Human Creativity in the Age of LLMs

Randomized Experiments on Divergent and Convergent Thinking

Harsh Kumar
harsh@cs.toronto.edu
University of Toronto
Toronto, Ontario, Canada

Jonathan Vincentius
jon.vincentius@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Ewan Jordan
ewan.jordan@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Ashton Anderson
ashton@cs.toronto.edu
University of Toronto
Toronto, Ontario, Canada
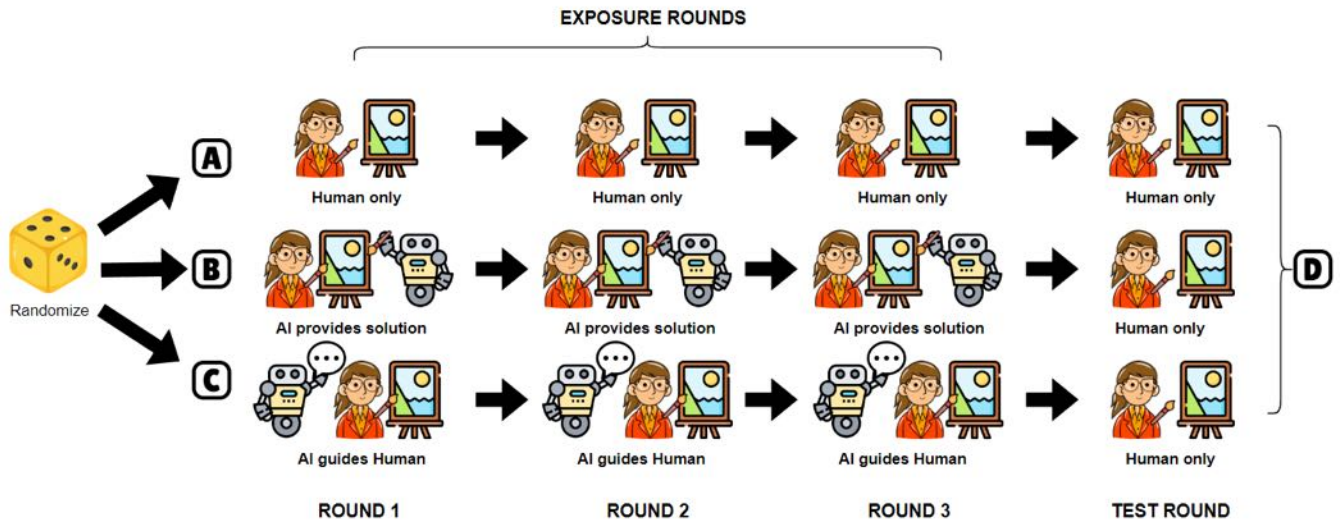
Figure 1: Experimental framework for measuring the impact of AI use on Human creativity. Participants engage in a series of *Exposure* rounds where they are randomized to either receive - (A) No assistance, (B) LLM solution (standard): This could be analogous to using a chat LLM such as ChatGPT for the task, or (C) LLM guidance (coach-like): In this case, participants receive response from a customized LLM which guides them through the creative process. Finally, in the last round, all participants are asked to do the same creative task without any assistance as a *Test*. (D) The performance and creative outputs in this unassisted round are the primary measures for evaluating the impact of using LLMs on Human cognition.

Abstract

Large language models are transforming the creative process by offering unprecedented capabilities to algorithmically generate ideas. While these tools can enhance human creativity when people co-create with them, it's unclear how this will impact unassisted human creativity. We conducted two large pre-registered parallel experiments involving 1,100 participants attempting tasks targeting the two core components of creativity, divergent and convergent thinking. We compare the effects of two forms of large language model (LLM) assistance—a standard LLM providing direct answers and a coach-like LLM offering guidance—with a control group receiving no AI assistance, and focus particularly on how all groups perform in a final, unassisted stage. Our findings reveal that while LLM assistance can provide short-term boosts in creativity during assisted tasks, it may inadvertently hinder independent creative performance when users work without assistance, raising concerns about the long-term impact on human creativity and cognition.

CCS Concepts

- **Human-centered computing** → **Empirical studies in HCI**; *Laboratory experiments*; • **Applied computing** → **Arts and humanities**.

Keywords

creativity, divergent thinking, convergent thinking, large language models, experiments

1 Introduction

The rise of generative artificial intelligence tools such as ChatGPT has the potential to upend the creative process. By their nature as *generative* systems, they offer an unprecedented capacity to algorithmically generate ideas, and perhaps even to create. State-of-the-art generative AI systems have reached proficiency levels that not only match human creativity in certain evaluations [19, 32], but can also enhance the creative output of knowledge workers [1, 2, 29]. We are living through a time of creative transformation, with AI visual art, AI music, and AI-enhanced videos rapidly proliferating.

But what does the era of generative AI hold for human creativity? When AI assistance becomes a regular part of creative processes, will human creativity change? There are natural concerns that reliance on generative AI might impair an individual's inherent ability to think creatively without assistance. Further, there is preliminary evidence that widespread dependence on similar generative AI tools could lead to a homogenization of thought. This could in turn reduce the diversity that drives collective innovation and stifle breakthroughs across various fields. However, there is also reason to believe that AI assistance could spark human creativity. Working with a fresh creative partner might be similarly stimulating regardless of whether the partner is human or machine. If this is the case, AI could prove to be a force for the flourishing of human creativity. Understanding how and whether co-creating with AI affects an individual's ability to generate creative ideas independently is therefore of critical importance.

This is a complex challenge. Creativity is inherently subjective and difficult to measure, making it hard to assess changes in an individual's creative ability. Factors like prior knowledge, experience, and environmental context all interplay with creativity, making it hard to isolate the effects of AI usage "in the wild". The rapid evolution of AI technologies further complicates matters, as their capabilities and effects are constantly changing. Without clear methods to evaluate these impacts, we cannot fully grasp how AI integration might alter the creative landscape.

In this work, we investigate the impact of Large Language Models (LLMs) assistance on human creativity by examining the two fundamental components of creativity: divergent and convergent thinking. Divergent thinking involves the generation of multiple, unique ideas, fostering exploration and innovation [48, 50]. Convergent thinking, in contrast, focuses on refining these ideas to select the most effective solutions [4, 24]. We investigate how different forms of LLM assistance influence these cognitive processes by comparing two types of LLMs—a standard LLM that provides direct answers out of the box, and a coach-like LLM that offers guidance and prompts to stimulate thinking—in contrast to a control condition with no LLM assistance.

Specifically, we address the following research questions:

RQ1 How do standard LLM assistance and coach-like LLM guidance, compared to no assistance, affect an individual's divergent thinking abilities when generating creative ideas independently?

RQ2 What are the impacts of standard LLM assistance and coach-like LLM guidance, versus no assistance, on an individual's convergent thinking skills in independently refining and selecting ideas?

To answer these questions, we designed and conducted two pre-registered parallel experiments with 1,100 participants that assess how these forms of LLM assistance influence unassisted human creativity, compared to a control group with no AI assistance. Figure 1 illustrates the high-level design of the experiments. In both experiments, participants were randomly assigned to one of three conditions: standard LLM assistance, coach-like LLM guidance, or no assistance (control). They engaged in a series of *exposure* rounds in which they completed creative tasks using their assigned form of LLM assistance. After a delay period, participants then completed the same type of creative tasks unassisted in the *test* rounds. This design allowed us to examine both the immediate effects of LLM assistance during the exposure rounds and the residual effects on unassisted creative performance during the test rounds.

For divergent thinking, we utilized the Alternate Uses Test (AUT), where participants were asked to come up with creative uses for common objects. We found that exposure to LLM assistance—whether providing ideas or strategies—did not enhance participants' originality or fluency in subsequent unassisted tasks. In some cases, it even led to decreased originality and reduced diversity of ideas, suggesting a potential homogenization effect where individuals generate more similar ideas after using LLM assistance. We employed the Remote Associates Test (RAT) for convergent thinking, which requires finding a word that connects three given words. Our findings indicate that while LLM assistance improved performance during the assisted tasks, it did not translate into better performance in subsequent unassisted tasks. Participants who received guidance from LLMs performed worse in the unassisted test rounds compared to those with no prior LLM exposure.

The paper contributes empirical findings on the impact of LLM assistance on human creativity, specifically focusing on divergent and convergent thinking. Specifically, our study (1) provides empirical evidence that LLM assistance boosts performance during assisted tasks but may hinder independent creative performance in unassisted tasks; (2) reveals the differential impacts of LLMs on divergent and convergent thinking, highlighting users' skepticism toward LLM assistance in divergent tasks and beneficial effects in convergent tasks; and (3) identifies persistent homogenization effects due to LLM-generated strategies, posing challenges for designing effective LLM coaching systems. The paper also offers design implications for developing LLM-based tools that enhance human creativity without undermining independent creative abilities.

2 Related Work

This paper builds on a rich body of literature on creativity, LLMs, and the impact of generative AI on human cognition. Although previous studies have laid the groundwork in these fields, our work makes novel contributions toward understanding the impact of different forms of LLM assistance on human creativity.

2.1 Theories of Human Creativity

Divergent and convergent thinking constitute critical components of the creative process [16, 18, 23, 31, 45, 59]. **Divergent thinking** involves generating a wide range of ideas, exploring multiple possibilities, and embracing unconventional approaches [4, 24, 48, 50]. In

contrast, **convergent thinking** focuses on narrowing down these ideas, selecting the most viable options, and refining them into coherent solutions. Using LLMs can differentially influence these processes, with distinct immediate and long-term effects on divergent and convergent thinking. We draw on key theories by Poincaré and Boden to frame our investigation [9, 20, 37]. Poincaré's four-phase model encapsulates creativity as both an unconscious and an active process. Divergent thinking is vital during the *Preparation* and *Incubation* phases, where the mind explores various possibilities [4, 22]. Convergent thinking becomes crucial in the *Insight* and *Revision* phases, refining ideas into viable solutions [17, 53]. This model underscores that creativity involves not only generating numerous ideas, but also selecting and refining them [28, 60].

Boden's theory further introduces measures of creativity, distinguishing between **P-creativity** (psychological creativity) and **H-creativity** (historical creativity) [8, 9]. P-creativity refers to ideas novel to the individual, while H-creativity refers to ideas novel within the broader context of human knowledge. This distinction allows us to assess creativity on both a personal and historical scale. Divergent thinking can be hypothesized to contribute primarily to P-creativity, where the generation of new ideas is the key [48–50]. In contrast, convergent thinking is essential for advancing these ideas toward H-creativity, where their novelty and value must be recognized within a broader context.

Our experiments explore how LLMs influence these aspects of creativity. To investigate divergent thinking, we utilize the Alternate Uses Test (AUT), where participants are prompted to generate as many creative uses as possible for a common object [21, 24]. For convergent thinking, we use the Remote Associates Test (RAT), where participants find a single word that connects three seemingly unrelated words, assessing their ability to converge on a correct and meaningful solution [35, 40, 58].

2.2 LLMs as Tools for Creativity

Large Language Models (LLMs) have demonstrated notable creative performance, often performing as well as, or even better than, average humans in various creative tasks. However, they still fall short when compared to the best human performers, particularly experts [30, 32]. For instance, Chakrabarty *et al.* [14] found that while LLMs performed competently in creative writing tasks, professionals consistently outperformed them. Anderson *et al.* [2] observed that ChatGPT users generated more ideas than those in a control condition. However, these ideas tended to be homogenized across different users, suggesting a limitation in the diversity of creative output when LLMs are used. These findings are particularly relevant when LLMs are employed to perform entire tasks on behalf of users.

Beyond performing tasks autonomously, LLMs can also be utilized to guide users through the *conceptual spaces* of creative thinking, serving as tools for 'creativity support' [1, 29, 43]. This approach positions LLMs not merely as replacements for human creativity but as enhancers of the creative process, helping users navigate and explore creative possibilities more effectively. The literature on Human-AI collaboration in creative tasks is rapidly growing. Lee *et al.* [36] introduced a dataset for analyzing GPT-3's use in creative and argumentative writing, suggesting that the HCI community

could foster more detailed examinations of LLMs' generative capabilities through such datasets. Suh *et al.* [55, 56] highlighted that the current interaction paradigm with LLMs tends to converge on a limited set of ideas, potentially stifling creative exploration. They proposed frameworks that facilitate the exploration of structured design spaces, allowing users to generate, evaluate, and synthesize many responses.

A key question remains: what happens to human creativity, particularly cognition, when humans repeatedly use LLMs for creative tasks—either directly to generate ideas or solutions, or as guides through the creative process? While there are concerns that this could lead to a deterioration of creative abilities, there is also optimism that these tools, if properly designed, could enhance human creativity while providing momentary assistance [19, 26]. However, empirical evidence to answer this question definitively is still lacking.

2.3 Generative AI and Human Cognition

The continued use of Generative AI is significantly impacting our society, particularly in how *culture* is created and propagated [3, 12]. These technologies are reshaping the production of cultural artifacts and the means through which culture is disseminated and experienced. The influence of Generative AI extends to our cognitive abilities, with effects that may be polarizing [26]. On the one hand, the pervasive use of AI tools might lead to a massive homogenization of creative output, a trend that could persist even after humans stop using these tools [2]. This raises concerns about the potential stifling of creativity and the reduction of diversity in thought and expression. On the other hand, Generative AI can unlock unprecedented creative growth and learning, allowing users to expand their cognitive horizons and engage in innovative forms of thought.

Hofman *et al.* [27] introduced a sports metaphor to conceptualize the spectrum of the impact of Generative AI on human cognition. They describe three distinct roles that AI can play: steroids, sneakers, and coach. "Steroids" represent AI as a tool that provides short-term performance gains, but with potential long-term detrimental effects. "Sneakers" symbolize AI tools that augment human skills without long-term adverse consequences. Lastly, the "Coach" role reflects AI as a guide that helps individuals improve their own capabilities, extending beyond immediate assistance to foster long-term cognitive growth. Collins *et al.* [15] proposed AI systems as 'thought partners,' designed to meet human expectations and complement our cognitive limitations. They outline several modes of collaborative thought in which humans and AI can engage, drawing insights from computational cognitive science to suggest how these partnerships can enhance human thinking.

Although the long-term effects of Generative AI use has been studied empirically in domains such as education (for e.g., [5, 34]), web search (for e.g., [54]), etc. There is a lack of empirical evidence on the effect of Generative AI tools on our convergent and divergent thinking abilities.

3 Experiment 1: Divergent Thinking

A critical aspect of creativity is the ability to generate a wide range of high-quality ideas, often called divergent thinking [24, 48]. LLMs

have the capacity to generate a significant quantity of ideas, often exceeding what an individual can produce, without the constraints of time or context. However, the impact of using LLM-generated ideas on an individual's ability to think divergently and come up with ideas independently remains underexplored. In addition to producing ideas, LLMs can offer structured frameworks that guide users in their creative processes, much like a coach [4]. To understand these dynamics, our first pre-registered[1] experiment investigates the effects of LLMs that either directly provide ideas or guide participants using a framework, helping them to develop their own ideas.

3.1 Experimental Design

We designed a three-condition, between-subjects experiment to understand how participants' unassisted divergent thinking is affected by the presence of LLM assistance during prior divergent thinking tasks (see Figure 2). The study was approved by the ethics board of the local university. We employed the Alternate Uses Test (AUT), which is the most widely used divergent thinking task [24]. Participants in this task were asked to come up with novel and creative uses for common everyday objects, outside of their intended use. They were told "The goal is to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different. Your ideas don't have to be practical or realistic; they can be silly or strange, even, so long as they are CREATIVE uses rather than ordinary uses...". For instance, an alternate use of pants might be as a makeshift flag.

3.1.1 Experimental conditions and treatment. Our experimental design involved two main phases: an exposure (paired) phase during which participants attempted the task with an LLM partner, and a test (solo) phase during which participants attempted the task on their own. Participants were given two-minute time frames per object to submit their ideas one by one, and could freely edit or delete any previously submitted ideas. During the exposure phase, participants are introduced to three objects, one after the other. Every participant was randomly assigned to receive one of three types of LLM responses:

None:	Control group. No LLM support provided.
List of Ideas:	GPT-4o generated a list of alternate uses for the given object. Seven randomly sampled uses were shown to the participant, which they could freely use in their responses. In other words, an LLM attempted the task and shared its responses with the participant.
List of Strategies:	GPT-4o, using a specialized system prompt (shown in Figure 11), generated seven strategies based on the SCAMPER technique [51]. In this condition, an LLM provided guidance to the participant but refrained from sharing explicit answers to the task.

The assigned type of LLM response appeared 5 seconds after the item was shown to the participant, and appeared character by character, similar to other chat LLM interfaces. Figure 2 shows the schema of the experiment design. Following the exposure phase, participants engaged in a brief distractor task to simulate forgetting,

playing a game of Snake for one minute. In this subsequent phase, participants were assigned the Alternate Uses Task for a new object selected at random, this time without LLM support.

Participants completed a pre-survey before beginning the experiment and a post-survey after completing the test phase, where we collected self-perceived creativity levels and their attitudes toward AI, along with other subjective measures (such as perceived difficulty of the test round, any strategies they utilized, and if there were any technical issues).

3.1.2 Stimuli. The items in each round were randomly sampled from five items: tire, pants, shoe, table, bottle. We chose these five particular objects as the originality scoring measure had the highest correlation with human judgements for these five objects [42]. Figure 3 shows the different responses shown to participants in different conditions.

3.1.3 Dependent Variables. AUT allows us to measure different dimensions of divergent thinking. As such, using LLMs may impact each of these dimensions differently. These dimensions include:

- **Originality (How original the idea is):** We measure the originality of each AUT idea with an existing fine-tuned GPT-3 classifier [42]. The model was fine-tuned with human judgments of AUT originality (where human raters judged, on a scale of 1 to 5, the originality of an idea given the object) and achieved an $r = 0.81$ overall correlation with human judgments. For our experiment, we chose the five items that had the highest accuracy for the model ($r > 0.88$).
- **Fluency (How many ideas):** We measure fluency by counting the number of ideas the participant generated in a given round.
- **Individual-level diversity of idea set (How different are ideas compared to each other for a given object):** We first embed all ideas using SBERT [46]. Idea diversity is the median pairwise cosine distance between idea embeddings in the idea set. This is a complementary measure to Originality. While Originality is a property of the idea, Diversity is a property of the idea set, so it is possible to have a diverse set of non-creative ideas if each individual idea is not creative (by originality metric) but different from one another [3].
- **Creative flexibility (How different are ideas in the Test round compared to the Exposure rounds):** Creative flexibility allows participants to switch between different concepts and perspectives. This is also sometimes referred to as **p-creativity**. In the framework of our experiment, we operationalize flexibility by checking how similar the ideas generated in the Test round are to the ideas generated in the Exposure rounds. We measure this by finding the maximum cosine similarity between the embedding of an idea in the Test set and all the ideas in the Exposure set. We use the maximum rather than a measure of central tendency because if a participant is inspired by an idea, it would likely be a single idea [47].
- **Group-level diversity of ideas (How different are the ideas generated by participants across the group for a given phase):** The goal of this analysis was to simulate how individuals contribute to idea generation as a group,
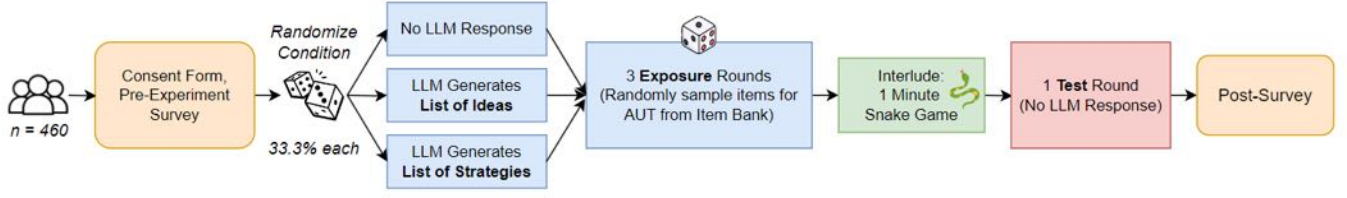
Figure 2: Schematic of design for Experiment 1 on divergent thinking.

with a focus on understanding the potential homogenization of ideas at the group level. Specifically, we aimed to study whether LLM assistance leads to a convergence of ideas, reducing diversity within a group. To achieve this, we performed 150 Monte Carlo runs, where we sampled 7 ideas per round, simulating group contribution dynamics. We calculated the median pairwise cosine distances between the embeddings of these ideas, using SBERT to encode them. This sampling process reflects the number of participants in each condition. The median pairwise cosine distance across each Monte Carlo run was used to evaluate diversity for each phase and condition. This approach allowed us to assess whether different forms of LLM assistance (e.g., providing direct ideas or acting as a coach) promote homogenization or help maintain diversity of ideas at the group level, even after stopping to use LLMs.

3.1.4 Analysis. We follow a long tradition of scoring responses to the AUT computationally [6, 7]. Following our pre-registration, we conducted a Kruskal-Wallis H test to compare the distribution of originality across the three conditions: None, List of Ideas, and List of Strategies, using a significance level of 0.05. In the event of a significant Kruskal-Wallis result, Dunn's test with Bonferroni correction was applied as a post-hoc test for pairwise comparisons between the conditions.

This same analysis procedure was applied for other dependent variables, including individual-level and group-level diversity, fluency, and creative flexibility. We report the significance of the overall Kruskal-Wallis H test and the results of Dunn's test for specific pairwise comparisons for each measure.

3.1.5 Participants. We recruited 460 participants from Prolific. Based on a power analysis using simulated and pilot data, we determined that this sample size will be necessary to achieve 70% power with a moderate effect size and a significance level of 0.05. The overall experiment took around 12 minutes to complete and participants were paid $1.57. Participants were based in the US or UK, and fluent in English. On average, participants felt they were more creative than 48.12% of the population, at the start of the experiment.

3.2 Results

We report on the analysis of 9,457 ideas generated by participants across all conditions and phases.

3.2.1 Originality. Figure 4a shows the average originality of ideas across conditions. In the *Exposure* phase, the mean originality was

similar across conditions, as indicated by the Kruskal-Wallis H test ($H(2) = 3.78$, $p = 0.151$). Interestingly, participants who received the *List of Ideas* performed slightly worse than other conditions, despite being exposed to LLM ideas with a mean originality of 3.25 (± 0.1). This is nearly 0.5 points higher than the average originality of ideas generated by participants in this condition (on a scale of 1 to 5). This discrepancy suggests that participants may struggle to accurately assess the quality of AI-generated ideas, or perhaps, when provided with AI ideas, they may prioritize coming up with their own, potentially lower-quality ideas. These findings highlight the need for designing effective reliance mechanisms that help users fully leverage the benefits of AI-generated ideas, beyond simply improving AI's output quality.

In the *Test* phase, the pattern shifts. The Kruskal-Wallis H test reveals significant differences in originality across conditions ($H(2) = 9.14$, $p = 0.010$). Post-hoc pairwise comparisons using Dunn's test with Bonferroni correction indicate that participants exposed to the *List of Strategies* performed significantly worse than those with no LLM exposure ($p = 0.009$). Additionally, the originality of participants in the *List of Strategies* condition decreased from the exposure to the test round, suggesting that they did not successfully internalize the strategies well enough to apply them independently. Neither the comparison between *List of Ideas* and *List of Strategies* ($p = 0.126$), nor the comparison between *List of Ideas* and *No LLM Response* ($p = 1.000$) showed significant differences. This suggests that the *List of Ideas* condition may fall somewhere in between, without a strong directional impact on originality. Overall, these results suggest that participants tended to perform better, in terms of originality, when they had no prior exposure to LLMs in the test phase.

3.2.2 Fluency. Figure 4b shows the average number of ideas generated by participants across different conditions. During the *Exposure* phase, the Kruskal-Wallis H test revealed significant differences in fluency across conditions ($H(2) = 8.57$, $p = 0.014$). Participants who received the *List of Strategies* produced significantly fewer ideas compared to those without LLM exposure, as indicated by Dunn's post-hoc test ($p = 0.011$). This may be because reading and applying strategies require more time, potentially reducing the number of ideas generated, even though the originality of these ideas remained comparable to the *No LLM Response* condition (as discussed earlier). Interestingly, participants who received the *List of Ideas* submitted nearly two fewer ideas than what was shown to them on average per round, being shown seven ideas each round, possibly supporting the hypothesis that individuals may prioritize generating their own ideas over simply adopting AI-provided ones.

Figure 3: Interface used in the divergent thinking experiment across all 3 exposure round conditions and test rounds.

In the *Test* phase, however, the Kruskal-Wallis H test did not indicate any significant differences in fluency across conditions ($H(2)$ = 1.14, p = 0.566), suggesting that prior exposure to LLMs may not significantly impact the quantity of ideas participants can produce independently in subsequent rounds of AUT.

3.2.3 Creative Flexibility. Figure 4c shows the average creative flexibility across conditions, which measures how different the ideas produced in the Test round are from those produced in the Exposure rounds (a higher value indicates more dissimilarity). The Kruskal-Wallis H test revealed a significant difference in creative flexibility across conditions ($H(2)$ = 6.31, p = 0.043). However, post-hoc pairwise comparisons using Dunn's test with Bonferroni correction did not reach significance after correcting for multiple comparisons.

Directionally, participants in the *List of Strategies* condition tended to produce ideas in the Test round that were more similar to those from the Exposure rounds compared to other conditions (p = 0.052 when compared to *No LLM Response*). Applying the same strategies to different objects may have led to more similar final outcomes across rounds.

3.2.4 Individual- and Group-level Diversity of Idea Set. Figure 5a shows the trend in the individual-level diversity of ideas produced by participants. In the *Exposure* phase, the Kruskal-Wallis H test revealed a significant difference between conditions ($H(2)$ = 9.95, p = 0.007). Post-hoc pairwise comparisons indicated that participants who received the *List of Strategies* produced significantly more similar ideas within each round compared to those in the *List of*
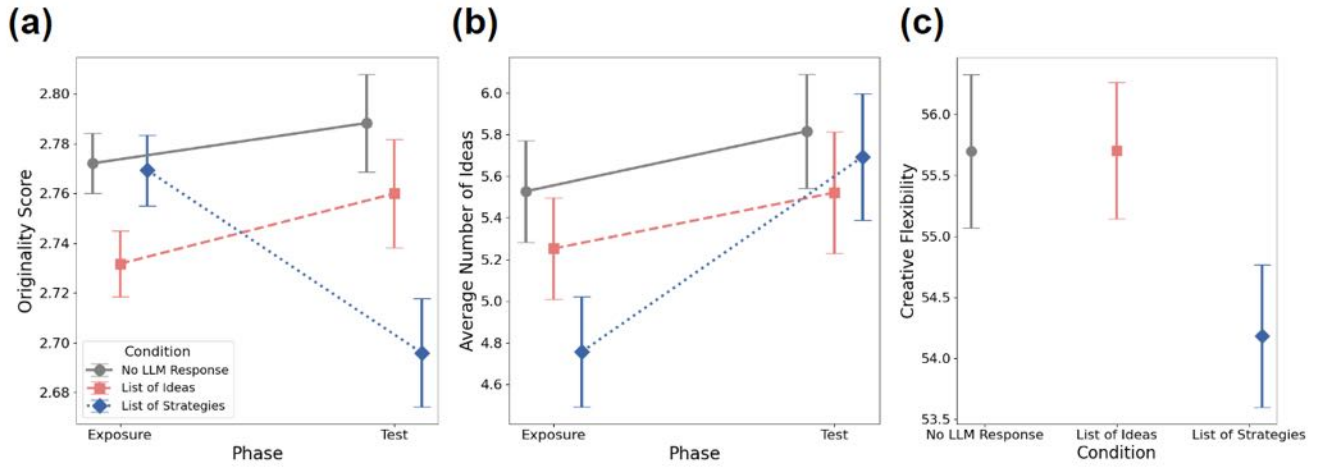
Figure 4: Plots of the Alternate Uses Task ideas for the various divergent thinking dimensions (Segmented by phase and/or LLM response type). The left figure shows idea originality scores, the middle figure indicates idea fluency, and the right figure presents participant creative flexibility.
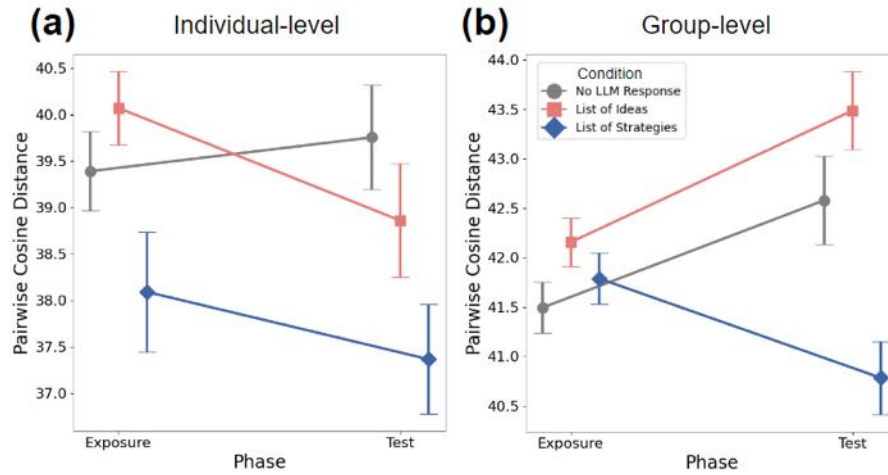

Figure 5: Plots of the average individual- and group-level median diversity, segmented by experiment conditions and phases. Higher values denote more difference between ideas. Error bars represent ± one standard error of mean.

Ideas condition ($p = 0.009$) and the *No LLM Response* condition ($p = 0.041$). In contrast, there was no significant difference between the *List of Ideas* and *No LLM Response* conditions ($p = 1.000$). The Kruskal-Wallis test for the *Test* phase also showed a significant difference between conditions ($H(2) = 7.71$, $p = 0.021$). Participants in the *List of Strategies* condition continued to generate more similar ideas compared to those in the *No LLM Response* condition ($p = 0.017$), although no significant difference was found between the *List of Strategies* and *List of Ideas* conditions ($p = 0.348$), or between *No LLM Response* and *List of Ideas* ($p = 0.747$).

Interestingly, while both LLM-assisted conditions showed a decline in idea diversity across rounds, participants in the *No LLM Response* condition maintained or even slightly improved their diversity. This result is somewhat unexpected, as it might have been assumed that unassisted participants would experience fatigue, leading to less varied ideas over time, whereas those who had support during the exposure rounds would be better equipped to produce diverse ideas.

We simulated group-level diversity by randomly sampling ideas across participants based on their condition, item, and phase. Figure 5b illustrates the trends in group-level diversity. In the *Exposure* rounds, the average median cosine distance between ideas was similar across all conditions ($p > 0.05$), which is unexpected. Participants in the *List of Ideas* condition, who had access to the same LLM-generated ideas, were expected to produce more similar ideas as a group. However, this may be due to participants not fully adopting the AI suggestions, as indicated by earlier findings. Additionally,
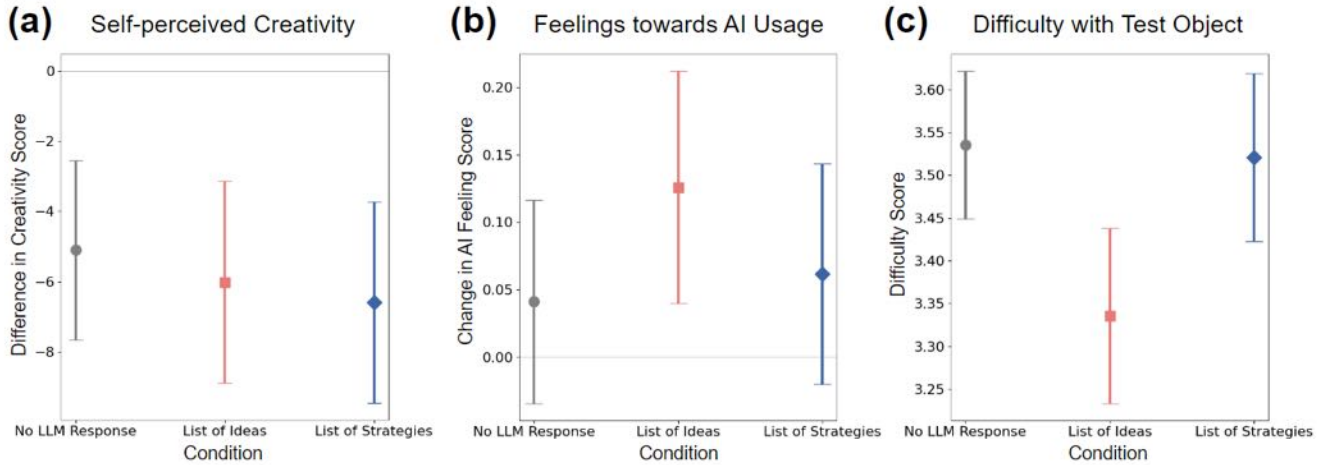
Figure 6: Plots of subjective measures collected before and after participants completed the Alternate Uses Task. The left figure shows participants' change in self perceived creativity levels (Based on how many % of humans they felt they were more creative than), the middle figure indicates how their feelings towards the increased use of AI computer programs in daily life changed (Between More concerned than excited/More excited than concerned/Equally excited and concerned), and the right figure presents how much difficulty they had in coming up with ideas for the test object, all segmented by the three LLM response types.

the random sampling of 7 ideas from a pool of 20 for each object may have contributed to the lack of overlap.

Interestingly, participants exposed to LLM-generated strategies during the exposure phase continued to generate more similar ideas in the test phase, even without LLM assistance, a phenomenon known as homogenization. This raises concerns that LLMs, which provide ubiquitous frameworks for thinking, could lead to reduced diversity in collective thinking, with people continuing to produce similar ideas even after they stop using the LLM.

3.2.5 Subjective Measures and Perceptions. Figure 6 summarizes participants' self-reported measures of creativity, feelings toward AI use, and the difficulty of generating uses for the *Test* object. Across all conditions, participants reported a decline in their perceived creativity after the experiment, a consistent drop that is somewhat unexpected. One might assume that receiving a list of AI-generated ideas would either further diminish creativity, as participants might feel they cannot match the AI, or conversely, seeing strategies during exposure rounds could boost their self-assessed creativity by providing structured guidance.

The change in participants' feelings toward AI remained relatively stable across conditions, with those in the *List of Ideas* condition reporting more than double the magnitude of change compared to the other groups. Interestingly, participants who received a list of ideas during the exposure rounds also found it easier to generate their own ideas during the test phase. This is surprising, as one might expect that exposure to AI-generated ideas would make it harder for participants to think independently in the test round, and that AI strategies in exposure would make it easier to independently come up with ideas.

4 Experiment 2: Convergent Thinking

The results from our first experiment highlight how LLMs shape our ability to generate ideas independently. However, creativity involves more than idea generation; it requires the capacity to identify and refine the most appropriate solutions within a given conceptual space. Convergent thinking is crucial for advancing from a broad set of ideas to selecting the most viable outcome. Motivated by this, our second pre-registered[2] experiment focuses on convergent thinking, testing not only LLM-generated answers but also the potential of LLMs to guide users toward solutions, simulating a coaching approach.

4.1 Experimental Design

We conducted a second three-condition, between-subjects experiment to understand how participants' unassisted convergent thinking performance are affected by the presence of LLM assistance during prior convergent thinking tasks (see Figure 7). We employed the Remote Associates Test (RAT), a widely recognized task for measuring convergent thinking. For each RAT problem, participants are presented with three words and asked to generate a fourth word that connects or fits all three words within a one-minute timeframe. For example, given the words *shelf, log*, and *worm*, the correct response would be 'book' (**book**shelf, log**book**, **book**worm). We selected this task due to its widespread use in prior research [16, 33], and its compatibility with LLMs. Similarly to the previous experiment, this study included both exposure (paired) and test (solo) phases. Figure 7 gives a high-level overview of the experiment schema. The experiment had participants complete a series of RAT problems alongside AI assistance (exposure phase), before

---
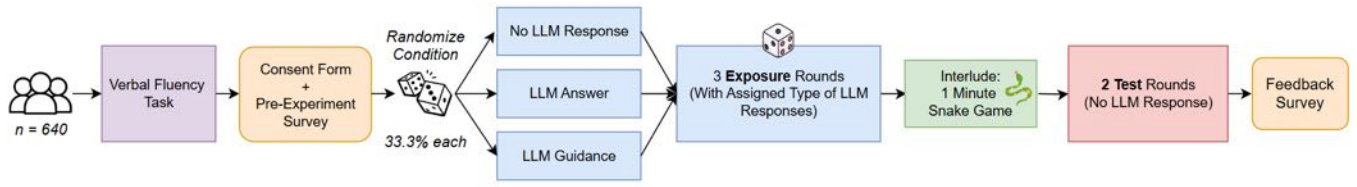
[2]https://aspredicted.org/V8K_ZTV

Figure 7: Schematic of design for Experiment 2 on convergent thinking.

completing unassisted RAT problems (test phase). The experiment was designed to give us insight into how convergent thinking skills are impacted by different levels of AI assistant, as well as by immediate prior use of AI assistance. Because a given RAT problem has a singular correct answer, we were able to define the metric for convergent thinking skills as accuracy on the RAT problems. We additionally collected perceptions and sentiments before and after the experiment to see how these were impact by completing the tasks alongside AI assistance or not.

4.1.1 Conditions and treatment. During the exposure phase, participants were required to solve three RAT problems consecutively. They were randomly assigned to one of three conditions, each providing different types of LLM assistance:

None:	Control group.
LLM Answer:	The answer generated by GPT-4o for the given set of words.
LLM Guidance:	GPT-4o was customized with a system prompt (shown in Figure 13) to generate possible associated words for each of the three given words. The model was instructed not to provide the solution directly but to encourage participants to make connections on their own, for example, by jotting the words down on paper.

After the exposure phase, participants played a game of Snake for one minute which served as a distractor task. In the test phase, all participants completed two additional RAT rounds (randomly selected) to attempt without any LLM assistance.

The *No LLM Response* (Figure 8a) condition serves as a baseline, reflecting the state of convergent thinking without AI assistance. The *LLM Answer* (Figure 8b) condition represents the scenario in which AI provides ready-made solutions. The *LLM Guidance* (Figure 8c) condition was designed to assist participants in navigating the conceptual space of the problem without directly revealing the answer. This approach aims to enable participants to solve the problem independently while still offering useful guidance in the moment. Research has shown that priming with related words and manually writing responses can enhance performance on RATs—strategies a coach might use for guidance [13, 41].

In all conditions, the assigned type of LLM response appeared 5 seconds after the question was shown to the participant, with the text being displayed character by character, mimicking typical chat-based LLM interfaces. In the *LLM Answer* and *LLM Guidance* conditions, the interface informed participants they could freely use the AI suggestions in their answers. Although the RAT has participants employ creative thinking, there is always one definitive answer. As such, we wanted to avoid the case in which a participant

was prompted with the correct answer by AI, acknowledged that it was the correct answer, but then felt as though they should attempt to find a different solution to be "more creative".

During the exposure rounds, participants were made to wait the full 1 minute before advancing to the next task, while in the test rounds, participants could advance once they submitted an answer. This was to discourage participants from speeding through the experiment without giving thought to their answers. Participants were never shown the correct answer to the RATs they completed (except as part of LLM responses, if applicable), nor were they informed if their answers were correct.

4.1.2 Stimulus. For the Remote Associates Test (RAT), we employed a widely-used dataset from Bowden and Jung-Beeman [11]. The subset we selected contained 45 questions equally distributed across three difficulty levels: easy, medium, and hard. For our study, we randomly selected questions from the easy and medium difficulty levels for each round, as pilot testing indicated that these were appropriate for our participant pool (crowdworkers). Figure 14 provides examples of the LLM responses presented to participants in the different experimental conditions.

4.1.3 Verbal Fluency and Surveys. At the start of the experiment, participants completed a verbal fluency task, where they were asked to list as many English words as possible within one minute, starting with a given letter (e.g., 'F'). We implemented this measure to account for participants with low English language proficiency, which can significantly confound performance on the RAT [25, 57]. Following this, participants received instructions for the RAT, along with additional information specific to each experimental condition. After the instructions, participants completed a survey assessing:

- **Self-reported creativity**, measured as a sliding value from 1-100, with the prompt: I am more creative than X% of humans.
- **Attitudes toward AI use in daily life**, measured as a multiple choice question with options; "More concerned than excited", "Equally excited and concerned", and "More excited than concerned".

After finishing the final test round, participants were surveyed again on these same questions, as well as:

- **Perceived test round difficulty**, measured with the prompt "How difficult was it to come up the associated word for the last two (test) tasks?" with options; "Very easy", "Somewhat easy", "Somewhat difficult", and "Very difficult".
- **Perceived helpfulness of exposure rounds** measured with the prompt "How helpful was the exposure phase (first
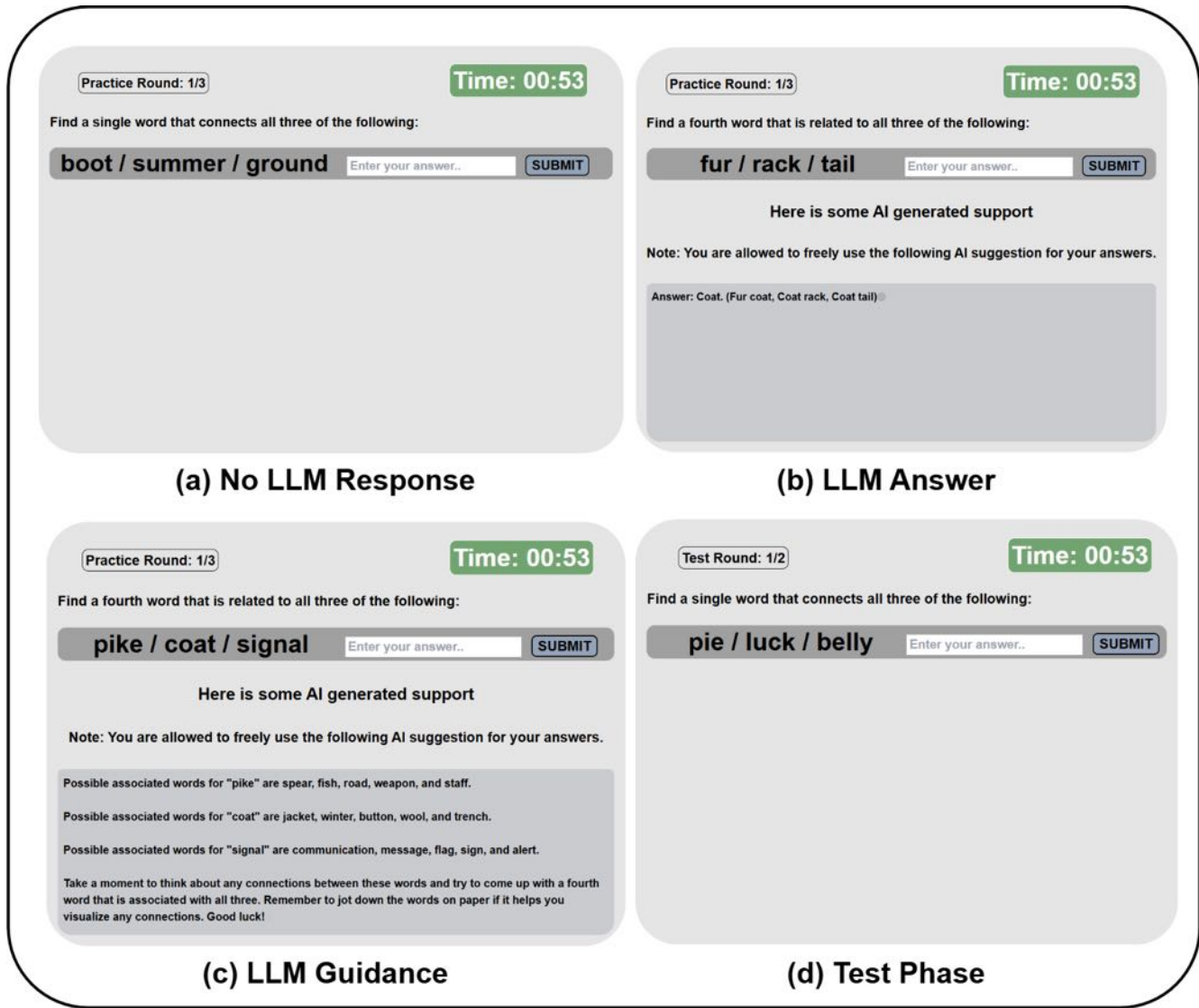
Figure 8: Interface used in the convergent thinking experiment across all 3 exposure round conditions and test rounds.

three questions)?" with options; "Not at all helpful", "A little helpful", and "Very helpful"

The post-experiment survey also contained a basic attention check.

4.1.4 Analysis. Following our pre-registration, we conducted an Analysis of Covariance (ANCOVA) to compare the average accuracy between the three conditions in test rounds, controlling for the number of words generated in the verbal fluency task as a covariate. We used Tukey's Honestly Significant Difference (HSD) test as a post-hoc test for pairwise comparisons between conditions after a significant ANCOVA result.

4.1.5 Participants. We recruited 640 participants from Prolific (of whom 99% passed the attention check). The sample size was determined through a power analysis using simulated data from pilot studies, ensuring 80% power for the main pre-registered test. Participants were based in the US or UK, and were fluent in English. The task took approximately 10 minutes to complete and the participants were paid $1.30. On average, the participants felt they were more creative than 46.4% of the population, and they were able to generate 13.3 words in the verbal fluency task.

4.2 Results

4.2.1 Accuracy in the Test Rounds. As shown in Figure 9, participants in the *LLM Answer* condition did better in the Exposure rounds in comparison to the other conditions. However, this trend changed in the Test rounds. Following our pre-registered analysis plan, the ANCOVA revealed a significant effect of verbal fluency on performance ($F(1, 1256) = 16.715, p < 0.001$) as well as a significant
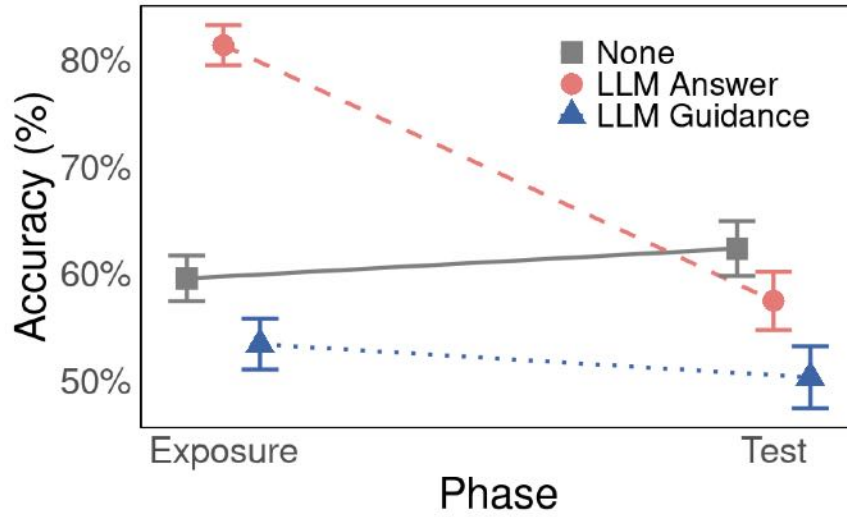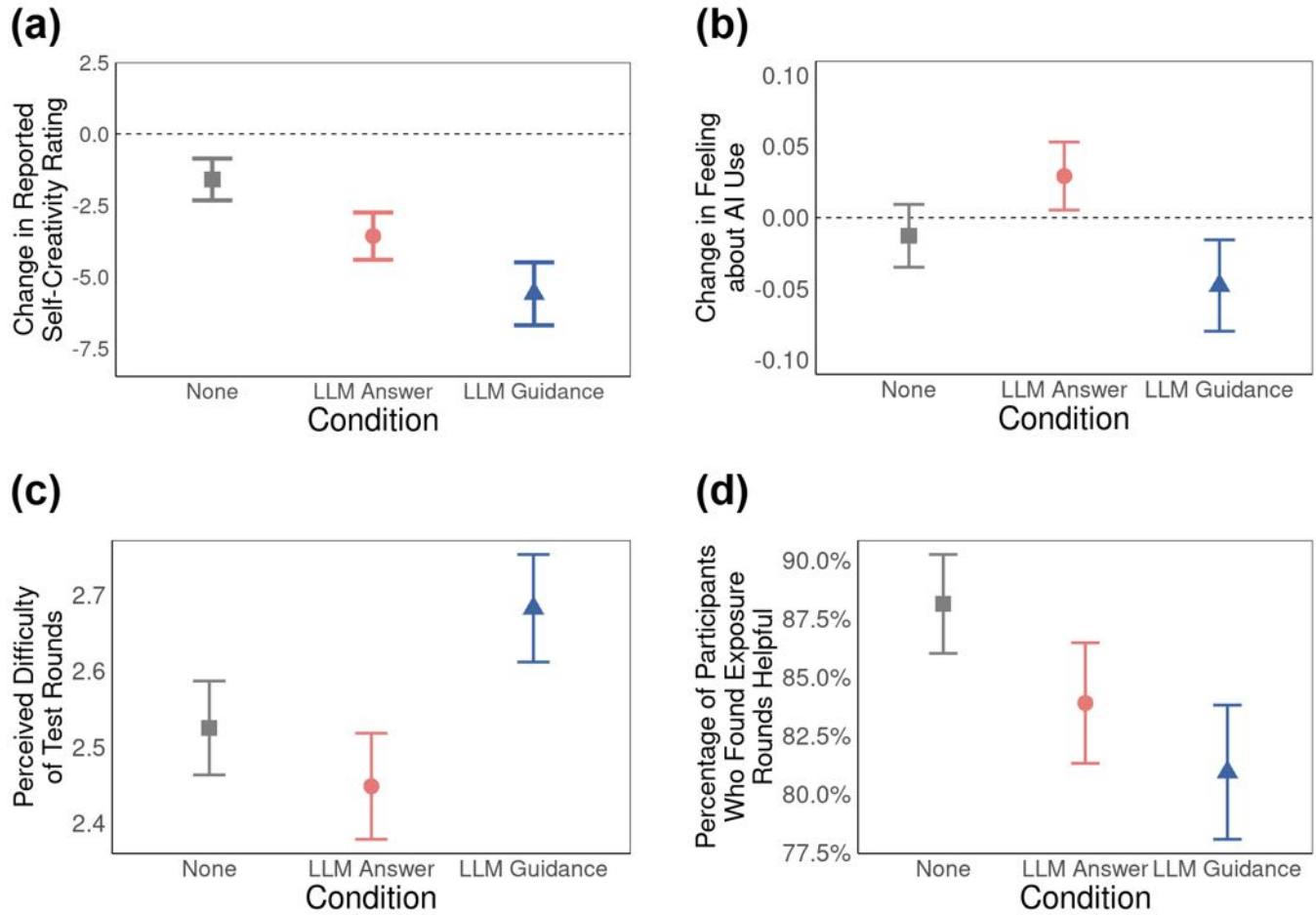
Figure 9: Plot of participant accuracy on the Remote Associates Task, segmented by exposure and test phase across all experimental conditions.

main effect of condition ($F(2, 1256) = 5.134, p = 0.006$). Post-hoc comparisons using Tukey's HSD test indicated that participants in the *LLM Guidance* condition performed significantly worse than those who did not receive any LLM responses in exposure rounds ($p = 0.005$), with a mean difference of -10.6 [95% CI: $-18.6, -2.7$]. While the *LLM Answer* condition did not differ significantly from the *No LLM Assistance* condition ($p = 0.682$), there was a trend suggesting that the *LLM Guidance* condition performed worse than the *LLM Answer* condition ($p = 0.064$).

This suggests that, although the difference between the *LLM Answer* and *LLM Guidance* conditions did not reach statistical significance levels, the *LLM Answer* condition may lie somewhere in between the other two conditions. However, this observation should be interpreted cautiously, given the lack of statistical significance in this comparison.

The positive performance of participants in the *LLM Answer* condition during exposure rounds underscores the effectiveness of LLMs in providing accurate solutions to the task. This suggests that LLMs are indeed proficient at generating correct answers and that participants are capable of recognizing when to leverage LLM advice effectively. Exposure to LLM assistance—whether in the form of direct answers or strategic guidance—did not translate into enhanced unaided performance and may have even been counterproductive. This phenomenon can be understood through the lens of convergent thinking, which often relies on achieving an 'aha' moment or insight. *LLM Guidance* might have disrupted this process by providing too much additional information to take in, thereby hindering the natural cognitive processes necessary for independent problem-solving. Similarly, *LLM Answer* could have impeded participants' engagement in their own creative thinking, as they might have become overly reliant on external solutions rather than developing their own problem-solving strategies.

4.2.2 Subjective Measures and Perceptions. Figure 10 shows participants' change in self-reported levels of creativity, change in feelings towards AI use perceived difficulty of test rounds, and perceived helpfulness of exposure rounds. In **plot (a)**, we see that across all experimental conditions, participants exhibited a reduction in self-reported creativity ratings from pre-experiment to post-experiment surveys. Both *LLM Answer* and *LLM Guidance* conditions saw a greater average reduction in creativity ratings, with *LLM Guidance* showing an average reduction more than twice as substantial as that observed in the *No LLM Assistance* condition. This suggests that reliance on LLMs, particularly in the *LLM Guidance* condition, may have undermined participants' confidence in their own creative abilities. The more substantial reduction in creativity ratings for *LLM Guidance* could indicate that receiving a wide range of possible word connections, rather than necessarily coming up with their own, diminished participants' sense of ownership over their creative process. Futrthermore, the less significant decrease observed in the *LLM Answer* condition could be from participants recognizing they were receiving direct answers which were completely detached from their own creative abilities. **Plot (b)** shows that the change in attitudes toward AI use in daily life stayed roughly constant between pre-experiment to post-experiment surveys. with *LLM Answer* seeing slight attitude increase. Feelings towards AI use was measured using a 3-point Likert scale (with values being "More concerned than excited", "Equally excited and concerned" "More excited than concerned"). This slight increase in positive attitudes toward AI in the *LLM Answer* condition could be attributed to participants recognizing the effectiveness of LLMs in completing tasks accurately, reinforcing their confidence in AI's capabilities. The reliable performance of the LLM may have led to a more favorable view of its potential for everyday use. In **plot (c)**, perceived difficulty of the test rounds was highest for the *LLM Guidance* condition, and slightly lower in the *No LLM Assistance* and *LLM Answer* conditions. That is to say, test rounds felt more difficult when the

Figure 10: Plots of participant accuracy on the Remote Associates Task, segmented by exposure and test phases across all experimental conditions. The top left figure shows participants' change in self perceived creativity ratings (Based on how many % of humans they felt they were more creative than), the top right figure indicates how their feelings towards the increased use of AI computer programs in daily life changed (Between More concerned than excited/More excited than concerned/Equally excited and concerned), the bottom left figure displays how much difficulty they had in coming up with answers for the test phase, and the bottom right figure presents how many participants found the exposure rounds helpful.

participant had completed previous tasks with guiding AI assistance, rather than just a straight answer, or no assistance at all. On average, perceived difficulty was slightly above 2.5 on a 4-point Likert scale (ranging from "very easy" to "very difficult"). meaning that participants found test rounds a little on the difficult side. The higher perceived difficulty in the *LLM Guidance* condition may be due to participants relying on the given word associations during the exposure rounds, which might have made it harder for them to independently generate associations in the test rounds where no assistance was given. Finally, **plot (d)** measured the percentage of participants who found the exposure helpful for their completion of the test rounds. Note that this was not explicitly measuring whether participants found the LLM assistance helpful, but rather whether completing the exposure rounds aided them in performing the test rounds. Participants who received no LLM assistance during the exposure rounds were the most likely to find these rounds helpful.

In comparison, the proportion of participants who found the exposure rounds helpful was roughly 5% lower in the *LLM Answer* condition and 7% lower in the *LLM Guidance* condition. This result may suggest that participants without LLM assistance had to rely more on their own problem-solving strategies during the exposure rounds, which could have enhanced their perceived usefulness of these rounds in preparing for the test phase. In contrast, those receiving LLM assistance might have become more dependent on the AI, leading to a lower perception of the exposure rounds' value for their independent performance during the test rounds.

5 Discussion

In our discussion, we first summarize the key findings of our work and emphasize our contribution as it relates to creativity and the evaluation of AI-based support tools. We then highlight our works'

contributions to the broader implications on fields involving co-creativity with AI. Finally, we discuss the limitations of our study and avenues for future research.

5.1 Key Findings

5.1.1 LLMs Boost Performance During Exposure, but Unassisted Participants Excel in Test. Across both experiments, we observed that LLM assistance was helpful during the exposure phase, aligning with previous research that highlights LLMs' ability to enhance creative task performance when users have access to AI-generated ideas or strategies [2, 14]. However, in the test phase, participants who had no prior exposure to LLMs consistently performed better (this was not statistically significant in all cases). In the divergent thinking task, participants who worked without LLM assistance generated more original ideas on average, and in the convergent thinking task, those without LLM exposure were better able to identify the correct connecting word compared to those who had LLM exposure. These findings suggest that while LLMs may provide short-term boosts in creativity during assisted tasks, they might inadvertently hinder independent creative performance when users are asked to perform without assistance. This raises important questions about the long-term impact of repeated LLM use on human creativity and cognition.

From a design perspective, it is critical to consider not just human-AI performance during exposure phases but also human performance in unassisted tasks after using AI. Systems should be designed with long-term human flourishing in mind, ensuring that the benefits of AI assistance do not come at the cost of diminished independent creative abilities. Ensuring that users can effectively transition from AI-supported creativity to autonomous creative work will help mitigate potential long-term harms associated with over-reliance on LLMs, as highlighted by concerns in the literature about cognitive decline with repeated AI use [26].

5.1.2 Differential Impact of LLMs on Divergent and Convergent Thinking. Our experiments reveal that the effects of LLMs vary significantly depending on the aspect of creativity being measured. In divergent thinking tasks, where participants were asked to generate a wide range of ideas, we observed more skepticism toward LLM assistance. Participants seemed less inclined to adopt AI-generated suggestions, which may be due to the nature of divergent thinking itself—encouraging exploration and unconventional approaches [24, 48]. In contrast, for the convergent thinking task, where participants were tasked with narrowing down the ideas to a single solution, LLM assistance during the exposure phase appeared more beneficial. This is consistent with the idea that convergent thinking is more structured and goal-oriented, making it easier for participants to recognize when LLMs are effectively guiding them toward the correct solution [40].

LLMs, however, may also hinder the creative process. In divergent thinking, the introduction of AI-generated ideas may distract participants, consume valuable cognitive resources, or prevent full engagement with the task. This aligns with theories of creativity such as Boden's concept of *conceptual spaces*, which emphasize the importance of users exploring and navigating creative possibilities independently [9]. Over-reliance on LLMs during divergent thinking may disrupt this exploration, leading to less original ideas, as

seen in the originality results. Similarly, in convergent thinking tasks, LLMs might steer participants toward specific solutions too quickly, reducing the need for deep engagement with the problem space [24].

These findings underscore the importance of carefully calibrating trust and reliance in human-LLM systems. Designers should incorporate measures that help users recognize when to trust and rely on LLMs, and when to prioritize their own cognitive processes. This approach can help balance the benefits of AI support with the risks of undermining the human creative process, ensuring that LLMs enhance rather than detract from creativity in the long term.

5.1.3 Persistent Homogenization of Ideas and the Challenges of Designing LLM Coaches. Existing research has shown that LLM usage can lead to the homogenization of ideas within groups, where participants tend to converge on similar outcomes when using AI-generated suggestions [2]. Our findings extend this concern, showing that even when people stop using LLMs that provide strategic frameworks for thinking, the homogenization effect can persist. In our divergent thinking experiment, participants who received LLM-generated strategies exhibited reduced diversity in their idea sets both during and after LLM use, suggesting that such frameworks may have a lasting impact on creative processes, potentially stifling the generation of more varied or unconventional ideas.

Interestingly, we did not find the same effect for participants who received direct ideas from a standard LLM. The *List of Ideas* condition did not lead to the same lasting homogenization once the LLM was no longer present. This indicates that while LLM-generated strategies can have long-term effects on creative diversity, providing ideas without a guiding framework may allow for more cognitive flexibility once participants stop using AI. In our convergent thinking experiment, participants who received LLM guidance during exposure performed worse in the test round compared to those who received direct answers. This highlights the complexity of designing LLMs as coaches or guides. While direct answers may not always seem ideal, in some cases, they can be more effective than offering frameworks or strategies. Designers of Human-AI systems must take these findings into account, ensuring that LLM interactions are structured to avoid unintentional long-term effects on cognitive diversity. Controlled experiments, such as those conducted here, can be useful methods in refining and optimizing the design of coach-like LLM systems.

5.2 Broader Implications for Fields Involving Human-AI Co-Creativity

When designing AI tools for co-creativity, it is crucial to consider their long-term impact on human cognitive abilities. Hofman *et al.* [27] introduced a useful metaphor—steroids, sneakers, and coach—to describe the spectrum of AI's role in human-AI collaboration. Our findings suggest that co-creative systems must be carefully designed to be coach-like to prevent unintended consequences, such as stifling human creativity, even after AI assistance is removed.

These insights have broad implications for various fields, from scientific discovery to the arts and humanities. In the context of AI

for science, for instance, there is ongoing excitement about building increasingly sophisticated models to accelerate the process of hypothesis generation and experimental execution [10, 38, 52]. However, the design of these models often overlooks their potential impact on scientists' creative abilities. Although some initial explorations have been conducted, the field lacks rigorous empirical evaluation of how AI systems affect human creativity in scientific discovery. The arts and humanities may face similar challenges [44]. For example, writers using the same LLM could produce homogenized content, even when AI is no longer part of their workflow. This underscores the need for AI systems that not only assist in the creative process but also promote long-term cognitive diversity, ensuring that human creativity thrives in collaboration with AI rather than becoming constrained by it.

Our work in a controlled, simple setting suggests that the performance of AI models alone is not enough to determine their value. The design of human-AI interactions—whether AI provides direct answers or encourages users to think critically—and the timing of performance evaluation (during AI use versus after) both significantly alter the narrative. These considerations are critical in ensuring that AI models enhance, rather than diminish, human creative potential. Beyond simply building superhuman AI, we must focus on how these tools influence human creativity, culture, and cognitive growth, aiming for AI systems that enrich and elevate human thought.

5.3 Limitations & Future Work

While our study provides valuable insights into the impact of LLMs on human creativity, there are several limitations that warrant further investigation.

Firstly, measuring creativity itself presents challenges. Though we employed both divergent and convergent thinking tasks (AUT and RAT) to capture a broad spectrum of creative processes, these tasks were limited to verbal responses and conducted over short time periods. Creativity, however, is a much more complex and nuanced phenomenon. Future work should expand the battery of tasks to include more natural, real-world creative activities, such as writing advertisements or solving complex problems. Additionally, creativity tasks that are non-verbal or visual, like those supported by diffusion models, could be explored. Experiments using visual creativity tasks, such as the Test of Creative Thinking Drawing Production (TCT-DP) or the Evaluation of Potential Creativity (EoPC) [39], would provide deeper insights into how LLMs impact creativity in non-text domains.

Our study also focused exclusively on the active, conscious aspects of creativity, largely due to the controlled lab environment. Yet, creativity often involves unconscious processes that are harder to capture in such settings. Another limitation lies in the short exposure period in our lab-based study. In real-world settings, exposure to LLMs is often prolonged, integrated into daily workflows, and thus, likely produces more significant and lasting effects on creativity. In contrast, our study's shorter exposure periods may have resulted in smaller effect sizes, limiting our ability to fully understand the long-term implications of LLM use. Future research should explore prolonged interactions with LLMs to better reflect

real-world scenarios and study potential cumulative impacts on creativity.

Additionally, the design of our LLM-based guidance presents limitations. In the convergent thinking task, the guidance provided was naturally more verbose than direct answers, which may have inadvertently influenced the results. Future studies should control for verbosity and ensure that the nature of guidance is standardized to more accurately assess the impact of LLM "coaching" versus direct assistance. Finally, although this study represents an early attempt to experimentally measure the impact of LLMs on human creativity, the static nature of our LLM interaction may not fully reflect real-world applications. In practical use cases, AI tools often engage in dynamic, interactive exchanges where users can refine inputs, seek clarifications, or adjust outputs. Future research should investigate how conversational and adaptive LLMs influence creativity over time, as these systems more closely resemble how people use AI in real-world creative processes.

6 Conclusion

Through this work, we sought to understand the impact of LLMs on human creativity. We conducted two parallel experiments on divergent and convergent thinking, two key components of creative thinking. Taken together, these experiments shed light on the complex relationship between human creativity and LLM assistance, suggesting that while AI can augment creativity, the mode of assistance matters greatly and can shape long-term creative abilities. In closing, we hope this work offers a template to experimentally evaluate the impact of generative AI on human cognition and creativity.

Acknowledgments

References

[1] Kat Agres, Stephen McGregor, Matthew Purver, and Geraint A Wiggins. 2015. Conceptualizing Creativity: From Distributional Semantics to Conceptual Spaces.. In *ICCC*. 118–125.
[2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. *arXiv preprint arXiv:2402.01536* (2024).
[3] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. 2024. How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment. *arXiv preprint arXiv:2401.13481* (2024).
[4] John Baer. 2014. *Creativity and divergent thinking: A task-specific approach*. Psychology Press.
[5] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. 2024. Generative AI Can Harm Learning. *Available at SSRN* 4895486 (2024).
[6] Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods* 53, 2 (2021), 757–780.
[7] Roger E Beaty, Dan R Johnson, Daniel C Zeitlen, and Boris Forthmann. 2022. Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal* 34, 3 (2022), 245–260.
[8] Margaret A Boden. 2007. Creativity in a nutshell. *Think* 5, 15 (2007), 83–96.
[9] Margaret A Boden. 2010. *Creativity and art: Three roads to surprise*. Oxford University Press.
[10] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature* 624, 7992 (Dec. 2023), 570–578. https://doi.org/10.1038/s41586-023-06792-0

[11] Edward M Bowden and Mark Jung-Beeman. 2003. Normative data for 144 compound remote associate problems. *Behav. Res. Methods Instrum. Comput.* 35, 4 (2003), 634–639.

[12] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. 2023. Machine culture. *Nature Human Behaviour* 7, 11 (2023), 1855–1868.

[13] Ingegerd Carlsson, Per Davidson, and Marianne Ors. 2019. Effects of a daytime nap on primed and repeated remote associates tests and relations with divergent creativity. *Creativity Research Journal* 31, 2 (2019), 207–214.

[14] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–34.

[15] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. 2024. Building Machines that Learn and Think with People. *arXiv preprint arXiv:2408.03943* (2024).

[16] Robert A Cortes, Adam B Weinberger, Richard J Daker, and Adam E Green. 2019. Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences* 27 (2019), 90–93.

[17] Arthur Cropley. 2006. In praise of convergent thinking. *Creativity research journal* 18, 3 (2006), 391–404.

[18] Herie B de Vries and Todd I Lubart. 2019. Scientific creativity: divergent and convergent thinking and the impact of culture. *The Journal of Creative Behavior* 53, 2 (2019), 145–155.

[19] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[20] Berys Gaut. 2003. Creativity and imagination. *The creation of art* (2003), 148–173.

[21] Kenneth J Gilhooly, Evridiki Fioratou, Susan H Anthony, and Victor Wynn. 2007. Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology* 98, 4 (2007), 611–625.

[22] Ken J Gilhooly, George Georgiou, and Ultan Devery. 2013. Incubation and creativity: Do something different. *Thinking & Reasoning* 19, 2 (2013), 137–149.

[23] Gabriela Goldschmidt. 2016. Linkographic evidence for concurrent divergent and convergent thinking in creative design. *Creativity research journal* 28, 2 (2016), 115–122.

[24] Joy P Guilford. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior* 1, 1 (1967), 3–14.

[25] Trey Hedden, Gary Lautenschlager, and Denise C. Park. 2005. Contributions of Processing Ability and Knowledge to Verbal Memory Tasks across the Adult Life-Span. *The Quarterly Journal of Experimental Psychology Section A* 58, 1 (2005), 169–190.

[26] Richard Heersmink. 2024. Use of large language models might affect our cognitive skills. *Nature Human Behaviour* 8, 5 (2024), 805–806.

[27] Jake M Hofman, Daniel G Goldstein, and David M Rothschild. 2023. Steroids, Sneakers, Coach: The Spectrum of Human-AI Relationships. *Available at SSRN 4578180* (2023).

[28] Bernhard Hommel, Lorenza S Colzato, Rico Fischer, and Ingrid K Christoffels. 2011. Bilingualism and creativity: Benefits in convergent thinking come with losses in divergent thinking. *Frontiers in psychology* 2 (2011), 273.

[29] Stacy Hsueh, Marianela Ciolfi Felice, Sarah Fdili Alaoui, and Wendy E Mackay. 2024. What Counts as 'Creative' Work? Articulating Four Epistemic Positions in Creativity-Oriented HCI Research. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[30] Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14, 1 (2024), 3440.

[31] Saskia Jaarsveld, Thomas Lachmann, and Cees Van Leeuwen. 2012. Creative reasoning across developmental levels: Convergence and divergence in problem creation. *Intelligence* 40, 2 (2012), 172–188.

[32] Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports* 13, 1 (2023), 13601.

[33] Wilma Koutstaal, Kara Kedrick, and Joshua Gonzalez-Brito. 2022. Capturing, clarifying, and consolidating the curiosity-creativity connection. *Scientific Reports* 12, 1 (2022).

[34] Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Math Education with Large Language Models: Peril or Promise? *Available at SSRN 4641653* (2023).

[35] Christine S Lee, Anne Corinne Huggins, and David J Therriault. 2014. A measure of creativity or intelligence? Examining internal and external structure validity evidence of the Remote Associates Test. *Psychology of Aesthetics, Creativity, and the Arts* 8, 4 (2014), 446.

[36] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[37] Paisley Livingston. 2009. Chapter seven. Poincaré's 'Delicate Sieve': On creativity and constraints in the arts. In *The idea of creativity*. Brill, 127–146.

[38] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292 [cs.AI] https://arxiv.org/abs/2408.06292

[39] Todd Lubart. 2016. CREATIVITY AND CONVERGENT THINKING: REFLECTIONS, CONNECTIONS AND PRACTICAL CONSIDERATIONS. *RUDN Journal of Psychology and Pedagogics* 4 (2016), 7–15.

[40] Martha T Mednick and Sharon Halpern. 1968. Remote associates test. *Psychological Review* (1968).

[41] Martha T Mednick, Sarnoff A Mednick, and Edward V Mednick. 1964. Incubation of creative performance and specific associative priming. *The Journal of Abnormal and Social Psychology* 69, 1 (1964), 84.

[42] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* (2023), 101356.

[43] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1051, 19 pages. https://doi.org/10.1145/3613904.3642105

[44] Nitin Rane and Saurabh Choudhary. 2024. Role and challenges of ChatGPT and similar generative artificial intelligence in arts and humanities. *Studies in Humanities and Education* 5, 1 (2024), 1–11.

[45] Olga M Razumnikova. 2020. Divergent versus convergent thinking. In *Encyclopedia of creativity, invention, innovation and entrepreneurship*. Springer, 759–765.

[46] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[47] Melissa Roemmele. 2021. Inspiration through observation: Demonstrating the influence of automatically generated text on creative writing. *arXiv preprint arXiv:2107.04007* (2021).

[48] Mark A Runco. 1991. *Divergent thinking*. Ablex Publishing Corporation Norwood, NJ.

[49] Mark A Runco et al. 2010. Divergent thinking, creativity, and ideation. *The Cambridge handbook of creativity* 413 (2010), 446.

[50] Mark A Runco and Selcuk Acar. 2012. Divergent thinking as an indicator of creative potential. *Creativity research journal* 24, 1 (2012), 66–75.

[51] Olivier Serrat and Olivier Serrat. 2017. The SCAMPER technique. *Knowledge solutions: tools, methods, and approaches to drive organizational performance* (2017), 311–314.

[52] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109 [cs.CL] https://arxiv.org/abs/2409.04109

[53] Dean Keith Simonton. 2015. On praising convergent thinking: Creativity as blind variation and selective retention. *Creativity Research Journal* 27, 3 (2015), 262–270.

[54] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744* (2023).

[55] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[56] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[57] Xinyi Wang. 2016. *The Effect of Language Ability in Creativity Assessment.* Ph. D. Dissertation.

[58] Ching-Lin Wu, Shih-Yuan Huang, Pei-Zhen Chen, and Hsueh-Chih Chen. 2020. A systematic review of creativity-related studies applying the remote associates test from 2000 to 2019. *Frontiers in psychology* 11 (2020), 573432.

[59] Weitao Zhang, Zsuzsika Sjoerds, and Bernhard Hommel. 2020. Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage* 210 (2020), 116572.

[60] Weili Zhu, Siyuan Shang, Weili Jiang, Meng Pei, and Yanjie Su. 2019. Convergent thinking moderates the relationship between divergent thinking and scientific creativity. *Creativity Research Journal* 31, 3 (2019), 320–328.

A Experiment 1 (Divergent Thinking)

A.1 LLM Configuration

A.1.1 LLM Model Specification.

- **model version**: gpt-4o
- **date of use**: August 2024

Configuration Settings:

- **temperature**: 0
- **max response**: 4000
- **top-p**: 0.95
- **frequency penalty**: 0
- **presence penalty**: 0

A.1.2 Prompt Design

"What are some creative uses for a [OBJECT]? The goal is to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different. List creative uses for a [OBJECT]." .

A.1.3 System Prompt. "List of Strategies" condition

You are an AI assistant tasked with the specific goal of teaching users to be more creative by applying the SCAMPER technique. SCAMPER is a powerful tool for creative thinking that involves exploring various ways to modify an object or idea to generate new possibilities. Your interactions should guide users through the SCAMPER process, which includes the following steps:

- **Substitute:** Encourage users to think about what elements of the object could be replaced with something else to improve or give it a new use.
- **Combine:** Prompt users to consider how they might combine the object with other items or ideas to create something new.
- **Adapt:** Ask users how the object could be adapted or modified to serve another purpose, or how it might be inspired by something else.
- **Modify:** Challenge users to think about how altering the shape, size, color, or any other attribute of the object could create new uses or improvements.
- **Put to another use:** Encourage users to brainstorm completely different uses for the object, beyond its original purpose.
- **Eliminate:** Guide users to consider what might happen if parts of the object were removed or simplified. How does this change its potential uses?
- **Reverse:** Ask users to consider what would happen if they reversed the object's purpose or use. Would it function differently or serve a new role?

In your guidance, encourage users to iterate through these steps, not necessarily in order, to explore the full range of creative possibilities. Don't mention SCAMPER explicitly in your response. Output a list that the user can follow. Don't include any examples, but only prompts or questions to help users ideate. Make your response concise and easy to follow for the users. Express each point in as few words as possible.

Figure 11: The system prompt for the "List of Strategies" condition (not visible to participants).

A.2 Sample LLM Responses

Figure 12: Sample LLM responses for the exposure rounds.

B Experiment 2 (Convergent Thinking)

B.1 LLM Configuration

B.1.1 LLM Model Specification.

- **model version**: gpt-4o
- **date of use**: August 2024

Configuration Settings:

- **temperature**: 0
- **max response**: 800
- **top-p**: 0.95
- **frequency penalty**: 0
- **presence penalty**: 0

B.1.2 Prompt Design.
"In this task, You will be shown sets of three words at a time. For each set of three words, you will be asked to generate a fourth word that connects or fits with all three words. Specifically, the word that you generate, when combined with each of the three stimulus words, should create three word-pairs, each of which makes a common compound word or phrase. Please try to be creative and appropriate in your responses.
Find a fourth word that is related to all three of the following: [word1], [word2], [word3]"

You are a supportive coach guiding participants through solving Remote Associates Test (RAT) problems. For each of the three given words, provide a concise response by listing five possible associated words. Encourage the participant to identify connections between these words to find the fourth word, but do not directly give the solution. Conclude by suggesting they consider if any word connects all three and nudge them to jot the words down on paper if it might help.

Figure 13: The system prompt for the *LLM Guidance* condition (not visible to participants).

B.2 Sample LLM Responses

Remote Associates Task	LLM Answer	LLM Guidance
flake, mobile, cone (snow)	Answer: Snow. (Snowflake, Snowmobile, Snowcone)	Possible associated words for "flake" are snow, cereal, skin, coconut, and pastry. Possible associated words for "mobile" are phone, sculpture, home, device, and car. Possible associated words for "cone" are ice cream, volcano, traffic, pine, and speaker. Now, try to identify any connections between these words and see if you can come up with a fourth word that is associated with all three. It might be helpful to jot down the words on paper to visually see any connections. Good luck!
dream, light, break (day)	Answer: Day. (Daydream, Daybreak, Daylight)	Possible associated words for "dream" are sleep, subconscious, nightmare, fantasy, and aspiration. Possible associated words for "break" are fracture, pause, vacation, interruption, and opportunity. Possible associated words for "light" are lamp, bright, illuminate, candle, and photon. Take a moment to think about any connections between these words and try to come up with a fourth word that is associated with all three. Remember to jot down the words on paper if it helps you visualize any connections. Keep up the good work!
aid, rubber, wagon (band)	Answer: Band. (Bandaid, Rubber Band, Bandwagon)	Possible associated words for "aid" are: help, support, assist, rescue, first. Possible associated words for "wagon" are: cart, horse, transport, carry, goods. Possible associated words for "rubber" are: elastic, tire, eraser, latex, ball. Now, can you identify any connections between these words? Is there a fourth word that could connect them? Remember, it might be helpful to jot down the words on paper and see if any connections jump out at you.

Figure 14: Sample LLM responses for the exposure rounds.