as many alternative uses as possible for a common object within a fixed time limit. Fluency was operationalized as the total number of non-redundant responses generated. Three trained raters independently counted the number of valid uses, and discrepancies were resolved through discussion until consensus was reached.

Although the Alternative Uses Task can be evaluated across multiple dimensions (e.g., originality, flexibility, elaboration), the present study focused specifically on fluency as the primary outcome of interest.

Trait Affect

The trait level effect was assessed with the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988). The scale measures two independent dimensions of affect, positive and negative, by asking individuals to rate their experience with 20 varying emotions (i.e., interested, ashamed, active) on a five-point scale (from very slightly or not at all to extremely). Following PANAS guidelines, instructions for measuring trait affect asked participants to reflect how much they felt each emotion over the past month. Positive emotion ratings were averaged to create a positive affect (PA) score (α = .857) and negative emotion ratings were averaged to create a negative affect (NA) score (α =.863).

Results

Means, standard deviations, and correlations among study variables are presented in Table 1 and 2. Consistent with prior research, quality and originality measures were positively correlated within each task phase (Mumford et al., 1991; Lubart, 2001). Specifically, generation quality and originality were strongly related ($r = .61$, $p < .01$), and similar patterns were observed for idea evaluation, where quality and originality were also strongly correlated ($r = .61$, $p < .01$). For the restaurant concept plan, quality was positively associated with both originality ($r = .32$, $p < .01$) and elegance ($r = .67$, $p < .01$), and originality was also positively related to elegance ($r = .35$, $p < .01$).

Across task phases, performance measures were moderately interrelated, such that higher quality and originality during idea generation were associated with higher performance during idea evaluation and implementation. These correlations were generally small to moderate in magnitude, suggesting that while related, creative performance across phases reflects partially distinct constructs.

Manipulation Checks

Manipulation checks were conducted through qualitative coding of chatbot interaction transcripts to assess whether participants engaged with the AI tool in accordance with their assigned condition. Two primary checks were performed to distinguish between AI-generated and AI-generated with prompt engineering conditions.

First, transcripts were coded for the presence of prompt engineering strategies. Specifically, coders assessed whether participants used techniques introduced in the instructions, such as prompt refinement, specifying constraints (e.g., requesting a certain number of ideas), or iteratively modifying prompts. Each transcript was evaluated for evidence of these strategies, and participants were classified based on whether they demonstrated adherence to one or more of the suggestions provided in the prompt engineering instructions.

Second, transcripts were coded to determine the mode of AI interaction among participants in the AI-generated and AI-generated with prompt engineering conditions. These codes were mutually exclusive, such that participants classified as prompt engineering could not also be classified as augmented or AI-generated, and vice versa. Interactions were categorized as AI-generated when participants relied on the chatbot to independently produce ideas with minimal user input, or as augmented when participants first generated their own ideas and subsequently used the chatbot to refine, expand, or evaluate those ideas. Results from these manipulation checks indicated substantial variability in how participants engaged with the AI tool. Although some participants in the prompt engineering condition demonstrated the intended strategies, many did not consistently apply these techniques. Similarly, across both AI conditions, participants varied in the extent to which they relied on AI-generated versus self-generated content. Together, these findings suggest that participants did not uniformly adhere to the prescribed interaction strategies, resulting in heterogeneous patterns of AI use across conditions.

Creating New GenAI Collaboration Groups

Given that participants in the GenAI conditions did not adhere to instructions, new categories were developed based on observed patterns of AI use. These behaviorally derived categories better captured how participants engaged with the AI tool and allowed for a more accurate examination of how different usage types relate to the proposed outcomes. Participants were categorized into groups based on manual coding of their AI chatbot transcripts, as explained in the manipulation check section. Differences in engagement were most pronounced during the idea generation phase but extended throughout the remainder of the task. Based on observed patterns of use, participants were classified into four independent groups: prompt engineering ($n = 33$, 8.1%), augmented ($n = 103$, 25.3%), AI-generated ($n = 128$, 31.4%), and no AI ($n = 143$, 35.1%). Participants in the prompt engineering group actively constructed and refined prompts to guide the AI's responses. Those in the augmented group generated their own initial ideas and used AI to elaborate or refine those ideas. In contrast, participants in the AI-generated group relied on the AI to produce ideas with minimal original input. For all subsequent regression analyses, AI Use Type was treated as a categorical predictor and dummy-coded. The AI group served as the reference category. Three dummy-coded variables were created to represent the remaining groups (prompt engineering, augmented, and AI-generated), with each coded as 1 when participants belonged to that group and 0 otherwise. This coding scheme allowed each AI use group to be compared directly with the no AI reference group.

Data Analysis Approach

Given this shift from manipulated conditions to behaviorally derived groups, linear regression was deemed more appropriate for addressing the research questions. Preliminary analyses indicated that none of the covariates were significantly correlated with the outcome variables; therefore, no covariates were included in the regression models. A series of linear regression analyses were conducted to examine the effects of AI use type (no AI [reference], AI-generated, augmented, and prompt engineering) on outcomes across generation, evaluation, implementation, self-efficacy, and attitudinal domains.

Generation Outcomes (RQ1, RQ2, RQ5)

To address RQ1 and RQ2; examining how AI use during idea generation influences idea quality, novelty, and broader creative outcomes and RQ5, which focuses on prompt engineering versus other forms of AI use, regression analyses were conducted on fluency, quality, and originality.

Fluency

AI use type significantly predicted fluency, $F(3, 400) = 26.207$, $p < .001$, $R^2 = .164$. Relative to the no AI category, the AI-generated category was a significant positive predictor of fluency ($\beta = 0.907$, $p < .001$), and the prompt engineering category also showed a significant positive effect ($\beta = 0.456$, $p < .001$). The augmented category was not significantly different from the no AI category ($\beta = 0.049$, $p=0.681$).

Quality

AI use type significantly predicted generation quality, $F(3, 398) = 9.814$, $p < .001$, $R^2 = 0.069$ (see Table 3). All AI use types outperformed the no AI group, with the prompt engineering group demonstrating the largest effect (β = 0.899, $p < .001$), followed by the augmented category (β = 0.473, $p < .001$) and the AI-generated category (β = 0.288, $p < .05$).

Originality

AI use type also significantly predicted originality, F(3, 398) = 4.128, $p <.05$, though the variance explained was modest ($R^2 = 0.030$; see Table 3). The prompt engineering category (β = 0.608, $p <.001$) and the augmented category (β = 0.300, $p <.05$) were associated with higher originality relative to the no AI category, whereas the AI-generated category was not significant.

Evaluation Outcomes (RQ3, RQ4, RQ6)

To address RQ3 and RQ4, which focus on how AI use during idea evaluation influences idea selection and creative outcomes, and RQ6 regarding prompt-based AI use, analyses were conducted on evaluation quality and originality.

Evaluation Quality

AI use type significantly predicted evaluation quality, $F(3, 399) = 16.117$, $p < .001$, $R^2 = .108$ (see Table 4). Relative to the no AI category, all AI use types were significant positive predictors of evaluation quality, including the AI-generated category (β = 0.644, $p < .001$), the augmented category (β = 0.683, $p < .001$), and the prompt engineering category (β = 0.812, $p < .001$).

Evaluation Originality

AI use type also significantly predicted evaluation originality, $F(3, 398) = 9.524$, $p < .001$, $R^2 = 0.067$ (see Table 4). All AI use types were associated with higher originality relative to the no AI category, including the AI-generated category (β = 0.411, $p < .001$), the augmented category (β = 0.637, $p < .001$), and the prompt engineering category (β = 0.507, $p < 0.01$).

Implementation Outcomes (RQ2, RQ4, RQ6)

Implementation outcomes (i.e., overall product quality, originality, and elegance) were generally weak and inconsistent. The overall model predicting implementation quality was not statistically significant (see Table 5).

Self-Efficacy and Attitudinal Outcomes (RQ7)

To address RQ7, examining the impact of AI use on self-efficacy and related attitudinal outcomes, results were mixed. AI use type did not significantly predict task-specific self-efficacy, and no individual category differed from the no AI category. In contrast, AI use type significantly predicted AI-related self-efficacy, $F(3, 401) = 3.636$, $p < 0.05$, although the effect size was small ($R^2 = .026$). Specifically, both the AI-generated category ($\beta = 0.364$, $p < 0.05$) and the prompt engineering category ($\beta = 0.410$, $p < 0.05$) were associated with higher AI self-efficacy relative to the no AI category, while the augmented category was not significant.

Attitudinal outcomes showed a similar pattern of limited effects. The model predicting satisfaction was not statistically significant overall

Discussion

The present study examined how GenAI influences creative performance across key stages of the creative process: idea generation, evaluation, and implementation. Results indicate that patterns of GenAI use, rather than simply access to AI, were more strongly associated with creative outcomes. These findings align with emerging research suggesting that the benefits of large language models depend not only on whether they are used, but also on how users strategically engage with them (e.g., Ivcevic & Grandinetti, 2024; Sun et al., 2025). Behaviorally derived usage patterns consistently predicted performance, with prompt engineering emerging as the most effective approach. These findings suggest that specific forms of human–AI collaboration play a critical role in shaping idea generation and evaluation outcomes.

These findings can be understood through a metacognitive framework of human–AI collaboration, which helps explain why certain patterns of GenAI use were more effective than others. Prompt engineering reflects a form of metacognitive regulation, in which individuals actively structure, monitor, and refine their interactions with AI to align outputs with task goals. Drawing on metacognitive theory (Flavell, 1979) and process models of creativity (Mumford et al., 1991), conceptualizing GenAI as a cognitive augmentation tool whose effectiveness depends on the degree of user control may be a useful way to think about human-GenAI collaboration on creative tasks. When individuals engage in higher-order regulation, such as specifying goals, iteratively refining prompts, and critically evaluating outputs, collaborating with AI is more likely to enhance both generative and evaluative processes. In contrast, relying primarily on AI-generated outputs appear to increase fluency without consistently improving higher-order creativity. Thus, differences in creative performance may be explained by variation in metacognitive engagement rather than the presence of AI itself. This interpretation is also consistent with recent organizational research showing that employees with stronger metacognitive strategies derive greater creative benefits from LLM use than employees who engage more passively with AI (Sun et al., 2025).

Distinct patterns of human–GenAI collaboration during idea generation were differentially associated with creative performance. Fully AI-generated use (i.e., minimal human collaboration) was most strongly associated with idea fluency, suggesting that offloading idea production to GenAI may enhance the quantity of ideas generated. This finding aligns with prior research demonstrating that AI systems can rapidly produce numerous associations and combinations (Koivisto & Grassini, 2023). However, improvements in idea generation quality and originality were more strongly associated with prompt engineering and augmented use, suggesting that generating a greater number of ideas does not necessarily translate into more novel and useful idea generation output. This pattern is also consistent with Doshi and Hauser (2024), who found that although generative AI increased individual creative performance, AI-assisted outputs became more similar across participants. Together, these findings suggest that AI may increase ideational fluency while producing fewer uniquely original solutions.

Additionally, the results point to a distinction between ideational quantity and quality, with increases in idea volume not reliably accompanied by improvements in quality or originality. Although AI can increase fluency, higher-order creativity depends on how individuals evaluate and refine those ideas. Prompt engineering, in particular, was consistently associated with stronger outcomes, likely because it requires active cognitive engagement, including some form of goal specification, iterative refinement, and strategic prompting. This pattern is consistent with models of creative problem solving that emphasize the role of problem construction and controlled processing in producing high-quality ideas (Mumford et al., 1991). Within the creative process, AI appears to function most effectively as a tool that supports and extends human cognition rather than as a substitute for it.

Most empirical research on GenAI and creativity has focused on idea generation. By contrast, the present study demonstrates that AI may also provide meaningful benefits during idea evaluation, suggesting that AI's value extends beyond producing ideas to supporting convergent thinking and refinement. Across all AI use types, participants demonstrated higher evaluation quality and originality compared to those who did not use AI. These results suggest that AI may be particularly effective in supporting convergent thinking processes, such as comparing alternatives, identifying strengths and weaknesses, and selecting the most promising ideas. Given that idea evaluation requires integrating multiple criteria and anticipating future outcomes (Mumford et al., 2002), GenAI's ability to process and synthesize large amounts of information may provide a distinct advantage during this stage.

Importantly, prompt engineering GenAI emerged as the most effective form of engagement, suggesting that structured and intentional interaction with AI enhances evaluative accuracy. This study's findings support the interpretation of prompt engineering as a form of metacognitive support, whereby individuals clarify evaluation criteria, refine decision rules, and engage more deeply with the selection process. Together, these results extend prior work by demonstrating that AI can support not only idea generation but also the refinement and selection of ideas, particularly when human input actively guides and structures its outputs (Huang et al., 2025).

In contrast to the robust findings for generation and evaluation, results related to implementation outcomes were weaker and less consistent. These findings suggest that the benefits of AI use may not extend as strongly to later stages of the creative process, particularly when AI is not available during the implementation planning phase. One explanation is that implementation requires a distinct set of cognitive and behavioral processes, including sustained effort, domain knowledge, and the ability to translate abstract ideas into coherent, fully developed outputs (Mumford et al., 1991; Mumford et al., 2002). While AI may facilitate early-stage ideation and evaluation, it may provide less sustained support for the integration and elaboration required during implementation. This interpretation aligns with process-based models of creativity, which emphasize that different stages rely on distinct cognitive mechanisms (Mumford et al., 1991). Accordingly, improvements in early-stage creative processes do not necessarily translate into improvements in final creative products. Together, these results reinforce the importance of differentiating between stages of the creative process, as tools that enhance one stage (e.g., idea fluency) may not translate to improvements in others (e.g., originality or implementation).

GenAI use appears to have limited and nuanced implications for self-efficacy and task-related attitudes. Rather than broadly enhancing individuals' confidence in their own creative abilities, engagement with GenAI may primarily reinforce confidence in using the technology itself. This pattern suggests that interacting with AI does not necessarily translate into stronger beliefs about one's own creative competence. This interpretation aligns with Albert Bandura's (1997) conceptualization of self-efficacy as domain-specific. Confidence gained through AI interaction may remain confined to the technological domain, rather than generalizing to broader perceptions of personal creative ability.

More broadly, the study's findings suggest a disconnect between performance and perception. While individuals may benefit from GenAI in terms of task outcomes, these gains do not appear to meaningfully shape how they evaluate their own abilities or the experience itself. This implies that the advantages of GenAI may be primarily functional rather than psychological, enhancing performance without substantially altering self-perceptions or attitudes.

Theoretical Contribution

The present study contributes to the literature on creativity and human–AI collaboration in several ways. By shifting from experimentally assigned conditions to behaviorally derived use types, this study highlights the importance of user engagement patterns in shaping creative outcomes. Rather than treating AI use as a binary condition, the findings suggest that *how* individuals interact with AI is a primary driver of performance, positioning human–AI collaboration as a heterogeneous and strategy-dependent process. This distinction is particularly important because much of the existing literature operationalizes AI use as a binary comparison (AI versus no AI). The present findings suggest that this approach obscures meaningful variation in how individuals collaborate with AI. Rather than representing a single intervention, GenAI use encompasses multiple patterns of interaction that differ substantially in their implications for creative performance.

These findings extend process-based models of creativity, particularly the model proposed by Mumford et al. (1991), which conceptualizes creativity as a sequence of cognitive operations including problem construction, information encoding, idea generation, evaluation, and implementation (Mumford et al., 2012). The present results suggest that AI does not uniformly influence these stages. Instead, its effects appear concentrated in idea generation and evaluation, where AI can rapidly expand the solution space and facilitate comparison across alternatives

Importantly, these findings suggest that human–AI collaboration may reshape the dynamics of the creative process by amplifying its inherently iterative structure. While traditional models (e.g., process analytic models; Mumford et al., 1991) recognize that individuals cycle between stages such as generation and evaluation, AI-supported contexts appear to make these cycles more rapid, explicit, and interactive. In these contexts, individuals engage in ongoing cycles of prompting, output evaluation, and refinement, creating a dynamic feedback loop between user and system. This pattern aligns with perspectives on extended and distributed cognition (e.g., Hutchins, 1995; Clark, 2008) and suggests that creative processing in AI-supported contexts may be better conceptualized as an interactive system rather than a sequence of discrete stages. As such, the present findings extend Mumford's model (1991) by positioning AI as a stage-contingent and interaction-dependent influence that reshapes how creative processes unfold.

Third, this study advances the conceptualization of prompt engineering by framing it as a metacognitive skill rather than a purely technical behavior. Within a process model framework, processes such as problem construction and evaluation are central to creative performance. The present findings suggest that effective prompt engineering requires individuals to actively engage in these processes: defining the problem, structuring inputs, and critically evaluating outputs, indicating that prompt engineering reflects higher-order cognitive regulation. This interpretation is consistent with research on metacognition and self-regulated problem solving (e.g., Zimmerman, 2000), and positions AI use as a context in which cognitive control is partially externalized but still user-driven. Additionally, recent research similarly suggests that effective LLM use depends on users' ability to strategically direct, monitor, and refine AI outputs rather than simply accepting generated responses (Ivcevic & Grandinetti, 2024; Sun et al., 2025).

Finally, the findings contribute to emerging theory on human–AI collaboration by highlighting a divergence between performance outcomes and self-perceptions. Although AI use was associated with improvements in creative performance, it did not meaningfully enhance individuals' confidence in their own creative abilities. Instead, engagement with AI appeared to strengthen domain-specific confidence in using the technology itself. This pattern is consistent with Albert Bandura's (1997) conceptualization of self-efficacy as domain-specific, and suggests that performance gains in AI-supported contexts may not translate into broader changes in self-beliefs. Theoretically, this points to a model of human–AI collaboration in which cognitive augmentation occurs without corresponding psychological internalization, raising important questions about how individuals attribute success in AI-assisted work.

Practical Implications

These findings suggest that organizations should move beyond training employees simply to use AI tools and instead focus on developing effective human-AI collaboration strategies. Training programs should emphasize prompt engineering, iterative refinement, critical evaluation of AI outputs, and metacognitive regulation, as these skills appear to differentiate effective from ineffective AI use. In particular, developing prompt-engineering skills may be critical for maximizing the benefits of AI in creative tasks. More broadly, training efforts should emphasize metacognitive skills that enable individuals to strategically collaborate with GenAI. This includes learning how to plan interactions with GenAI (e.g., when to rely on AI for idea expansion), monitor outputs (e.g., evaluating the quality and relevance of generated ideas), and adapting strategies (e.g., engaging in iterative prompt refinement and adjusting inputs). Teaching individuals to regulate their interaction with GenAI in this way may be key to effectively leveraging AI-generated outputs in creative work. In this sense, effective AI use involves selecting among different modes of collaboration, such as delegating idea generation, augmenting one's own thinking through iterative prompting, or using AI as an evaluative partner to compare and refine alternatives. Teaching individuals to recognize and flexibly apply these collaboration strategies may be particularly important for complex or ill-defined creative tasks.

Limitations and Future Directions

Several limitations should be noted. First, because participants were reclassified based on observed behavior, the present analyses are more consistent with a quasi-experimental design, and causal inferences should be made cautiously. However, this approach enhances ecological validity by capturing how individuals naturally engage with AI tools, which may differ from strictly controlled experimental conditions. Second, the use of an undergraduate sample may limit generalizability to other populations, particularly those with greater domain expertise.

Third, the study focused on a single creative task, which may not capture the range of activities involved in human–AI collaboration across applied settings.

In addition, the broader organizational context in which AI is used introduces practical constraints that were not captured in the present study. For example, organizations may need to establish clear guidelines regarding what types of information can be shared with AI systems, particularly when working with proprietary or sensitive content. As AI tools become more integrated into creative workflows, concerns related to intellectual property protection and data security will become increasingly important. Future research should examine how such constraints shape patterns of AI use and whether limitations on information sharing influence the effectiveness of human–AI collaboration.

Future research should also examine how patterns of AI develop over time and whether repeated exposure leads to changes in creative performance or self-efficacy. Additionally, investigating underlying cognitive mechanisms, such as metacognitive regulation and cognitive load, may provide further insight into how and when AI is most beneficial. An important direction for future work involves understanding the role of iterative refinement in AI-supported creativity. While iterative prompting and revision appear central to effective AI use, it remains unclear how much refinement is optimal and when additional iteration may yield diminishing returns or even hinder performance. Examining the balance between exploration and over-refinement may help clarify how individuals can most effectively engage with AI in creative tasks.

Conclusion

In conclusion, this study demonstrates that generative AI can enhance creative performance, particularly during idea generation and evaluation, but that these benefits depend critically on how the tool is used. Prompt engineering, characterized by active and strategic engagement with AI, emerged as the most effective form of use, highlighting the importance of human input in human–AI collaboration. These findings reinforce the idea that AI is not inherently creative but can serve as a powerful tool for augmenting human creativity when used thoughtfully and intentionally.

References

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior 1. *Journal of Applied Social Psychology*, *32*(4), 665-683.

Amabile, T. M. (1983). A theoretical framework. *The Social Psychology of Creativity*, 65-96.

Amabile, T. M. (1988). A model of creativity and innovation in organizations. *Research in organizational behavior*, *10*.

Amabile, T. M., Conti, R., Coon, H., Lazenby, J., & Herron, M. (1996). Assessing the work environment for creativity. *Academy of management journal*, *39*(5), 1154-1184.

Baidoo-Anu, D., Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*. 7(1), 52-62.

Bandura, A., & Wessels, S. (1997). *Self-efficacy* (pp. 4-6). Cambridge: Cambridge University Press.

Basadur, M., Pringle, P., Speranzini, G., & Bacot, M. (2000). Collaborative problem solving through creativity in problem definition: Expanding the pie. *Creativity and Innovation Management*, *9*(1), 54-76.

Benedek, M., Beaty, R. E., Schacter, D. L., & Kenett, Y. N. (2023). The role of memory in creative ideation. *Nature Reviews Psychology*, *2*(4), 246-257.

Brophy, D. R. (1998). Understanding, measuring, and enhancing individual creative problem-solving efforts. *Creativity Research Journal*, *11*(2), 123-150.

Cropley, A. (2006). In praise of convergent thinking. *Creativity research journal*, *18*(3), 391-404

Cropley, D. H., & Marrone, R. L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*.

Doshi, A. R., & Hauser, O. P. (2023). *Generative artificial intelligence enhances creativity but reduces the diversity of novel content*. arXiv.

Forgeard, M. J., & Eichner, K. V. (2014). Creativity as a target and tool for positive interventions. *The Wiley Blackwell handbook of positive psychological interventions*, 135-154.

Girotra, K., Terwiesch, C., & Ulrich, K. T. (2010). Idea generation and the quality of the best idea. *Management science*, *56*(4), 591-605.

Haase, J., & Hanel, P. H. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, *33*(3), 100066.

Ivcevic Pringle, Z., & Grandinetti, M. (2024). *Artificial intelligence as a tool for creativity*. *Journal of Creativity, 34*(1), Article 100083.

Johnson, G., & Connelly, S. (2014). Negative emotions in informal feedback: The benefits of disappointment and drawbacks of anger. *Human Relations*, *67*(10), 1265-1290.

Kampylis, P. G., & Valtanen, J. (2010). Redefining creativity—analyzing definitions, collocations, and consequences. *The Journal of Creative Behavior*, *44*(3), 191-214.

Kaufman, J. C. (2016). *Creativity 101*. Springer Publishing Company.

Kubany, E. S., & Watson, S. B. (2003). Guilt: Elaboration of a multidimensional model. *Psychological Record*, *53*(1), 51-90.

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of vocational behavior*, *45*(1), 79-122.

Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. *The Journal of Creative Behavior*, *41*(1), 1-27.

Lubart, T. I. (2001). Models of the creative process: Past, present and future. *Creativity research journal*, *13*(3-4), 295-308.

Magni, F., Park, J., & Chao, M. M. (2024). Humans as creativity gatekeepers: Are we biased GenAIAInst AI creativity?. *Journal of Business and Psychology*, *39*(3), 643-656.

Medeiros, K. E., Marrone, R. L., Joksimovic, S., Cropley, D. H., & Siemens, G. (2023). Promises and realities of artificial creativity. In *Handbook of Organizational Creativity* (pp. 275-289). Academic Press.

Mumford, M. D., & Gustafson, S. B. (1988). Creativity syndrome: Integration, application, and innovation. *Psychological Bulletin*, *103*(1), 27.

Mumford, M. D., Hester, K., & Robledo, I. (2012). Methods in creativity research: Multiple approaches, multiple levels. In *Handbook of organizational creativity* (pp. 39-65). Academic Press.

Mumford, M. D., & McIntosh, T. (2017). Creative thinking processes: The past and the future. *The Journal of Creative Behavior*, *51*(4), 317-322.

Mumford, M. D., Mobley, M. I., Reiter-Palmon, R., Uhlman, C. E., & Doares, L. M. (1991). Process analytic models of creative capacities. *Creativity Research Journal, 4,* 91–122. https://doi.org/10.1080/10400419109534380

Mumford, M. D., Lonergan, D. C., & Scott, G. (2002). Evaluating creative ideas: Processes, standards, and context. *Inquiry: Critical thinking across the disciplines*, *22*(1), 21-30.

Unsworth, K. L., & Clegg, C. W. (2010). Why do employees undertake creative action?. *Journal of occupational and organizational psychology*, *83*(1), 77-99.

Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, *7*(11), 1836-1838.

Riggs, Matt L., and Knight, Patrick A. "The impact of perceived group success-failure on motivational beliefs and attitudes: A causal model." *Journal of Applied Psychology* 79.5 (1994): 755.

Runco, M. A. (2007). Creativity: Theories and Themes: Research. *Development and Practice. Amsterdam: Elsevier*.

Runco, Mark A., and Garrett J. Jaeger. "The standard definition of creativity." *Creativity research journal* 24, no. 1 (2012): 92-96.

Simonton, D. K. (2000). Creativity: Cognitive, personal, developmental, and social aspects. *American Psychologist*, *55*(1), 151.

Smith, G. F. (1998). Idea-generation techniques: A formulary of active ingredients. *The Journal of Creative Behavior*, *32*(2), 107-134.

Sun, S., Li, Z. A., Foo, M. D., Zhou, J., & Lu, J. G. (2025). How and for whom using generative AI affects creativity: A field experiment. *Journal of Applied Psychology*.

Tierney, P., & Farmer, S. M. (2011). Creative self-efficacy development and creative performance over time. *Journal of applied psychology*, *96*(2), 277.

Wang, Y. Y., & Chuang, Y. W. (2023). Artificial intelligence self-efficacy: Scale development and validation. *Education and Information Technologies*, 1-24.White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. *arXiv*. https://doi.org/10.48550/arXiv.2302.11382

Table 1
Means and Standard Deviations of Creative Performance Outcomes by AI Use Type

Outcome	AI-Generated *(n = 128)*	Augmented *(n = 103)*	Prompt Engineering *(n = 33)*	No AI *(n = 143)*
Generation Fluency	6.60 (5.86)	2.90 (2.82)	4.66 (3.92)	2.69 (2.20)
Generation Quality	2.71 (1.10)	2.92 (1.26)	3.40 (1.11)	2.38 (0.98)
Generation Originality	2.23 (0.76)	2.33 (0.95)	2.59 (0.93)	2.08 (0.79)
Evaluation Quality	2.69 (1.29)	2.74 (1.37)	2.90 (1.23)	1.88 (0.93)
Evaluation Originality	2.22 (0.91)	2.43 (0.99)	2.31 (0.86)	1.85 (0.80)
Implementation Quality	2.90 (0.91)	3.08 (0.95)	3.25 (0.90)	2.90 (0.73)
Implementation Originality	2.50 (0.97)	2.58 (1.00)	2.93 (1.04)	2.45 (0.99)
Implementation Elegance	2.44 (0.93)	2.57 (0.89)	2.73 (0.76)	2.47 (0.87)

Note. Values represent means with standard deviations shown in parentheses. AI use groups were behaviorally derived from participant interactions with the chatbot. AI-generated = participants who primarily relied on GenAI to generate ideas; Augmented = participants who generated their own ideas and used GenAI to elaborate, refine, or evaluate them; Prompt Engineering = participants who actively constructed and iteratively refined prompts to guide GenAI responses; No AI = participants who completed the task without using GenAI.

Table 2

Correlations Among Outcome Variables and Covariates

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Generation Fluency	—															
2. Generation Quality	-.02	—														
3. Generation Originality	.05	.61**	—													
4. Evaluation Quality	.09	.36**	.32**	—												
5. Evaluation Originality	.03	.34**	.59**	.61**	—											
6. Implementation Quality	-.02	.26**	.17**	.29**	.16**	—										
7. Implementation Originality	.01	.13**	.34**	.14**	.31**	.32**	—									
8. Implementation Elegance	-.04	.18**	.11*	.17**	.13*	.67**	.35**	—								
9. Satisfaction with AI	.00	0.09	-0.07	0.03	-0.06	0.09	-0.07	0.1	—							
10. Utility of AI	.00	.14*	-0.03	0.01	-0.04	0.09	-0.05	0.12	.86**	—						
11. Task Self-Efficacy	.02	0.03	0.07	-0.02	0.04	.12*	.17**	.11*	0.02	0.03	—					
12. AI Self-Efficacy	.09	-0.05	-0.05	0.01	-0.03	-0.02	0.02	-0.07	.31**	.32**	.17**	—				
13. Positive Affect	-.05	0.05	0.01	-0.05	-0.02	0.03	0.01	0.01	0.1	0.08	0.06	0.02	—			
14. Negative Affect	-.02	0.05	0.01	-0.01	-0.05	0.08	-0.04	0.02	0.02	0.01	-.10*	-0.05	.55**	—		
15. Trait Fluency	.05	0.07	0.06	0.04	0.04	0.08	0.05	0.08	-0.12	-.19**	0.05	-.14**	0.02	0	—	
16. Intelligence	.03	0.04	0.08	-0.01	0.03	0.05	.16**	0.07	-0.06	-0.06	0.02	-0.03	-.13**	.12*	.15**	—

Note. Values represent Pearson correlations. Generation, evaluation, and implementation variables reflect performance across stages of the creative process.* p < .05. ** p < .01.

Table 3

Regression Results for Idea Generation Outcomes

Predictor	Fluency β	Quality β	Originality β
AI Generated	0.907***	0.288*	0.187
Augmented	0.049	0.473***	0.300*
Prompt Engineering	0.456*	0.187***	0.608**

Note. Coefficients are standardized (β). The reference group is the no AI condition. Fluency: $R^2 = .164$, $F(3, 400) = 26.207$, $p < .001$. Quality: $R^2 = .069$, $F(3, 398) = 9.814$, $p < .001$. Originality: $R^2 = .030$, $F(3, 398) = 4.128$, $p < .01$. $p < .05$*, $p < .01$**, $p < .001$***.

Table 4

Regression Results for Idea Evaluation Outcomes

Predictor	Quality β	Originality β
AI Generated	0.644***	0.411**
Augmented	0.683***	0.637**
Prompt Engineering	0.812***	0.507*

Note. Coefficients are standardized (β). The reference group is the no AI condition. Quality: $R^2 = .108$, $F(3, 399) = 16.117$, $p < .001$.Originality: $R^2 = 0.067$, $F(3, 398) = 9.524$, $p < .001$. $p < .05$*, $p < .01$**, $p < .001$***.

Table 5

Regression Results for Plan Outcomes

Predictor	Quality β	Originality β	Elegance β
AI Generated	-0.009	0.059	-0.034
Augmented	0.208	0.134	0.122
Prompt Engineering	0.401*	0.487*	0.295

Note. Coefficients are standardized (β). The reference group is the no AI condition. Quality: $R^2 = 0.017$, $F(3, 397) = 2.344$, $p = .073$. Originality: $R^2 = .017$, $F(3, 397) = 2.248$, $p = .082$. Elegance: $R^2 = .009$, $F(3, 396) = 1.238$, $p = .295$. $p < .05$*, $p < .01$**, $p < .001$***.

Appendix A

Prompt Engineering Information Sheet

What is ***Prompt Engineering***?

Prompt engineering is the skill of writing clear and precise instructions for AI Chatbots, so the tool can understand and respond the way you want. Being mindful of the way you phrase requests for the AI chatbot can play a major role in validity, accuracy, and desirability of AI output.

How to be Mindful of Prompt Engineering

Think about your intent

- Consider the type of interaction you want to have with AI.
- What type of task are you asking the AI to do?

Think about the Chatbots understanding

- Consider what AI chatbots are particularly good at and what they struggle with.
- For example, an AI Chatbot could be great at pulling relevant information and coming up with ideas, but weak at providing detail for the ideas they generate.
- Continue to ask questions until you get the information you need.

Think about Vocab

- Consider what vocabulary the AI model is best suited to understand.
- For example, AT&T's "Ask AT&T" AI software is better suited for technological and company-specific verbiage than ChatGPT.
- If you are not getting the response you want, consider rephrasing your prompt to more generic vocabulary.

Think about Clarity and specificity

- Consider word choice and attempt to make AI requests as direct as possible.
- Ask the Chatbot for a specific number of ideas at a time rather than using words like some or many ideas.

Think about Constraints

- Consider limitations of the AI model (Ex. Word count, character output, format)
- Ask the Chatbot for 300 words related topic rather than a simple request for information on the topic

Alternating Between High Level and Low-Level Questions

- Test the AI's answering ability by varying factual questions with complex questions.
- Ask for varying levels of detail.

Appendix B

Benchmark Rating Scale Example

Rating Scales and Benchmarks for Idea Generation

1. Quality

Definition: The extent to which the idea is realistic, practical, and appropriate for the situation

Markers:

- Logical – Is the idea coherent and reasonable (does it make sense)?
- Coherence- Is the response coherent?
- Usefulness—Is the idea a realistic and appropriate response for addressing the task?

Scale:
1 = Poor quality
3 = Average quality
5 = Excellent quality

Benchmarks:
• 1 (Poor): The idea is of <u>poor quality</u>. The idea is illogical and incomplete. The key issues are not addressed. The plan was not well thought out and will not address the task at hand.

"On a beach, seafood, glass windows, inside and outside, modern"

• 3 (Average): The idea is of <u>average quality</u>. The plan addresses the task but may contain irrelevant elements. The idea makes general sense overall and is likely to address the situation.

"Tony's Pizza and Pasta: small hole in the wall vibe, local, old Italian restaurant vibes with the red and white checker tablecloth and the weird drip candles"

• 5 (Excellent): The idea is of <u>excellent quality</u>. This response is of excellent quality. The response is very clear and coherent. The idea is very likely to address and fix the problem.

"Enchanted Garden: A restaurant with an indoor garden theme, featuring greenery, hanging lights, and garden-style seating. The menu could focus on farm-to-table dishes with an emphasis on fresh produce and floral infusions in both food and drinks, like lavender lemonade or rose-infused desserts."

2. Originality

Definition: The degree to which the response is novel, original, and unexpected.

Markers:

- Unexpectedness – Is the idea novel, imaginative, unpredictable, or innovative?
- Novelty/Originality – Is the idea new, original, or not indicative of the "typical" plan?

Scale:
1 = Poor originality
3 = Average originality
5 = Excellent originality

1	2	3	4	5
Poor Originality		***Average originality***		***Excellent originality***

Benchmarks:

• 1 (Poor): The idea is of poor originality. The response is very predictable and does not provide any unique ideas. The response lacks elaboration and descriptiveness.

"Asian Cuisine"

• 3 (Average): The idea is of average originality. The response has a few original and unique elements; however, it still contains predictable concepts. The response is somewhat descriptive.

"*Artisan Pizzeria: A cozy pizzeria that offers unique toppings and local ingredients. The theme could include an open kitchen where guests can watch their pizzas being made, and the décor can feature local art pieces. Options might include seasonal pizzas or global-inspired flavors."*

• 5 (Excellent originality): The plan is of excellent originality. The response is clearly unique and has elements that appear to be wholly original.

"Concept 1): Interactive dinner. This includes having your guests "choose" the menu of the day in a series of games. The atmosphere will be fun and exciting but with some modern touches as well. For this concept, the dinner will be a tasting menu and how your guests will "choose" the menu is lets say the waiter has a deck of cards, and then he has your guests draw two cards from that deck. Those two cards are the options for the tasting menus, and then the waiter labels each card with the option 1 or 2 and then has the guests pick that card."

3. Fluency

Definition: The total number of relevant, meaningful, and non-redundant ideas generated by the participants

Examples:

Response	Farm-to-table, Pop-Up Dining, Food Truck Park. Culinary Theatre, Gastro Pub, Ethnic Cuisine, Health-Conscious Café, Themed Diner, Interactive Dining, Dessert Bar	A French style restaurant that serves brunch and dinner. The lighting would be soft and live quiet music would be played. The theme of the restaurant would be Paris. Paris during the day for brunch, and Paris at night time during dinner. An interactive restaurant that is primarily focused fir kids. They will be presented with a story and the meal will correspond to the story. The theme of the restaurant will be whatever story id being told at that time.
Number of Ideas	10 Total Items	2 total items