

1
2
3
4
5
6
7

Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking

8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

ANONYMOUS AUTHOR(S)*

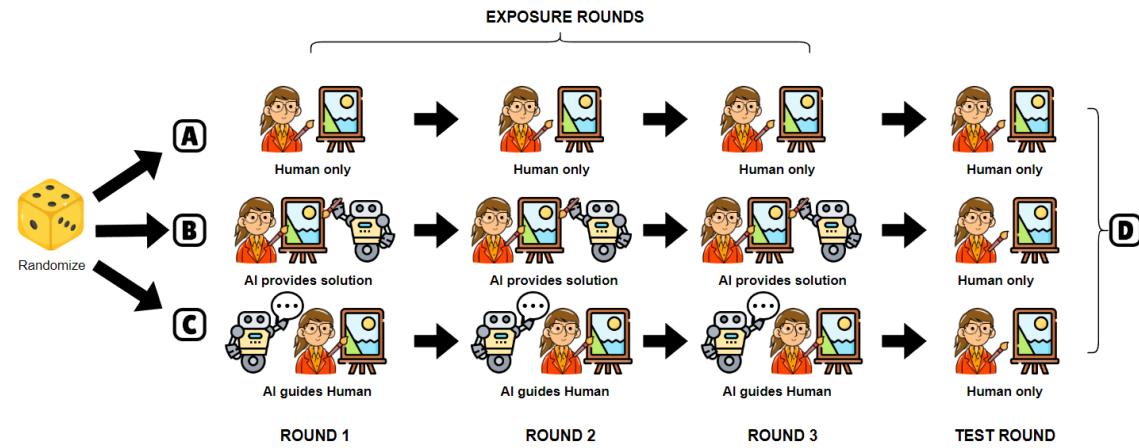


Fig. 1. Experimental framework for measuring the impact of AI use on Human creativity. Participants engage in a series of *Exposure* rounds where they are randomized to either receive - (A) No assistance, (B) LLM solution (standard): This could be analogous to using a chat LLM such as ChatGPT for the 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.

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.

Additional Key Words and Phrases: creativity, divergent thinking, convergent thinking, large language models, experiments

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

53 ACM Reference Format:

54 Anonymous Author(s). 2018. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking.
 55 In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New
 56 York, NY, USA, 24 pages. <https://doi.org/XXXXXXX.XXXXXXX>
 57

58 1 Introduction

59 The rise of generative artificial intelligence tools such as ChatGPT has the potential to upend the creative process.
 60 By their nature as *generative* systems, they offer an unprecedented capacity to algorithmically generate ideas, and
 61 perhaps even to create. State-of-the-art generative AI systems have reached proficiency levels that not only match
 62 human creativity in certain evaluations [? ?], but can also enhance the creative output of knowledge workers [? ? ?].
 63 We are living through a time of creative transformation, with AI visual art, AI music, and AI-enhanced videos rapidly
 64 proliferating.
 65

66 But what does the era of generative AI hold for human creativity? When AI assistance becomes a regular part of
 67 creative processes, will human creativity change? There are natural concerns that reliance on generative AI might
 68 impair an individual's inherent ability to think creatively without assistance. Further, there is preliminary evidence that
 69 widespread dependence on similar generative AI tools could lead to a homogenization of thought. This could in turn
 70 reduce the diversity that drives collective innovation and stifle breakthroughs across various fields. However, there is
 71 also reason to believe that AI assistance could spark human creativity. Working with a fresh creative partner might
 72 be similarly stimulating regardless of whether the partner is human or machine. If this is the case, AI could prove
 73 to be a force for the flourishing of human creativity. Understanding how and whether co-creating with AI affects an
 74 individual's ability to generate creative ideas independently is therefore of critical importance.
 75

76 This is a complex challenge. Creativity is inherently subjective and difficult to measure, making it hard to assess
 77 changes in an individual's creative ability. Factors like prior knowledge, experience, and environmental context all
 78 interplay with creativity, making it hard to isolate the effects of AI usage "in the wild". The rapid evolution of AI
 79 technologies further complicates matters, as their capabilities and effects are constantly changing. Without clear
 80 methods to evaluate these impacts, we cannot fully grasp how AI integration might alter the creative landscape.
 81

82 In this work, we investigate the impact of Large Language Models (LLMs) assistance on human creativity by
 83 examining the two fundamental components of creativity: divergent and convergent thinking. Divergent thinking
 84 involves the generation of multiple, unique ideas, fostering exploration and innovation [? ?]. Convergent thinking, in
 85 contrast, focuses on refining these ideas to select the most effective solutions [? ?]. We investigate how different forms
 86 of LLM assistance influence these cognitive processes by comparing two types of LLMs—a standard LLM that provides
 87 direct answers out of the box, and a coach-like LLM that offers guidance and prompts to stimulate thinking—in contrast
 88 to a control condition with no LLM assistance.
 89

90 Specifically, we address the following research questions:
 91

92 **RQ1** How do standard LLM assistance and coach-like LLM guidance, compared to no assistance, affect an individual's
 93 divergent thinking abilities when generating creative ideas independently?
 94

95 **RQ2** What are the impacts of standard LLM assistance and coach-like LLM guidance, versus no assistance, on an
 96 individual's convergent thinking skills in independently refining and selecting ideas?
 97

98 To answer these questions, we designed and conducted two pre-registered parallel experiments with 1,100 participants
 99 that assess how these forms of LLM assistance influence unassisted human creativity, compared to a control group with
 100 no AI assistance. Figure 1 illustrates the high-level design of the experiments. In both experiments, participants were
 101 Manuscript submitted to ACM
 102
 103
 104

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 [? ? ? ? ?]. **Divergent thinking** involves generating a wide range of ideas, exploring multiple possibilities, and embracing unconventional approaches [? ? ?]. 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 [? ? ?]. 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 [? ?]. Convergent thinking becomes crucial in the *Insight* and *Revision* phases, refining ideas into viable solutions [? ?]. This model underscores that creativity involves not only generating numerous ideas, but also selecting and refining them [? ?].

Boden's theory further introduces measures of creativity, distinguishing between **P-creativity** (psychological creativity) and **H-creativity** (historical creativity) [? ?]. P-creativity refers to ideas novel to the individual, while

¹⁵⁷ H-creativity refers to ideas novel within the broader context of human knowledge. This distinction allows us to assess
¹⁵⁸ creativity on both a personal and historical scale. Divergent thinking can be hypothesized to contribute primarily to
¹⁵⁹ P-creativity, where the generation of new ideas is the key [? ? ?]. In contrast, convergent thinking is essential for
¹⁶⁰ advancing these ideas toward H-creativity, where their novelty and value must be recognized within a broader context.
¹⁶¹

¹⁶² Our experiments explore how LLMs influence these aspects of creativity. To investigate divergent thinking, we
¹⁶³ utilize the Alternate Uses Test (AUT), where participants are prompted to generate as many creative uses as possible
¹⁶⁴ for a common object [? ?]. For convergent thinking, we use the Remote Associates Test (RAT), where participants
¹⁶⁵ find a single word that connects three seemingly unrelated words, assessing their ability to converge on a correct and
¹⁶⁶ meaningful solution [? ? ?].
¹⁶⁷

¹⁶⁹ 2.2 LLMs as Tools for Creativity

¹⁷¹ Large Language Models (LLMs) have demonstrated notable creative performance, often performing as well as, or
¹⁷² even better than, average humans in various creative tasks. However, they still fall short when compared to the best
¹⁷³ human performers, particularly experts [? ?]. For instance, Chakrabarty *et al.* [?] found that while LLMs performed
¹⁷⁴ competently in creative writing tasks, professionals consistently outperformed them. Anderson *et al.* [?] observed that
¹⁷⁵ ChatGPT users generated more ideas than those in a control condition. However, these ideas tended to be homogenized
¹⁷⁶ across different users, suggesting a limitation in the diversity of creative output when LLMs are used. These findings
¹⁷⁷ are particularly relevant when LLMs are employed to perform entire tasks on behalf of users.
¹⁷⁸

¹⁷⁹ Beyond performing tasks autonomously, LLMs can also be utilized to guide users through the *conceptual spaces*
¹⁸⁰ of creative thinking, serving as tools for ‘creativity support’ [? ? ?]. This approach positions LLMs not merely as
¹⁸¹ replacements for human creativity but as enhancers of the creative process, helping users navigate and explore creative
¹⁸² possibilities more effectively. The literature on Human-AI collaboration in creative tasks is rapidly growing. Lee *et al.*
¹⁸³ [?] introduced a dataset for analyzing GPT-3’s use in creative and argumentative writing, suggesting that the HCI
¹⁸⁴ community could foster more detailed examinations of LLMs’ generative capabilities through such datasets. Suh *et al.* [?]
¹⁸⁵ [?] highlighted that the current interaction paradigm with LLMs tends to converge on a limited set of ideas, potentially
¹⁸⁶ stifling creative exploration. They proposed frameworks that facilitate the exploration of structured design spaces,
¹⁸⁷ allowing users to generate, evaluate, and synthesize many responses.
¹⁸⁸

¹⁸⁹ A key question remains: what happens to human creativity, particularly cognition, when humans repeatedly use
¹⁹⁰ LLMs for creative tasks—either directly to generate ideas or solutions, or as guides through the creative process? While
¹⁹¹ there are concerns that this could lead to a deterioration of creative abilities, there is also optimism that these tools, if
¹⁹² properly designed, could enhance human creativity while providing momentary assistance [? ?]. However, empirical
¹⁹³ evidence to answer this question definitively is still lacking.
¹⁹⁴

¹⁹⁵ 2.3 Generative AI and Human Cognition

¹⁹⁶ The continued use of Generative AI is significantly impacting our society, particularly in how *culture* is created and
¹⁹⁷ propagated [? ?]. These technologies are reshaping the production of cultural artifacts and the means through which
¹⁹⁸ culture is disseminated and experienced. The influence of Generative AI extends to our cognitive abilities, with effects
¹⁹⁹ that may be polarizing [?]. On the one hand, the pervasive use of AI tools might lead to a massive homogenization of
²⁰⁰ creative output, a trend that could persist even after humans stop using these tools [?]. This raises concerns about the
²⁰¹ potential stifling of creativity and the reduction of diversity in thought and expression. On the other hand, Generative
²⁰²

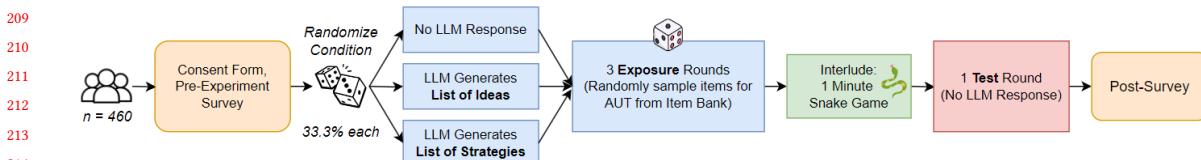


Fig. 2. Schematic of design for Experiment 1 on divergent thinking.

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.* [?] 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.* [?] 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., [? ?]), web search (for e.g., [?]), 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 [? ?]. 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 [?]. To understand these dynamics, our first pre-registered¹ 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 [?]. 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.

¹https://aspredicted.org/4X6_6X6

261 Your ideas don't have to be practical or realistic; they can be silly or strange, even, so long as they are CREATIVE uses
 262 rather than ordinary uses...". For instance, an alternate use of pants might be as a makeshift flag.
 263

264 *3.1.1 Experimental conditions and treatment.* Our experimental design involved two main phases: an exposure (paired)
 265 phase during which participants attempted the task with an LLM partner, and a test (solo) phase during which
 266 participants attempted the task on their own. Participants were given two-minute time frames per object to submit
 267 their ideas one by one, and could freely edit or delete any previously submitted ideas. During the exposure phase,
 268 participants are introduced to three objects, one after the other. Every participant was randomly assigned to receive
 269 one of three types of LLM responses:
 270

271 *None:* Control group. No LLM support provided.

272 *List of Ideas:* GPT-4o generated a list of alternate uses for the given object. Seven randomly
 273 sampled uses were shown to the participant, which they could freely use in
 274 their responses. In other words, an LLM attempted the task and shared its
 275 responses with the participant.

276 *List of Strategies:* GPT-4o, using a specialized system prompt (shown in Figure 11), generated
 277 seven strategies based on the SCAMPER technique [?]. In this condition, an
 278 LLM provided guidance to the participant but refrained from sharing explicit
 279 answers to the task.

280 The assigned type of LLM response appeared 5 seconds after the item was shown to the participant, and appeared
 281 character by character, similar to other chat LLM interfaces. Figure 2 shows the schema of the experiment design.
 282 Following the exposure phase, participants engaged in a brief distractor task to simulate forgetting, playing a game of
 283 Snake for one minute. In this subsequent phase, participants were assigned the Alternate Uses Task for a new object
 284 selected at random, this time without LLM support.

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

289
 290 *3.1.2 Stimuli.* The items in each round were randomly sampled from five items: tire, pants, shoe, table, bottle. We chose
 291 these five particular objects as the originality scoring measure had the highest correlation with human judgements for
 292 these five objects [?]. Figure 3 shows the different responses shown to participants in different conditions.
 293

294
 295 *3.1.3 Dependent Variables.* AUT allows us to measure different dimensions of divergent thinking. As such, using LLMs
 296 may impact each of these dimensions differently. These dimensions include:
 297

- 298 • **Originality (How original the idea is):** We measure the originality of each AUT idea with an existing
 299 fine-tuned GPT-3 classifier [?]. The model was fine-tuned with human judgments of AUT originality (where
 300 human raters judged, on a scale of 1 to 5, the originality of an idea given the object) and achieved an $r = 0.81$
 301 overall correlation with human judgments. For our experiment, we chose the five items that had the highest
 302 accuracy for the model ($r > 0.88$).
 303
- 304 • **Fluency (How many ideas):** We measure fluency by counting the number of ideas the participant generated
 305 in a given round.
 306

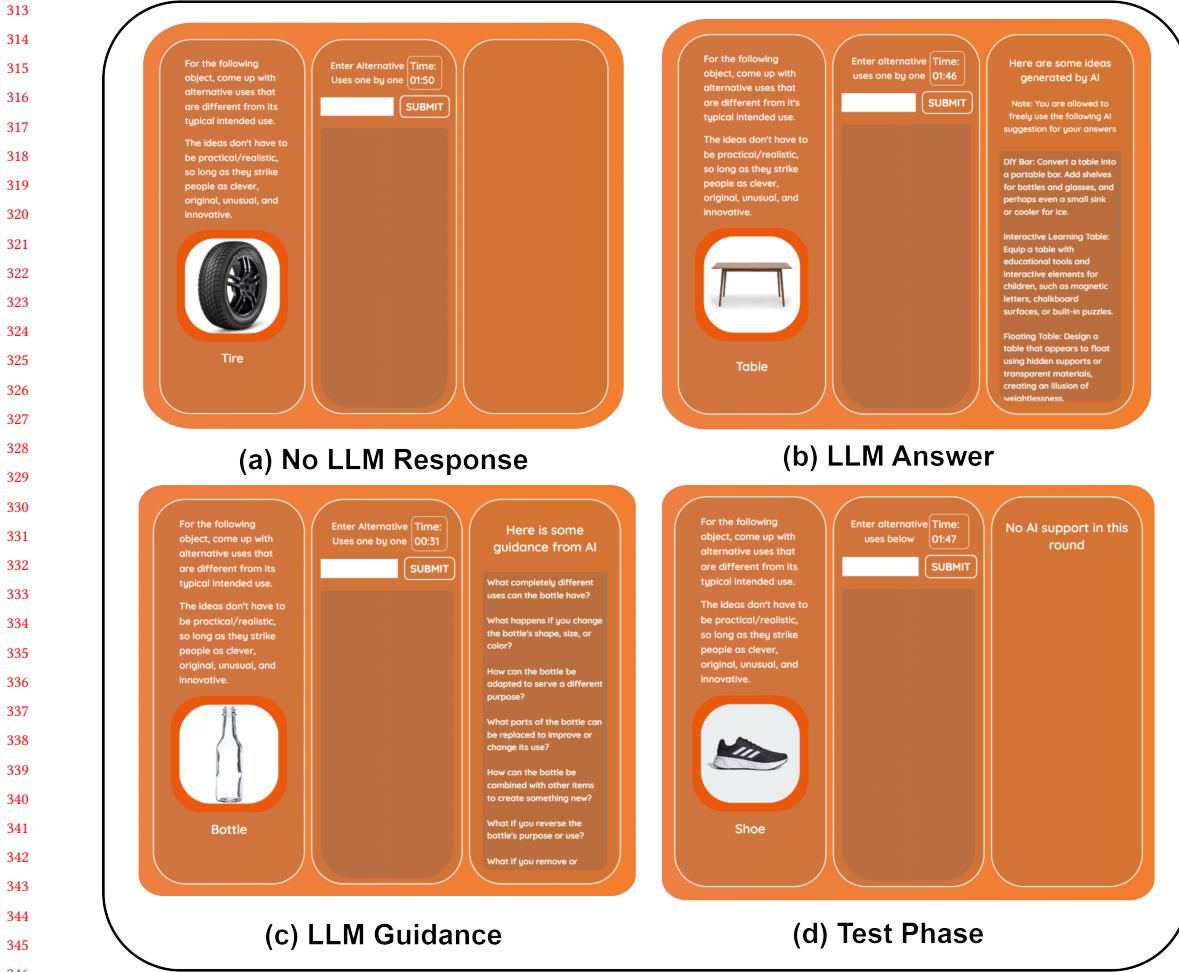


Fig. 3. Interface used in the divergent thinking experiment across all 3 exposure round conditions and test rounds.

- 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 [?]. 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 [?].
- 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

365 and all the ideas in the Exposure set. We use the maximum rather than a measure of central tendency because if
 366 a participant is inspired by an idea, it would likely be a single idea [?].
 367

- 368 • **Group-level diversity of ideas (How different are the ideas generated by participants across the group
 369 for a given phase):** The goal of this analysis was to simulate how individuals contribute to idea generation as
 370 a group, with a focus on understanding the potential homogenization of ideas at the group level. Specifically,
 371 we aimed to study whether LLM assistance leads to a convergence of ideas, reducing diversity within a group.
 372 To achieve this, we performed 150 Monte Carlo runs, where we sampled 7 ideas per round, simulating group
 373 contribution dynamics. We calculated the median pairwise cosine distances between the embeddings of these
 374 ideas, using SBERT to encode them. This sampling process reflects the number of participants in each condition.
 375 The median pairwise cosine distance across each Monte Carlo run was used to evaluate diversity for each phase
 376 and condition. This approach allowed us to assess whether different forms of LLM assistance (e.g., providing
 377 direct ideas or acting as a coach) promote homogenization or help maintain diversity of ideas at the group level,
 378 even after stopping to use LLMs.
 379

380
 381
 382 • **3.1.4 Analysis.** We follow a long tradition of scoring responses to the AUT computationally [? ?]. Following our
 383 pre-registration, we conducted a Kruskal-Wallis H test to compare the distribution of originality across the three
 384 conditions: None, List of Ideas, and List of Strategies, using a significance level of 0.05. In the event of a significant
 385 Kruskal-Wallis result, Dunn's test with Bonferroni correction was applied as a post-hoc test for pairwise comparisons
 386 between the conditions.
 387

388 This same analysis procedure was applied for other dependent variables, including individual-level and group-level
 389 diversity, fluency, and creative flexibility. We report the significance of the overall Kruskal-Wallis H test and the results
 390 of Dunn's test for specific pairwise comparisons for each measure.
 391

392
 393 • **3.1.5 Participants.** We recruited 460 participants from Prolific. Based on a power analysis using simulated and pilot
 394 data, we determined that this sample size will be necessary to achieve 70% power with a moderate effect size and a
 395 significance level of 0.05. The overall experiment took around 12 minutes to complete and participants were paid \$1.57.
 396 Participants were based in the US or UK, and fluent in English. On average, participants felt they were more creative
 397 than 48.12% of the population, at the start of the experiment.
 398

400 3.2 Results

401 We report on the analysis of 9,457 ideas generated by participants across all conditions and phases.
 402

403
 404 • **3.2.1 Originality.** Figure 4a shows the average originality of ideas across conditions. In the *Exposure* phase, the mean
 405 originality was similar across conditions, as indicated by the Kruskal-Wallis H test ($H(2) = 3.78, p = 0.151$). Interestingly,
 406 participants who received the *List of Ideas* performed slightly worse than other conditions, despite being exposed to
 407 LLM ideas with a mean originality of 3.25 (± 0.1). This is nearly 0.5 points higher than the average originality of ideas
 408 generated by participants in this condition (on a scale of 1 to 5). This discrepancy suggests that participants may struggle
 409 to accurately assess the quality of AI-generated ideas, or perhaps, when provided with AI ideas, they may prioritize
 410 coming up with their own, potentially lower-quality ideas. These findings highlight the need for designing effective
 411 reliance mechanisms that help users fully leverage the benefits of AI-generated ideas, beyond simply improving AI's
 412 output quality.
 413

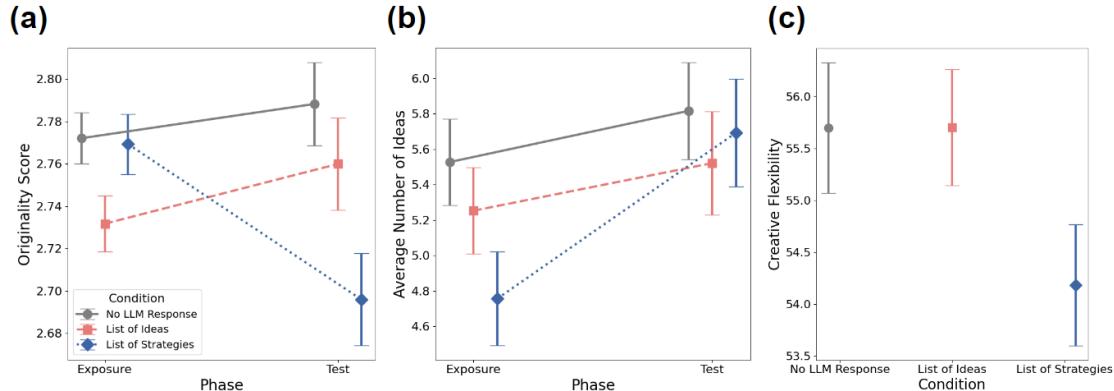


Fig. 4. Plots of the Alternate Uses Task ideas for the various divergent thinking dimensions (Segmented by phase and/or LLM response type). The left figure shows idea originality scores, the middle figure indicates idea fluency, and the right figure presents participant creative flexibility.

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.

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

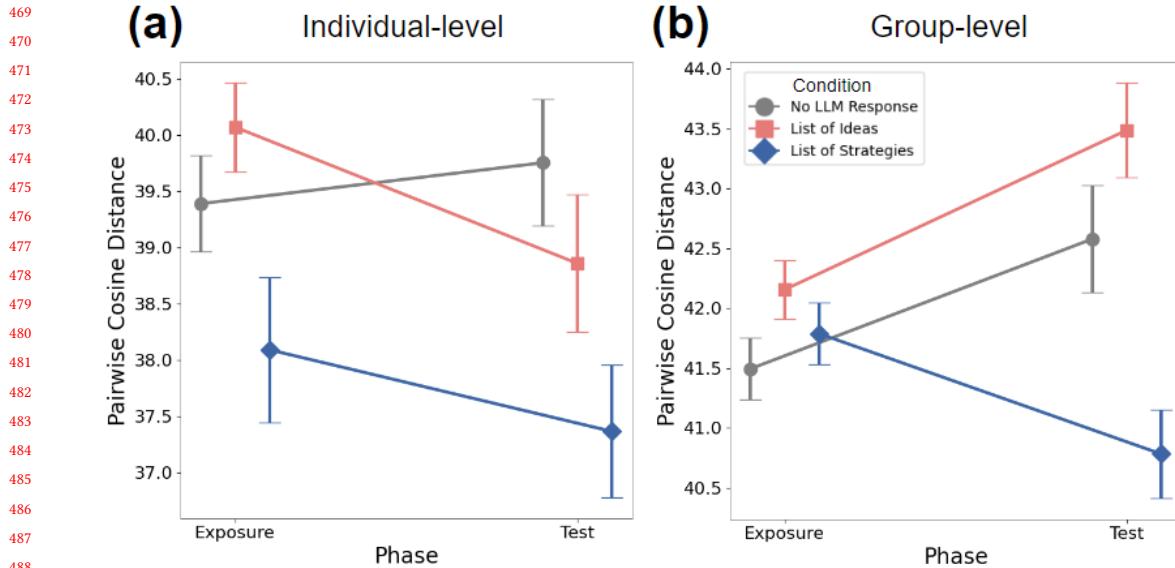


Fig. 5. Plots of the average individual- and group-level median diversity, segmented by experiment conditions and phases. Higher values denote more difference between ideas. Error bars represent \pm one standard error of mean.

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 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.

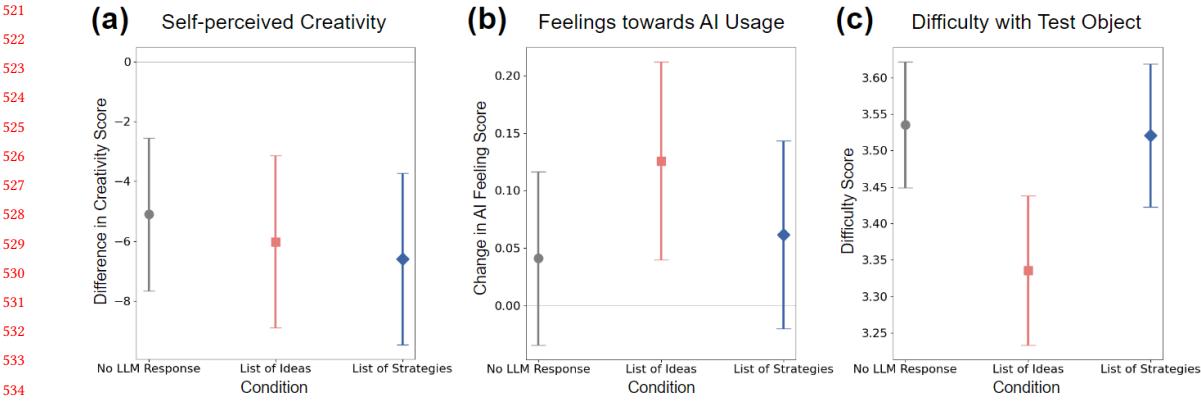


Fig. 6. Plots of subjective measures collected before and after participants completed the Alternate Uses Task. The left figure shows participants' change in self perceived creativity levels (Based on how many % of humans they felt they were more creative than), the middle figure indicates how their 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.

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, 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.

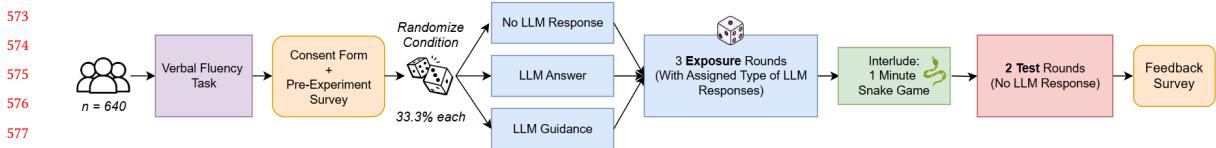


Fig. 7. Schematic of design for Experiment 2 on convergent thinking.

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-registered² 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' (*bookshelf*, *logbook*, *bookworm*). We selected this task due to its widespread use in prior research [? ?], 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 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 impacted 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.

²https://aspredicted.org/V8K_ZTV

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

628 The *No LLM Response* (Figure 8a) condition serves as a baseline, reflecting the state of convergent thinking without
629 AI assistance. The *LLM Answer* (Figure 8b) condition represents the scenario in which AI provides ready-made solutions.
630 The *LLM Guidance* (Figure 8c) condition was designed to assist participants in navigating the conceptual space of
631 the problem without directly revealing the answer. This approach aims to enable participants to solve the problem
632 independently while still offering useful guidance in the moment. Research has shown that priming with related words
633 and manually writing responses can enhance performance on RATs—strategies a coach might use for guidance [? ?].
634

635 In all conditions, the assigned type of LLM response appeared 5 seconds after the question was shown to the
636 participant, with the text being displayed character by character, mimicking typical chat-based LLM interfaces. In the
637 *LLM Answer* and *LLM Guidance* conditions, the interface informed participants they could freely use the AI suggestions
638 in their answers. Although the RAT has participants employ creative thinking, there is always one definitive answer. As
639 such, we wanted to avoid the case in which a participant was prompted with the correct answer by AI, acknowledged
640 that it was the correct answer, but then felt as though they should attempt to find a different solution to be “more
641 creative”.

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

647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
4.1.2 Stimulus. For the Remote Associates Test (RAT), we employed a widely-used dataset from Bowden and Jung-Beeman [?]. 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.

660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
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 [? ?]. 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:

677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728

(a) No LLM Response

(b) LLM Answer

(c) LLM Guidance

(d) Test Phase

Fig. 8. Interface used in the convergent thinking experiment across all 3 exposure round conditions and test rounds.

- **Perceived test round difficulty**, measured with the prompt “How difficult was it to come up with 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 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.

Manuscript submitted to ACM

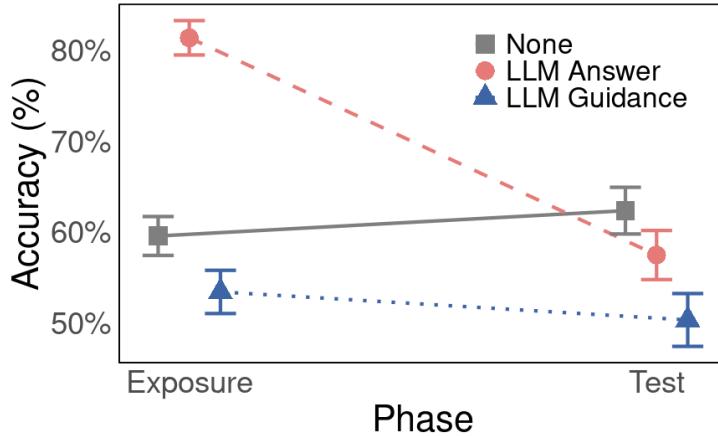


Fig. 9. Plot of participant accuracy on the Remote Associates Task, segmented by exposure and test phase across all experimental conditions.

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 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*

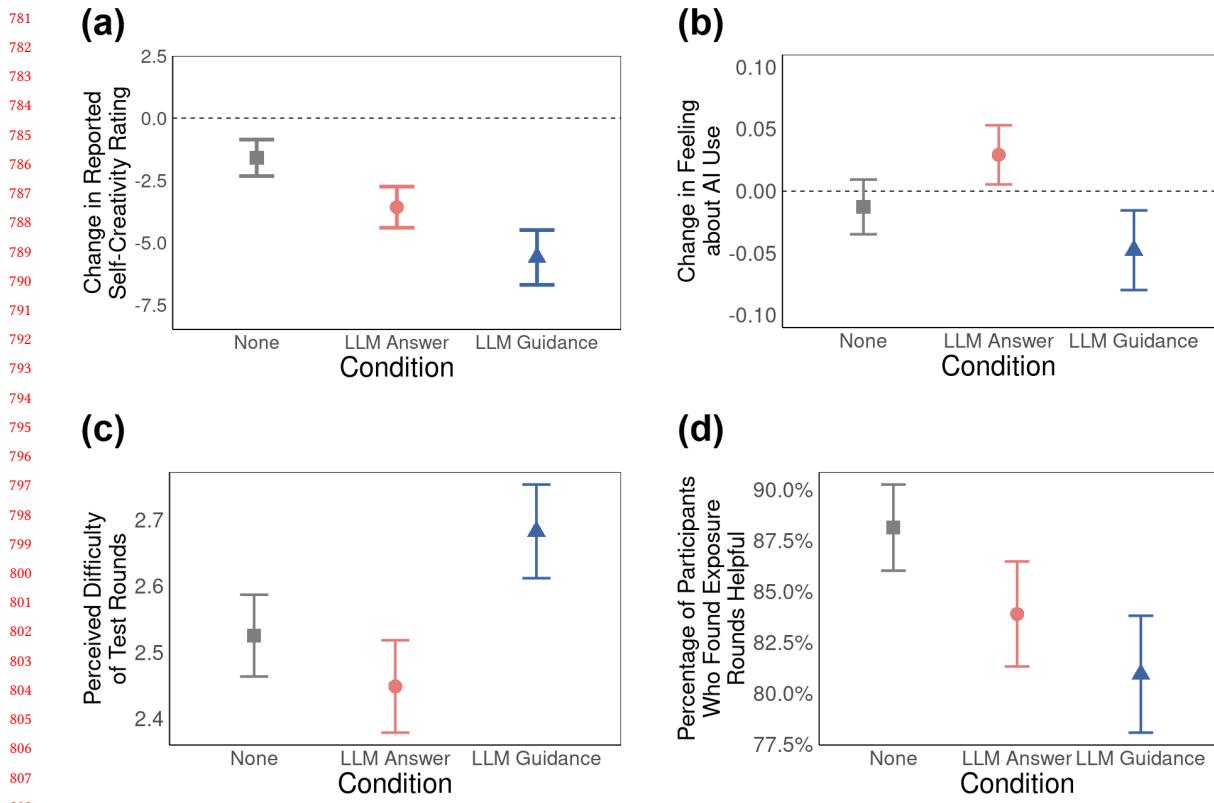


Fig. 10. Plots of participant accuracy on the Remote Associates Task, segmented by exposure and test phases across all experimental conditions. The top left figure shows participants' change in self perceived creativity ratings (Based on how many % of humans they felt they were more creative than), the 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.

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. Furthermore, 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 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

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 [??]. 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.

885 From a design perspective, it is critical to consider not just human-AI performance during exposure phases but also
 886 human performance in unassisted tasks after using AI. Systems should be designed with long-term human flourishing
 887 in mind, ensuring that the benefits of AI assistance do not come at the cost of diminished independent creative abilities.
 888 Ensuring that users can effectively transition from AI-supported creativity to autonomous creative work will help
 889 mitigate potential long-term harms associated with over-reliance on LLMs, as highlighted by concerns in the literature
 890 about cognitive decline with repeated AI use [?].
 891

892 **5.1.2 Differential Impact of LLMs on Divergent and Convergent Thinking.** Our experiments reveal that the effects of
 893 LLMs vary significantly depending on the aspect of creativity being measured. In divergent thinking tasks, where
 894 participants were asked to generate a wide range of ideas, we observed more skepticism toward LLM assistance.
 895 Participants seemed less inclined to adopt AI-generated suggestions, which may be due to the nature of divergent
 896 thinking itself—encouraging exploration and unconventional approaches [? ?]. In contrast, for the convergent thinking
 897 task, where participants were tasked with narrowing down the ideas to a single solution, LLM assistance during the
 898 exposure phase appeared more beneficial. This is consistent with the idea that convergent thinking is more structured
 899 and goal-oriented, making it easier for participants to recognize when LLMs are effectively guiding them toward the
 900 correct solution [?].
 901

902 LLMs, however, may also hinder the creative process. In divergent thinking, the introduction of AI-generated ideas
 903 may distract participants, consume valuable cognitive resources, or prevent full engagement with the task. This aligns
 904 with theories of creativity such as Boden’s concept of *conceptual spaces*, which emphasize the importance of users
 905 exploring and navigating creative possibilities independently [?]. Over-reliance on LLMs during divergent thinking may
 906 disrupt this exploration, leading to less original ideas, as seen in the originality results. Similarly, in convergent thinking
 907 tasks, LLMs might steer participants toward specific solutions too quickly, reducing the need for deep engagement with
 908 the problem space [?].
 909

910 These findings underscore the importance of carefully calibrating trust and reliance in human-LLM systems. Designers
 911 should incorporate measures that help users recognize when to trust and rely on LLMs, and when to prioritize their
 912 own cognitive processes. This approach can help balance the benefits of AI support with the risks of undermining the
 913 human creative process, ensuring that LLMs enhance rather than detract from creativity in the long term.
 914

915 **5.1.3 Persistent Homogenization of Ideas and the Challenges of Designing LLM Coaches.** Existing research has shown
 916 that LLM usage can lead to the homogenization of ideas within groups, where participants tend to converge on similar
 917 outcomes when using AI-generated suggestions [?]. Our findings extend this concern, showing that even when people
 918 stop using LLMs that provide strategic frameworks for thinking, the homogenization effect can persist. In our divergent
 919 thinking experiment, participants who received LLM-generated strategies exhibited reduced diversity in their idea sets
 920 both during and after LLM use, suggesting that such frameworks may have a lasting impact on creative processes,
 921 potentially stifling the generation of more varied or unconventional ideas.
 922

923 Interestingly, we did not find the same effect for participants who received direct ideas from a standard LLM. The
 924 *List of Ideas* condition did not lead to the same lasting homogenization once the LLM was no longer present. This
 925 indicates that while LLM-generated strategies can have long-term effects on creative diversity, providing ideas without a
 926 guiding framework may allow for more cognitive flexibility once participants stop using AI. In our convergent thinking
 927 experiment, participants who received LLM guidance during exposure performed worse in the test round compared to
 928 those who received direct answers. This highlights the complexity of designing LLMs as coaches or guides. While direct
 929 answers may not always seem ideal, in some cases, they can be more effective than offering frameworks or strategies.
 930 Manuscript submitted to ACM
 931
 932
 933
 934
 935
 936

937 Designers of Human-AI systems must take these findings into account, ensuring that LLM interactions are structured
938 to avoid unintentional long-term effects on cognitive diversity. Controlled experiments, such as those conducted here,
939 can be useful methods in refining and optimizing the design of coach-like LLM systems.
940

941 942 **5.2 Broader Implications for Fields Involving Human-AI Co-Creativity**

943 When designing AI tools for co-creativity, it is crucial to consider their long-term impact on human cognitive abilities.
944 Hofman *et al.* [?] introduced a useful metaphor—steroids, sneakers, and coach—to describe the spectrum of AI’s role in
945 human-AI collaboration. Our findings suggest that co-creative systems must be carefully designed to be coach-like to
946 prevent unintended consequences, such as stifling human creativity, even after AI assistance is removed.
947

948 These insights have broad implications for various fields, from scientific discovery to the arts and humanities. In the
949 context of AI for science, for instance, there is ongoing excitement about building increasingly sophisticated models
950 to accelerate the process of hypothesis generation and experimental execution [??]. However, the design of these
951 models often overlooks their potential impact on scientists’ creative abilities. Although some initial explorations have
952 been conducted, the field lacks rigorous empirical evaluation of how AI systems affect human creativity in scientific
953 discovery. The arts and humanities may face similar challenges [?]. For example, writers using the same LLM could
954 produce homogenized content, even when AI is no longer part of their workflow. This underscores the need for AI
955 systems that not only assist in the creative process but also promote long-term cognitive diversity, ensuring that human
956 creativity thrives in collaboration with AI rather than becoming constrained by it.
957

958 Our work in a controlled, simple setting suggests that the performance of AI models alone is not enough to determine
959 their value. The design of human-AI interactions—whether AI provides direct answers or encourages users to think
960 critically—and the timing of performance evaluation (during AI use versus after) both significantly alter the narrative.
961 These considerations are critical in ensuring that AI models enhance, rather than diminish, human creative potential.
962 Beyond simply building superhuman AI, we must focus on how these tools influence human creativity, culture, and
963 cognitive growth, aiming for AI systems that enrich and elevate human thought.
964

965 966 **5.3 Limitations & Future Work**

967 While our study provides valuable insights into the impact of LLMs on human creativity, there are several limitations
968 that warrant further investigation.
969

970 Firstly, measuring creativity itself presents challenges. Though we employed both divergent and convergent thinking
971 tasks (AUT and RAT) to capture a broad spectrum of creative processes, these tasks were limited to verbal responses
972 and conducted over short time periods. Creativity, however, is a much more complex and nuanced phenomenon.
973 Future work should expand the battery of tasks to include more natural, real-world creative activities, such as writing
974 advertisements or solving complex problems. Additionally, creativity tasks that are non-verbal or visual, like those
975 supported by diffusion models, could be explored. Experiments using visual creativity tasks, such as the Test of Creative
976 Thinking Drawing Production (TCT-DP) or the Evaluation of Potential Creativity (EoPC) [?], would provide deeper
977 insights into how LLMs impact creativity in non-text domains.
978

979 Our study also focused exclusively on the active, conscious aspects of creativity, largely due to the controlled lab
980 environment. Yet, creativity often involves unconscious processes that are harder to capture in such settings. Another
981 limitation lies in the short exposure period in our lab-based study. In real-world settings, exposure to LLMs is often
982 prolonged, integrated into daily workflows, and thus, likely produces more significant and lasting effects on creativity.
983 In contrast, our study’s shorter exposure periods may have resulted in smaller effect sizes, limiting our ability to fully
984

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.

References

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

Manuscript submitted to ACM

1041 You are an AI assistant tasked with the specific goal of teaching users to be more creative by applying the SCAMPER technique.
1042 SCAMPER is a powerful tool for creative thinking that involves exploring various ways to modify an object or idea to generate new
1043 possibilities. Your interactions should guide users through the SCAMPER process, which includes the following steps:

- 1044
- 1045 • **Substitute:** Encourage users to think about what elements of the object could be replaced with something else to improve or
1046 give it a new use.
 - 1047 • **Combine:** Prompt users to consider how they might combine the object with other items or ideas to create something new.
 - 1048 • **Adapt:** Ask users how the object could be adapted or modified to serve another purpose, or how it might be inspired by
1049 something else.
 - 1050 • **Modify:** Challenge users to think about how altering the shape, size, color, or any other attribute of the object could create
1051 new uses or improvements.
 - 1052 • **Put to another use:** Encourage users to brainstorm completely different uses for the object, beyond its original purpose.
 - 1053 • **Eliminate:** Guide users to consider what might happen if parts of the object were removed or simplified. How does this
1054 change its potential uses?
 - 1055 • **Reverse:** Ask users to consider what would happen if they reversed the object's purpose or use. Would it function differently
1056 or serve a new role?

1057
1058 In your guidance, encourage users to iterate through these steps, not necessarily in order, to explore the full range of creative
1059 possibilities. Don't mention SCAMPER explicitly in your response. Output a list that the user can follow. Don't include any examples,
1060 but only prompts or questions to help users ideate. Make your response concise and easy to follow for the users. Express each point in
1061 as few words as possible.

1062
1063 Fig. 11. The system prompt for the “List of Strategies” condition (not visible to participants).

1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092

A.2 Sample LLM Responses

Alternate Uses Task	LLM Ideas	LLM Strategies
Tire	<p>Garden Planter: Cut a tire in half and use it as a unique planter for flowers or vegetables.</p> <p>Swing: Hang a tire from a tree to create a fun outdoor swing for kids.</p> <p>Coffee Table: Stack a few tires on top of each other, paint them, and add a glass top to create a unique coffee table.</p> <p>Sandbox: Fill a large tire with sand to create a small sandbox for children.</p> <p>Tire Ottoman: Stuff a tire with cushioning material and cover it with fabric to create a unique ottoman.</p> <p>Tire Pond: Dig a hole in your backyard, place a tire in it, and fill it with water to create a small pond.</p> <p>Tire Climbing Wall: Attach several tires to a wall to create a climbing wall for kids.</p>	<p>What can you replace on the tire to improve it or give it a new use?</p> <p>How can you combine the tire with other items or ideas to create something new?</p> <p>How can the tire be adapted or modified to serve another purpose?</p> <p>How can altering the shape, size, or color of the tire create new uses or improvements?</p> <p>What completely different uses can you brainstorm for the tire, beyond its original purpose?</p> <p>What might happen if parts of the tire were removed or simplified? How does this change its potential uses?</p> <p>What would happen if you reversed the tire's purpose or use? Would it function differently or serve a new role?"</p>
Pants	<p>Plant Pot Cover: Use pants to cover unsightly plant pots. The pockets can also hold gardening tools.</p> <p>Art Canvas: Use the fabric of the pants as a canvas for painting or drawing.</p> <p>Hammock for Small Pets: Sew the ends of the legs together to create a cozy hammock for small pets.</p> <p>Wine Bottle Bag: Use a pant leg as a chic, recycled wine bottle gift bag.</p> <p>Patchwork Quilt: Combine various pants to create a unique patchwork quilt.</p> <p>DIY Bag: Convert an old pair of jeans into a stylish tote bag.</p> <p>Halloween Decorations: Stuff the pants with hay or newspapers and pair with a shirt to make a scarecrow.</p>	<p>What could you replace the fabric or parts of the pants with to create something new?</p> <p>How can you combine pants with another item to make a unique product?</p> <p>How could you modify pants to serve a completely different purpose?</p> <p>What happens if you change the shape, size, or color of the pants?</p> <p>What entirely different function could pants serve?</p> <p>What if you remove certain parts of the pants? How does this change their use?</p> <p>What if you reverse the way pants are used or worn? What new function could they serve?</p>

Fig. 12. Sample LLM responses for the exposure rounds.

1145 **B Experiment 2 (Convergent Thinking)**

1146 **B.1 LLM Configuration**

1148 *B.1.1 LLM Model Specification.*

- 1150
 1151
 1152 • **model version:** gpt-4o
 1153 • **date of use:** August 2024
 1155
 1156
 1157

1158 Configuration Settings:

- 1159
 1160
 1161
 1162 • **temperature:** 0
 1163 • **max response:** 800
 1164 • **top-p:** 0.95
 1165 • **frequency penalty:** 0
 1166 • **presence penalty:** 0
 1168

1169
 1170
 1171 *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
 1172 be asked to generate a fourth word that connects or fits with all three words. Specifically, the word that you generate,
 1173 when combined with each of the three stimulus words, should create three word-pairs, each of which makes a common
 1174 compound word or phrase. Please try to be creative and appropriate in your responses.
 1175

1176 Find a fourth word that is related to all three of the following:

1177 [word1], [word2], [word3]"
 1179

1180
 1181
 1182
 1183
 1184
 1185
 1186
 1187 You are a supportive coach guiding participants through solving Remote Associates Test (RAT) problems. For each of the three given
 1188 words, provide a concise response by listing five possible associated words. Encourage the participant to identify connections between
 1189 these words to find the fourth word, but do not directly give the solution. Conclude by suggesting they consider if any word connects
 1190 all three and nudge them to jot the words down on paper if it might help.
 1192

1193 Fig. 13. The system prompt for the *LLM Guidance* condition (not visible to participants).
 1194

B.2 Sample LLM Responses

Remote Associates Task	LLM Answer	LLM Guidance
flake, mobile, cone (snow)	Answer: Snow. (Snowflake, Snowmobile, Snowcone)	<p>Possible associated words for "flake" are snow, cereal, skin, coconut, and pastry.</p> <p>Possible associated words for "mobile" are phone, sculpture, home, device, and car.</p> <p>Possible associated words for "cone" are ice cream, volcano, traffic, pine, and speaker.</p> <p>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!</p>
dream, light, break (day)	Answer: Day. (Daydream, Daybreak, Daylight)	<p>Possible associated words for "dream" are sleep, subconscious, nightmare, fantasy, and aspiration.</p> <p>Possible associated words for "break" are fracture, pause, vacation, interruption, and opportunity.</p> <p>Possible associated words for "light" are lamp, bright, illuminate, candle, and photon.</p> <p>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!</p>
aid, rubber, wagon (band)	Answer: Band. (Bandaid, Rubber Band, Bandwagon)	<p>Possible associated words for "aid" are: help, support, assist, rescue, first.</p> <p>Possible associated words for "wagon" are: cart, horse, transport, carry, goods.</p> <p>Possible associated words for "rubber" are: elastic, tire, eraser, latex, ball.</p> <p>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.</p>

Fig. 14. Sample LLM responses for the exposure rounds.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Manuscript submitted to ACM