nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01953-1

An empirical investigation of the impact of ChatGPT on creativity

Received: 11 September 2023

Accepted: 10 July 2024

Published online: 12 August 2024

Check for updates

Byung Cheol Lee [1,3] ✉ & Jaeyeon (Jae) Chung [2,3]

This paper investigates the potential of ChatGPT for helping humans tackle problems that require creativity. Across five experiments, we asked participants to use ChatGPT (GPT-3.5) to generate creative ideas for various everyday and innovation-related problems, including choosing a creative gift for a teenager, making a toy, repurposing unused items and designing an innovative dining table. We found that using ChatGPT increased the creativity of the generated ideas compared with not using any technology or using a conventional Web search (Google). This effect remained robust regardless of whether the problem required consideration of many (versus few) constraints and whether it was viewed as requiring empathetic concern. Furthermore, ChatGPT was most effective at generating incrementally (versus radically) new ideas. Process evidence suggests that the positive influence of ChatGPT can be attributed to its capability to combine remotely related concepts into a cohesive form, leading to a more articulate presentation of ideas.

The recent advancement in generative artificial intelligence (AI) technology has captured worldwide interest. One notable AI, ChatGPT, developed by OpenAI, attracted more than 100 million users within just 2 months of its launch[1,2]. Generative AI has shown remarkable performance in objective knowledge assessments, such as the bar exam and US medical licensing exams[3,4], and its use has broadened to encompass a variety of fields, including coding, language translation and now creative tasks. For example, Copy.ai, an AI-powered copywriting service, claims to assist in writing blog and social media content, as well as aiding in email marketing. Whether in design, writing or engineering, the role of generative AI as a support tool to enhance human creativity continues to expand. Given this societal shift, we investigated the following question: can AI chatbots such as ChatGPT improve human response to problems that require creative thinking?

To explore this question, we surveyed lay individuals ($n = 100$) and discovered that they remain somewhat sceptical about the effectiveness of AI chatbots in creative tasks (Supplementary Note A). Specifically, although many acknowledge the utility of chatbots in translating languages (74%), programming (46%) and professional writing (24.2%), only a small fraction (7%) see value in their use for generating creative ideas. As this scepticism persists, the empirical study of generative AI's role in creative tasks is still in its early stages. Research findings on its performance are varied, with some studies suggesting that it is limited and others finding it to be remarkable, resulting in a wide range of opinions about AI's potential in creativity tasks[5–8]. However, research has yet to fully explore the application of AI chatbots in daily life, where the technology is often compared to other conventional approaches to creative problem-solving. This research therefore aims to examine whether using ChatGPT can enhance creativity beyond what is achievable without it, or when using traditional tools such as standard Web search engines.

Creativity is commonly viewed as the generation of original and appropriate ideas[9–13]. Originality refers to a new idea's deviation from common ideas, combining existing knowledge in novel ways through analogical thinking; when the idea is original, it is deemed novel, surprising and uncommon[10]. Meanwhile, appropriateness relates to the feasibility and practicality of the generated idea; when an idea is appropriate, it can be used to solve a problem and to improve the given context[11,13,14]. Either of these two—originality or appropriateness—is an outcome of creative thinking.

Regardless of whether it focuses on originality or appropriateness, creative problem-solving commonly involves connecting associative

[1]Department of Marketing, Bauer College of Business, University of Houston, Houston, TX, USA. [2]Department of Marketing, Jones Graduate School of Business, Rice University, Houston, TX, USA. [3]These authors contributed equally: Byung Cheol Lee, Jaeyeon (Jae) Chung. ✉e-mail: blee30@uh.edu

elements to create new combinations of concepts[15–19]. Creative ideas thus often result from the cohesive association of two seemingly unrelated pieces of knowledge[17]. To achieve this, individuals rely on techniques such as brainstorming, sketching and storyboarding or digital tools such as Web services and interactive whiteboards[20,21]. In fact, our pilot survey revealed that, when faced with a problem requiring creative thinking, a vast majority (68%) of lay consumers seek assistance from online resources such as Google (Supplementary Note B).

With a specific focus on ChatGPT, this paper examines whether AI chatbots can substantially enhance creative problem-solving beyond what is achievable without them. Specifically, we compare ChatGPT with a commonly used resource—namely, traditional Web search engines without AI chatbot functionality. While traditional Web search engines such as Google have made accessing information easier, enabling users to find information with just a few keystrokes, the tasks of navigating through search results and synthesizing information from multiple sources still largely require human effort. Designed to generate text on the basis of the context, ChatGPT goes beyond traditional Web searches by integrating disparate pieces of knowledge into coherent, articulate verbal responses, rather than merely presenting a list of relevant pages. For example, recent studies examining ChatGPT's performance without human intervention have found that ChatGPT demonstrates substantial creative capabilities[22–24]. Consequently, we argue that ChatGPT can substantially help users solve creative problems by improving the associative process involved in creative problem-solving.

Our experiments focus on incentive-compatible idea contests that require verbal output, considering ChatGPT's capabilities as a large language model. To avoid any potential biases from ChatGPT's prior exposure to the tasks, we avoided publicly available tests of creativity such as the Alternative Uses Test or Torrance Test. Instead, we designed problem-solving tasks, based on the paradigm adapted from creativity literature[13,25]. In all experiments, participants were randomly assigned to either use or not use ChatGPT assistance. Their submitted ideas were later rated by external judges on originality and appropriateness sub-dimensions, which were averaged to compute the overall creativity score. It is important to note that our focus in this paper is the impact of ChatGPT's assistance on the creativity of an idea, rather than on task efficiency and productivity, which concern the ratio between input and output in a performance task[26,27].

Results

Experiment 1

We recruited MTurk participants (n = 233; mean age, 42.29 years; 52% female) and had them engage in a creative idea contest adapted from prior creativity research[13]. They were asked to use three items—paper clips, water bottles and paper bags— to come up with a new toy for a child between 5 and 11 years old (see Supplementary Note C for the survey materials and Supplementary Note D for the sample responses). Crucially, the participants were instructed to use either ChatGPT (the ChatGPT-assisted condition) or Google (the Web-search-assisted condition) to help them come up with their idea. By 'Web search', we mean traditional search engines—prior to the introduction of AI chatbot functionality such as Gemini. We selected Web search as a baseline because it is one of the most commonly used aids (Supplementary Note B) and because it can also provide access to external information in response to user queries. Indeed, the two technologies—ChatGPT and Web search—have been the subject of recent media discussions speculating whether the rise of ChatGPT will disrupt and change people's reliance on Web search engines[28,29]. The participants' responses from the ChatGPT-assisted condition and the Web-search-assisted condition were evaluated by external judges hired from the same population (n = 331; see Methods for more information), following prior research[30].

The results revealed that ideas from the ChatGPT-assisted condition received higher creativity ratings than those from the Web-search-assisted condition (mean for the ChatGPT-assisted condition ($M_{\text{ChatGPT-assisted}}$) = 4.28, s.d. = 0.81 versus $M_{\text{Web-search-assisted}}$ = 3.89, s.d. = 0.80; two-sided t-test: $t_{231} = 3.68$; $P = 2.87 \times 10^{-4}$; Cohen's d = 0.49; 95% confidence interval (CI), (0.22, 0.75)). We also compared the ratings for the sub-dimensions and found that both originality ($t_{231} = 3.39$; $P = 8.23 \times 10^{-4}$; d = 0.45; 95% CI, (0.18, 0.71)) and appropriateness ($t_{231} = 3.49$; $P = 5.80 \times 10^{-4}$; d = 0.46; 95% CI, (0.20, 0.72)) scores were higher in the ChatGPT-assisted condition than in the Web-search-assisted condition (see Supplementary Note E for analyses of sub-dimensions for the current experiment and all the remaining experiments). Additional analyses revealed that the ChatGPT-assisted condition not only had a higher average creativity score but also had a higher percentage of the ideas with top-ranked creativity scores (see Supplementary Note F for this additional analysis).

As a robustness check, we hired three expert judges who had at least 5 years of experience in businesses, product management and consulting. The results were highly consistent with those from the judges in the online participation pool ($M_{\text{ChatGPT-assisted}}$ = 3.06, s.d. = 1.40 versus $M_{\text{Web-search-assisted}}$ = 2.65, s.d. = 0.98; $t_{231} = 2.61$; P = 0.010; d = 0.34; 95% CI, (0.08, 0.61); see Supplementary Note G for the details).

Experiment 1 demonstrates that using ChatGPT leads to more creative ideas than conventional methods, such as Web search—a comparison that mirrors the ongoing debate about whether ChatGPT will outperform Web search[28,29]. However, one may question the validity of comparing ChatGPT to conventional Web search. That is, one might argue that creativity is inherently human, and the most creative ideas may emerge when people do not receive assistance from any external technology. To address this possibility, Experiment 2A includes a Human-only condition as another benchmark.

Experiment 2A

This experiment used a three-level between-participants (ChatGPT-assisted, Web-search-assisted and Human-only) survey design. Prolific participants (n = 291; mean age, 39.80 years; 44% female) were randomly assigned to one of the three conditions. Participants in the ChatGPT-assisted condition were told that they could use the ChatGPT website, while those in the Web-search-assisted condition were told that they could use the Google search engine. Participants in the Human-only condition were not provided with any information about external Web resources; they were simply instructed to come up with a creative idea on their own. For the creativity task, we provided a task that many participants can relate to—they were asked to repurpose unused household items (an old tennis racket and a garden hose) and come up with a new way of using the products. To examine the creativity of the submitted ideas while explicitly controlling for the quantity (that is, the number of ideas generated), we asked the participants to submit just one idea. Each idea was evaluated by external judges hired from the same population pool (n = 232; mean age, 44.53 years; 54% female). This experiment was preregistered.

A one-way analysis of variance (ANOVA) revealed a significant main effect of the condition ($F_{2,288} = 13.56$, $P = 2.35 \times 10^{-6}$, $\eta_p^2 = 0.09$). Again, ideas submitted by participants in the ChatGPT-assisted condition were rated as more creative ($M_{\text{ChatGPT-assisted}}$ = 4.56, s.d. = 0.75) than those in the Web-search-assisted condition ($M_{\text{Web-search-assisted}}$ = 3.98, s.d. = 0.94; $t_{288} = 4.97$; $P = 1.16 \times 10^{-6}$; d = 0.71; 95% CI, (0.42, 1.00)) and the Human-only condition ($M_{\text{Human-only}}$ = 4.11, s.d. = 0.74; $t_{288} = 3.85$; $P = 1.45 \times 10^{-4}$; d = 0.56; 95% CI, (0.27, 0.85)). This pattern of results was again observed when using expert judges' ratings. Ideas from the ChatGPT-assisted condition were rated as more creative than those from the Web-search-assisted condition ($t_{288} = 3.51$; $P = 5.22 \times 10^{-4}$; d = 0.50; 95% CI, (0.22, 0.79); Bayes factor, 33.3), although they were not significantly more creative than those from the Human-only condition ($t_{288} = 1.79$; P = 0.075; d = 0.26; 95% CI, (−0.03, 0.54); Bayes factor, 0.71; Supplementary Note G). Overall, when the Human-only condition was included as an additional benchmark, two insights emerged: first, the

use of ChatGPT improves creativity beyond what a person can achieve without any Web aid; and second, ChatGPT facilitates creativity, rather than Web search impairing it.

Experiment 2B

Experiment 2B had three goals. First, we sought to replicate the findings on ChatGPT's assistance with creativity, by again comparing this condition with the Human-only condition. Second, we explored whether it is necessary to have additional human modification—in the form of editing or refining ChatGPT's outputs—for an idea to be seen as creative. To investigate this, we included a ChatGPT-only condition that featured raw initial responses from ChatGPT without any human editing or refinement. We predicted that ChatGPT, whether in collaboration with humans (the ChatGPT-assisted condition) or operating independently (the ChatGPT-only condition), would outperform the Human-only approach, due to its ability to combine remotely related concepts into coherent, articulate responses. Third, we aimed to identify the type of creativity that is most likely to be enhanced by ChatGPT. The literature distinguishes two forms of creativity: incremental and radical[31,32]. Incrementally new ideas enhance existing concepts, improving on what already exists. In contrast, radically new ideas are not just enhancements; they are disruptively new and different, breaking away from the existing paradigm[31,32]. ChatGPT, with its ability to access existing concepts from its database and combine them in a cohesive way, is likely to be effective at generating incrementally new ideas. However, its capacity to generate radically new ideas—those that represent a substantial departure from existing concepts—may be limited. We tested this hypothesis by classifying the type of creativity featured in each response (incrementally new versus radically new versus not new).

This experiment used a three-level between-participants (ChatGPT-assisted, Human-only and ChatGPT-only) survey design. For the creativity task, Prolific participants were asked to create a toy for a 7-year-old child using three items: a paper bag, a leftover construction brick and an unused fan. As in Experiment 2A, we asked the participants (n = 200; mean age, 41.68 years; 52% female) to submit one creative idea. Participants in the ChatGPT-assisted condition were instructed to use the ChatGPT website. In the Human-only condition, participants were instructed to respond to the task without using any external Web resources. For the ChatGPT-only condition, we requested 100 responses from GPT-3.5 by entering the instruction for the ideation task on the ChatGPT website, following prior research[5]: the initial responses were simply collected without editing. We used GPT-3.5 because a pilot study revealed that the majority of Prolific participants only have access to the free, default version of GPT-3.5 (and not the paid, premium version, GPT-4.0). Indeed, this was confirmed in our main experiment; the majority of participants (98.04%) reported having used GPT-3.5. After all ideas were collected, these ideas were evaluated by external judges hired from Prolific (n = 241; mean age, 42.27 years; 56% female). This experiment was preregistered.

A one-way ANOVA revealed a significant main effect of the condition ($F_{2,297}$ = 117.00, $P = 4.36 \times 10^{-38}$, η_p^2 = 0.44). Replicating the findings from the previous experiment, ideas submitted by participants in the ChatGPT-assisted condition were rated as more creative ($M_{\text{ChatGPT-assisted}}$ = 4.84, s.d. = 0.47) than those from participants in the Human-only condition ($M_{\text{Human-only}}$ = 3.66, s.d. = 0.89; t_{297} = 13.30; $P = 5.56 \times 10^{-32}$; d = 1.88; 95% CI, (1.56, 2.20)). Additionally, as expected, ideas from the ChatGPT-only condition were rated as more creative ($M_{\text{ChatGPT-only}}$ = 4.83, s.d. = 0.40) than those from the Human-only condition ($M_{\text{Human-only}}$ = 3.66, s.d. = 0.89; t_{297} = 13.21; $P = 1.18 \times 10^{-31}$; d = 1.88; 95% CI, (1.56, 2.20)). Surprisingly, we found no credible evidence of a difference between the creativity ratings in the ChatGPT-assisted condition ($M_{\text{ChatGPT-assisted}}$ = 4.84, s.d. = 0.47) and those in the ChatGPT-only condition ($M_{\text{ChatGPT-only}}$ = 4.83, s.d. = 0.30; t_{297} = 0.03; P = 0.980; d < 0.01; 95% CI, (−0.27, 0.28); Bayes factor, 0.15). This pattern of findings was replicated when expert judges' scores were used for the analysis ($M_{\text{ChatGPT-assisted}}$ = 3.85 versus $M_{\text{Human-only}}$ = 2.67; t_{297} = 12.54; $P = 3.22 \times 10^{-29}$; d = 1.77; 95% CI, (1.46, 2.09); $M_{\text{ChatGPT-only}}$ = 3.87 versus $M_{\text{Human-only}}$ = 2.67; t_{297} = 12.71; $P = 7.77 \times 10^{-30}$; d = 1.81; 95% CI, (1.49, 2.12); Supplementary Note G).

We also examined the type of creativity enhanced by ChatGPT. Expert judges scored ideas as not new at all, incrementally new (that is, providing new combinations of existing ideas) or radically new (that is, going beyond mere new combinations to truly new ideas). Analysis of expert judges' ratings revealed that ChatGPT is particularly effective at generating incrementally new ideas but less effective at generating radically new ideas (see Supplementary Note G for the details).

In sum, Experiment 2B replicates the positive influence of ChatGPT's assistance on creativity. Importantly, Experiment 2B adds two critical insights. First, ChatGPT is particularly effective at generating incrementally new ideas but less effective at producing radically new ideas. This makes sense because ChatGPT uses existing knowledge to make connections; thus, the submitted ideas are likely to be an incrementally improved version of what already exists. Second, ChatGPT outperformed the Human-only approach both when paired with additional human modifications (ChatGPT-assisted condition) and when operating in isolation (ChatGPT-only condition). This finding suggests that additional human modifications to ChatGPT's output may not necessarily make an idea more creative. Indeed, when we conducted a follow-up analysis among participants in the ChatGPT-assisted condition, their self-reported level of modification of ChatGPT's initial responses was not correlated with higher creativity ratings (r = 0.11, P = 0.291; for further discussion, see Supplementary Note H). On the basis of this evidence, we infer that using (versus not using) ChatGPT increases creativity, while the participants' additional modification of ChatGPT's responses may not further increase creativity.

What makes ideas based on ChatGPT responses creative? We posited that the positive impact of ChatGPT use on creativity may be driven by ChatGPT's ability to synthesize different concepts into articulate, cohesive responses. We shed light on this process in the next experiment.

Experiment 3

Experiment 3 delves into the process mechanism that underpins our findings. Creative problem-solving often involves accessing a wide range of distant pieces of knowledge and then linking them in a cohesive way[12,20]. We propose that using ChatGPT (versus search engines that just present a list of relevant pages) can aid in this process[33,34]. That is, we propose that using ChatGPT (versus Web search) is likely to aid in developing a more coherent and clearer conceptualization of an idea (that is, it serves as an 'idea exposition aid'), which, in turn, should increase how well the idea is articulated (that is, 'idea articulateness').

To test this idea, we randomly assigned MTurk participants (n = 194; mean age, 43.11 years; 50% female) to either the ChatGPT-assisted or the Web-search-assisted condition. The participants were instructed to use either ChatGPT or a Web search to come up with a creative idea for a dining table that did not exist on the market. We then measured our first theoretical mediator by asking them the degree to which they felt that idea exposition was possible through the use of ChatGPT (or Web search). Later, ideas were evaluated by external judges hired from MTurk (n = 158; mean age, 43.11 years; 50% female). To fully capture the process mechanism, we also asked these judges to rate how articulate each idea seemed (Methods and Supplementary Note I). This measure served as the second mediator.

Replicating the previous findings, the ideas from the ChatGPT-assisted (versus Web-search-assisted) condition were rated as more creative ($M_{\text{ChatGPT-assisted}}$ = 4.60, s.d. = 0.72 versus $M_{\text{Web-search-assisted}}$ = 4.35, s.d. = 0.59; t_{192} = 2.59; P = 0.010; d = 0.37; 95% CI, (0.09, 0.66)). Importantly, participants reported that ChatGPT (versus Web search) enabled idea exposition to a greater extent ($M_{\text{ChatGPT-assisted}}$ = 5.16, s.d. = 1.96 versus $M_{\text{Web-search-assisted}}$ = 2.94, s.d. = 1.59;

https://doi.org/10.1038/s41562-024-01953-1

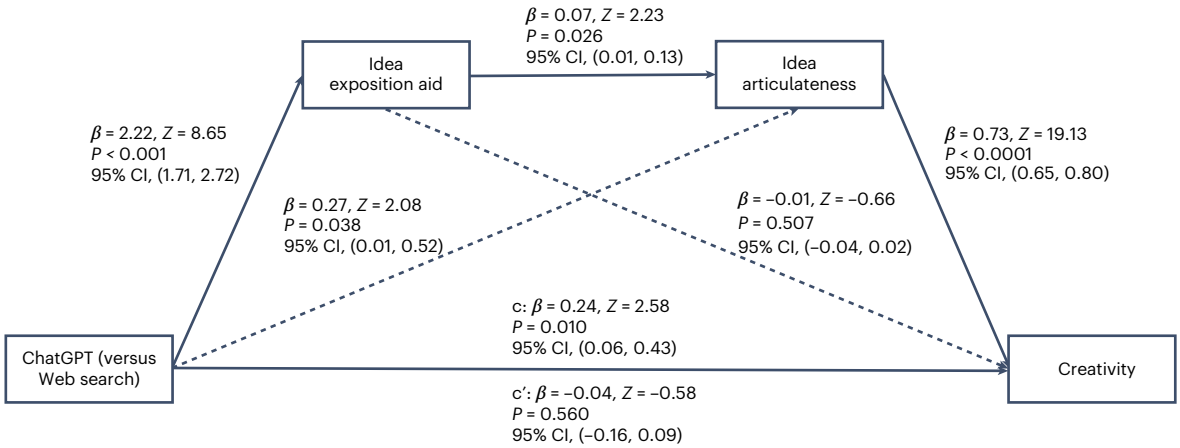

Fig. 1 | Process evidence. A serial mediation model demonstrates that the effect of ChatGPT on creativity is mediated by idea exposition aid and idea articulateness (Experiment 3). The mediation model was tested using the lavaan package (version 0.6.12) for R with 5,000 bootstrap samples. Two-sided *Z*-tests were performed to test the significance of the coefficients. Path c denotes the total effect, and path c′ denotes the direct effect.

$t_{192} = 8.69$; $P = 1.54 \times 10^{-15}$; $d = 1.25$; 95% CI, (0.94, 1.56)), which in turn improved the external judges' ratings of the extent to which the submitted ideas were articulate ($M_{\text{ChatGPT-assisted}} = 4.80$, s.d. = 0.90 versus $M_{\text{Web-search-assisted}} = 4.38$, s.d. = 0.58; $t_{192} = 3.86$; $P = 1.53 \times 10^{-4}$; $d = 0.55$; 95% CI, (0.27, 0.84)). A serial mediation analysis confirmed that the effect of ChatGPT on creativity was mediated by idea exposition aid and idea articulateness (indirect effect, 0.11; 95% 5,000 bootstrap CI, (0.01, 0.22)). The significant indirect effect was replicated when we repeated the analyses using expert judges' ratings (indirect effect, 0.20; 95% CI, (0.09, 0.34); Fig. 1; see Supplementary Note G for the details). We also tested for and ruled out potential alternative explanations, such as the perception of having a conversation, task engagement or affect (see Methods and Supplementary Note I for the details). Overall, these findings indicate that ChatGPT enhances creativity by accessing disparate concepts and combining them in a cohesive and articulate manner.

The next two experiments investigate the role of two possible moderators. We examined whether ideation tasks that have a high number of constraints and ideation tasks that require empathetic consideration hinder ChatGPT's ability to outperform conventional Web search.

Experiment 4

Experiment 4 examines the role of constraints imposed in creative problem-solving. Past research on creativity suggests that people become more creative when faced with many (versus few) task constraints. For example, if someone is deciding what to cook for a regular guest who has no dietary restrictions versus a vegan guest who has diabetes, high cholesterol and a peanut allergy, the latter case is likely to require more creative thinking. This occurs because constraints force people to deviate from conventional thinking patterns in favour of a more creative processing approach[13]. On the basis of this reasoning, one might expect that a high number of task constraints would lead participants in the Web-search-assisted condition to improve their ideation, resulting in participants in both conditions performing equally well on the creativity task. However, we hypothesized that ChatGPT's capability to integrate remote pieces of knowledge into cohesive, articulate verbal responses would allow it to outperform conventional Web search, even under conditions of substantial task constraints.

This experiment used a two (technology: ChatGPT-assisted versus Web-search-assisted) by two (level of constraint: Low versus High) survey design. Prolific participants were assigned to one of the two website conditions and were asked to come up with one creative gift idea for a teenage girl. We manipulated the level of constraints by instructing them to consider either two or six constraints during the ideation task[13]. In the Low (versus High) constraint condition, we provided two (versus six) constraints randomly selected out of the list of nine constraints—these constraints included sentimental value, sustainability and self-expression goals (for the full list of nine items, see Supplementary Note J). The creativity of each idea was evaluated by external judges recruited from Prolific ($n = 321$; mean age, 43.45 years; 51% female). This experiment was preregistered.

As predicted, we again found that ChatGPT outperformed Web search. A 2 × 2 ANOVA on creativity score revealed a significant main effect of ChatGPT ($F_{1,396} = 5.06$, $P = 0.025$, $\eta_p^2 = 0.01$). The main effect of the constraint level ($F_{1,396} = 1.53$, $P = 0.218$, $\eta_p^2 < 0.01$) and the interaction ($F_{1,396} = 0.01$, $P = 0.904$, $\eta_p^2 < 0.01$) were not significant. We replicated the previous findings that ideas submitted by participants in the ChatGPT-assisted (versus Web-search-assisted) condition were rated as more creative in both the High constraint conditions ($M_{\text{ChatGPT-assisted}} = 4.73$, s.d. = 0.66 versus $M_{\text{Web-search-assisted}} = 4.53$, s.d. = 0.68; $t_{396} = 2.25$; $P = 0.025$; $d = 0.32$; 95% CI, (0.04, 0.60)) and the Low constraint conditions ($M_{\text{ChatGPT-assisted}} = 4.62$, s.d. = 0.57 versus $M_{\text{Web-search-assisted}} = 4.40$, s.d. = 0.61; $t_{396} = 2.40$; $P = 0.017$; $d = 0.34$; 95% CI, (0.06, 0.63)). A robustness check using expert judges again revealed highly similar results ($M_{\text{ChatGPT-assisted}} = 3.30$, s.d. = 1.01 versus $M_{\text{Web-search-assisted}} = 2.95$, s.d. = 1.05; $t_{396} = 3.52$; $P = 4.87 \times 10^{-4}$; $d = 0.35$; 95% CI, (0.15, 0.55); Supplementary Note G).

Experiment 5

The final experiment investigates whether ChatGPT can enhance creativity in tasks commonly viewed as requiring emotional empathy, which is defined as sensing the feelings of others and showing an active interest in their concerns[35–37]. In our daily lives, we encounter many situations that demand emotional empathy, such as when one has to prepare a personalized condolence gift or to understand another person's perspective. Such empathy-demanding scenarios have been perceived as requiring uniquely human traits; thus, people often perceive AI as less adept at addressing these problems than other types of problems[38,39]. Some might therefore view AI chatbots as less useful in creative tasks that demand substantial empathetic understanding. However, contrary to this popular belief, we hypothesized that ChatGPT can still enhance creative thinking, regardless of whether empathetic understanding is viewed as necessary. That is, even when a problem is expected to require an understanding of others' feelings, ChatGPT can still help improve creativity by combining remote pieces of knowledge into cohesive, articulate responses.

To test our prediction, we ran an experiment using a two (technology: ChatGPT-assisted versus Web-search-assisted) by two (empathy: Baseline versus High) between-participants design. Prolific participants ($n = 383$; mean age, 40.24 years; 47% female) were randomly

https://doi.org/10.1038/s41562-024-01953-1

Table 1 | Summary statistics

Experiment	Condition	Mean	s.d.	Statistic	*P*	Cohen's d	95% CI
1	ChatGPT-assisted	4.28	0.81	t_{231}=3.68	2.87×10^{-4}	0.49	(0.22, 0.75)
	Web-search-assisted	3.89	0.80				
2A	ChatGPT-assisted	4.56	0.75				
	Web-search-assisted	3.98	0.94	t_{288}=4.97	1.16×10^{-6}	0.71	(0.42, 1.00)
	Human-only	4.11	0.74	t_{288}=3.85	1.45×10^{-4}	0.56	(0.27, 0.85)
2B	ChatGPT-assisted	4.84	0.47	t_{297}=13.30	5.56×10^{-32}	1.88	(1.56, 2.20)
	ChatGPT-only	4.83	0.40	t_{297}=13.21	1.18×10^{-31}	1.88	(1.56, 2.20)
	Human-only	3.66	0.89				
3	ChatGPT-assisted	4.60	0.72	t_{192}=2.59	1.03×10^{-2}	0.37	(0.09, 0.66)
	Web-search-assisted	4.35	0.59				
4	ChatGPT-assisted × High constraints	4.73	0.66	t_{396}=2.25	2.51×10^{-2}	0.32	(0.04, 0.60)
	Web-search-assisted × High constraints	4.53	0.68				
	ChatGPT-assisted × Low constraints	4.62	0.57	t_{396}=2.40	1.68×10^{-2}	0.34	(0.06, 0.63)
	ChatGPT-assisted × Low constraints	4.40	0.61				
5	ChatGPT-assisted × High empathy	4.47	0.73	t_{379}=5.26	2.44×10^{-7}	0.77	(0.48, 1.07)
	Web-search-assisted × High empathy	3.83	0.97				
	ChatGPT-assisted × Baseline empathy	4.29	0.68	t_{379}=7.02	1.04×10^{-11}	1.01	(0.72, 1.30)
	Web-search-assisted × Baseline empathy	3.45	0.96				

Two-sided *t*-tests were performed without adjustments made for multiple comparisons.

assigned to either the ChatGPT-assisted or the Web-search-assisted condition. Participants in the Baseline empathy condition read a scenario in which they were asked to submit one idea to repurpose two unused items (an old flashlight and a hair spray). Those in the High empathy condition read the same scenario and were additionally informed that these items were adored and frequently used by one's daughter when she was young. The empathy manipulation was successful in altering the perceptions of the level of empathy required to solve the problem (see Supplementary Note K for the details). Ideas were evaluated by external judges from the same population pool (n = 310; mean age, 43.68 years; 59% female). This experiment was preregistered.

As predicted, we found that using ChatGPT outperformed conventional Web search. A 2 × 2 ANOVA on creativity score revealed a significant main effect of technology manipulation ($F_{1,379}$ = 49.26, $P = 1.04 \times 10^{-11}$, η_p^2 = 0.12). The main effect of the empathy manipulation ($F_{1,379}$ = 2.43, P = 0.120, $\eta_p^2 < 0.01$) and the interaction ($F_{1,379}$ = 1.34, P = 0.248, $\eta_p^2 < 0.01$) were not significant. ChatGPT led to an increase in creativity regardless of whether the task was expected to require a high level of empathy ($M_{\text{ChatGPT-assisted}}$ = 4.47, s.d. = 0.73 versus $M_{\text{Web-search-assisted}}$ = 3.83, s.d. = 0.97; t_{379} = 5.26; $P = 2.44 \times 10^{-7}$; d = 0.77; 95% CI, (0.48, 1.07)) or not ($M_{\text{ChatGPT-assisted}}$ = 4.29, s.d. = 0.68 versus $M_{\text{Web-search-assisted}}$ = 3.45, s.d. = 0.96; t_{379} = 7.02; $P = 1.04 \times 10^{-11}$; d = 1.01; 95% CI, (0.72, 1.30)). The findings were again replicated when the data were coded by expert judges ($M_{\text{ChatGPT-assisted}}$ = 3.96, s.d. = 0.96 versus $M_{\text{Web-search-assisted}}$ = 3.20, s.d. = 0.96; t_{379} = 7.68; $P = 1.37 \times 10^{-13}$; d = 0.79; 95% CI, (0.58, 1.00); Supplementary Note G).

Discussion

We found that using (versus not using) ChatGPT can increase the creativity of responses to problem-solving tasks. This positive effect is robust across various types of tasks, including generating creative gift ideas, repurposing items and designing innovative dining tables. Surprisingly, ChatGPT also led to superior performance in creativity tasks requiring a high number of task constraints and even in tasks commonly viewed as requiring empathetic considerations. Our findings remained robust when the ideas were evaluated by lay consumers and by expert judges (see Table 1 for the summary of findings).

This work contributes to the expanding field of research focused on ChatGPT's effectiveness in creative idea generation[3–8]. It does so in three important ways. First, we focus on realistic everyday scenarios that demand creativity, demonstrating that ChatGPT can assist humans in a variety of real-world problems in daily life. This focus emphasizes the practical significance of our findings for everyday users, moving beyond stylized tests such as the Alternative Uses or Torrance Test[5–7]. Second, our study is among the first to empirically provide process evidence explaining how ChatGPT facilitates creativity. We demonstrate that ChatGPT's strength lies in its ability to bring together diverse concepts in a clear and coherent manner. Lastly, we identify the particular aspect of creativity where ChatGPT excels. Specifically, our results show that ChatGPT is highly effective at generating ideas that are incrementally new rather than radically new. This result probably stems from ChatGPT's proficiency in combining various concepts from its database rather than inventing entirely new concepts from scratch.

Surprisingly, ChatGPT showed competence in tasks that people expect to require empathy. Prior beliefs have often emphasized the uniquely human ability to understand emotions, empathize and 'read between the lines', leading to expectations that ChatGPT would fall short in scenarios that appear to require such emotional understanding. Contrary to these expectations, however, our results show that ChatGPT demonstrates remarkable performance in these very tasks. This finding highlights ChatGPT's potential for assisting not only in analytical tasks but also in tasks that traditionally require emotional understanding.

We note another finding that may surprise some readers: we found that ideas sourced directly from ChatGPT were judged to be as creative as those subsequently modified by human participants (Experiment 2B). This observation was further supported by a follow-up analysis in Experiment 1, which is detailed in Supplementary Note H. Overall, our results suggest that further human modifications to ChatGPT's outputs do not necessarily enhance creativity. This result may hint at the possibility that ChatGPT can serve as an automated agent for creativity, generating creative ideas without any human intervention. However, it is important to note that our studies involved everyday tasks performed by laypeople, who might lack the in-depth domain expertise required for making concrete modifications. This limitation opens an avenue for future research to explore the role of domain experts. For example, experienced product designers or engineers might be better equipped than lay consumers to provide meaningful modifications and to further improve ChatGPT's answers.

This paper not only introduces exciting opportunities but also raises numerous questions that warrant further exploration. First, future research could explore whether ChatGPT systematically improves efficiency or productivity in creative performance—for example, by examining the ratio between task inputs and outputs, or the speed of output production. Second, we encourage future researchers to investigate alternative forms of creative idea generation, such as group brainstorming, and how groups may benefit from ChatGPT assistance. Third, we suggest exploring various conditions that enhance creativity, focusing on factors such as the nature of the task (for example, verbal or figural) and individual differences (for example, personality traits, education level or professional experiences in the field). Also, while this paper focused on the impact of ChatGPT's assistance on verbal forms of creativity, its impact on other creative task domains such as video creation warrants further investigation. Finally, future research can analyse how different creativity tasks affect the variation in effect sizes. For instance, the effect was stronger in experiments that involved repurposing a set of items (with Cohen's *d* values exceeding 0.5 in Experiments 2A, 2B and 5) than in experiments that did not specify a set of items to use (with Cohen's *d* values at or below 0.4 in Experiments 3 and 4).

To summarize, we found evidence that the use of AI chatbots such as ChatGPT can improve creative problem-solving. We highlight ChatGPT's positive impact in a context where laypeople benefit from using it in everyday scenarios. More broadly, the emergence of advanced AI technologies introduces a wide range of unknown questions regarding their potential and risks. Our investigation is one of the early efforts to address those questions.

Methods

This research was approved by the Institutional Review Board at Rice University (IRB-FY2022-122). Informed consent was obtained from all participants. The participants voluntarily took part in each survey in exchange for a small nominal fee and had the right to decide whether or not to participate in the study. The risks associated with participation were minimal. The participants were informed that the potential benefit of the study is to contribute to an academic journal.

Experiment 1

MTurk participants in the USA participated in a survey titled 'Social Science Survey'. They were randomly assigned to one of the two experimental conditions: ChatGPT-assisted or Web-search-assisted. They were blind to the hypotheses and conditions of the experiment. They engaged in a creative idea-generation contest, in which they were asked to use three items—paper clips, water bottles and paper bags—to come up with a new toy for a child between 5 and 11 years old (see Supplementary Note C for the survey materials). The participants could submit as many ideas as they liked, and there was no time limit on the task. They entered their ideas into an open-ended text box. We collected 260 complete responses from MTurk. After we removed 27 participants who provided non-sensical responses, 233 participants (mean age, 42.29 years; 52% female) were retained for analysis. No statistical methods were used to predetermine sample sizes, but our sample sizes are larger than those reported in previous research[13,22].

The participants entered around 1.69 ideas on average, and there was no difference in the number of ideas entered between the ChatGPT-assisted and the Web-search-assisted condition ($t_{231} = 1.42$, $P = 0.149$, $d = 0.19$). Sample ideas from each condition are available in Supplementary Note D.

Creativity ratings. The ideas submitted by the participants were evaluated by external judges recruited from the same population as the participants (MTurk; $n = 331$; mean age, 43.11 years; 57% female). Each idea was evaluated using three items for originality (original/innovative/creative; 1, not at all; 4, neutral; 7, very much) and three items for appropriateness (practical/effective/useful; 1, not at all; 4, neutral; 7, very much). These six items were averaged to compute the creativity score ($\alpha = 0.96$). As participants could have submitted more than one idea, the scores for each idea were aggregated to calculate an average creativity score for each participant, following the procedure in previous research[13].

As a robustness check, we hired three expert judges with more than 5 years of professional experience in product management, consulting and business strategy. The rationale for involving expert judges was twofold: to ensure that the ideas submitted were recognized as creative by professionals in the field and to mitigate concerns that external judges, drawn from the same participant pool, might not fully understand the concept of creativity or could mistake verbosity for creativity. These expert judges evaluated all ideas in terms of creativity using an item for originality (1, not at all original/innovative/creative; 7, very original/innovative/creative) and another item for appropriateness (1, not at all useful/effective/implementable; 7, very useful/effective/implementable). Additionally, these expert judges scored each idea in terms of articulateness (1, low in articulateness; 7, very articulate) and verbosity (1, used appropriate number of words; 7, used more words than necessary). Further details on the measures and analyses conducted are available in Supplementary Note G.

Experiment 2A

Prolific participants in the USA were randomly assigned to one of the three experimental conditions: ChatGPT-assisted, Web-search-assisted or Human-only. The participants were blind to the hypotheses and conditions of the experiment. They engaged in an idea-generation contest. The participants were asked to submit a creative idea to repurpose unused household items—an old tennis racket and a garden hose (see Supplementary Note C for the survey materials). The participants were asked to submit one idea, and there was no time limit on the task. They entered their idea into an open-ended text box.

We preregistered to collect responses from 300 participants via Prolific on 26 June 2023 (https://aspredicted.org/7B2_WXQ). We requested 300 responses, and 298 complete responses were recorded at the end of data collection. This deviation was beyond our control and was due to some participants submitting the survey without finishing it entirely. We then applied the preregistered exclusion criteria, removing six participants who provided non-sensical responses and one participant who failed the attention check. This resulted in a final sample of

291 participants (mean age, 39.80 years; 44% female) for analysis. No statistical methods were used to predetermine sample sizes, but our sample sizes are larger than those reported in previous research[13,22].

Creativity ratings. External judges recruited from Prolific (n = 232; mean age, 44.53 years; 54% female) rated the creativity of each idea using the same six-item scale used in Experiment 1, and the creativity score for each idea was computed by averaging the six items as in the previous experiment (α = 0.93).

As a robustness check, three expert judges rated all ideas in terms of creativity, articulateness and verbosity, as in Experiment 1. See Supplementary Note G for the details.

Experiment 2B

This experiment used a three-level between-participants (ChatGPT-assisted, Human-only and ChatGPT-only) survey design. Prolific participants in the USA were randomly assigned to one of two conditions (ChatGPT-assisted or Human-only). They were blind to the hypotheses and conditions of the experiment. They engaged in an idea-generation contest. The participants were asked to submit one creative toy idea for a 7-year-old child using all three items; a paper bag, a leftover construction brick and an unused fan (see Supplementary Note C for the survey materials). The participants were asked to submit one idea, and there was no time limit on the task. Participants in the Human-only condition were instructed not to use any external websites for the idea-generation task; they simply typed in the ideas that they generated by themselves. Participants in the ChatGPT-assisted condition were instructed to use the ChatGPT website for the idea-generation task. Specifically, they were asked to interact with ChatGPT until they found a satisfactory response and then make more modifications as needed on the basis of the response from ChatGPT. These participants also reported how much they relied on this external resource during the idea generation. For the ChatGPT-only condition, we directly entered the same prompt given to the participants into ChatGPT-3.5, obtaining 100 responses by entering the prompt 100 times. We did not modify the content obtained from ChatGPT. We preregistered to post our survey on Prolific until we collected 200 responses meeting specific criteria (that is, passing an attention check and providing sensical responses) and to collect an additional 100 responses from ChatGPT-3.5 on 24 January 2024 (https://aspredicted.org/J7P_ZFN). The sample size was determined via a priori power analysis using G*Power to ensure 80% power to detect an effect size of d = 0.4 at the 5% α level. We concluded data collection upon reaching our target of 200 responses (mean age, 41.68 years; 52% female) from Prolific and 100 responses from ChatGPT-3.5.

Creativity ratings. External judges (n = 241; mean age, 42.27 years; 56% female) rated the creativity of each idea using the same six-item scale used in Experiment 1, and the creativity score for each idea was computed by averaging the six items as in the previous experiment (α = 0.95).

As a robustness check, three expert judges rated all ideas in terms of creativity, articulateness and verbosity, as in the previous experiments. Additionally, in this experiment, we asked the expert judges to rate whether the ideas were not new at all, incrementally new (that is, providing new combinations of existing ideas) or radically new (that is, going beyond mere new combinations to truly new ideas)[31,32].

Experiment 3

MTurk participants in the USA were randomly assigned to one of two experimental conditions: ChatGPT-assisted or Web-search-assisted. The participants were blind to the hypotheses and conditions of the experiment. They engaged in an idea-generation contest in which they were asked to come up with a creative idea for a dining table that did not exist on the market (see Supplementary Note C for the survey materials). The participants were asked to submit one idea, and there was no time limit on the task. They entered their idea into an open-ended text box. We collected 208 complete responses from MTurk. After we removed 1 participant who provided non-sensical responses and 13 participants who failed the attention check, 194 participants (mean age, 43.11 years; 50% female) were retained for analysis. No statistical methods were used to predetermine sample sizes, but our sample sizes are larger than those reported in previous research[13,22].

Creativity ratings. External judges (n = 158; mean age, 43.11 years; 50% female) rated the creativity of each idea using the same six-item scale used in Experiment 1 (α = 0.77). As a robustness check, three expert judges rated all ideas in terms of creativity, articulateness and verbosity. See Supplementary Note G for the details.

Process measures. After all participants submitted their ideas, we measured our first theoretical mediator (idea exposition aid) by asking them the degree to which idea exposition was possible through the use of ChatGPT (or Web search) (three items; α = 0.98). The survey also included other exploratory questions—including the extent to which ChatGPT (or Web search) helped them consider important aspects of a dining table (three items; α = 0.97), their perception of having a conversation with the chatbot (three items; α = 0.93) and their level of engagement in the creativity task (three items, α = 0.84)—as well as the PANAS affect scale. All of these items and their means are reported in Supplementary Note I.

To fully capture the process mechanism, we also asked the external judges to rate how articulate the idea seemed (idea articulateness) using five items (for example, "This idea is clearly presented", five items, α = 0.90; see Supplementary Note I for all items). This measure served as the second mediator. We ran a path analysis using the lavaan package[40] in R.

As a robustness check, three expert judges rated all ideas in terms of creativity, articulateness and verbosity. We later used these scores to run a serial mediation model and confirmed that the mediation results replicated as well. See Supplementary Note G for the details.

Experiment 4

Prolific participants in the USA were randomly assigned to one condition in the two (technology: ChatGPT-assisted or Web-search-assisted) by two (level of constraints: Low or High) between-participants design. They were blind to the hypotheses and conditions of the experiment. They engaged in an idea-generation contest in which they were asked to come up with a creative gift idea for a teenage girl (see Supplementary Note C for the survey materials). The participants were asked to submit one idea, and there was no time limit on the task. They entered their idea into an open-ended text box. We preregistered to collect responses from 400 participants via Prolific on 16 June 2023 (https://aspredicted.org/FYM_VRZ). We set the Prolific survey to allow 405 responses, anticipating some incomplete submissions. We received 405 complete responses. We then applied the preregistered exclusion criteria, removing five participants who provided non-sensical responses. This resulted in the intended sample of 400 participants (mean age, 38.72 years; 34% female) for analysis. No statistical methods were used to predetermine sample sizes, but our sample sizes are larger than those reported in previous research[13,22].

Creativity ratings. External judges (n = 321; mean age, 43.45 years; 51% female) rated the creativity of each idea using the same six-item scale used in Experiment 1 (α = 0.79). As a robustness check, three expert judges rated all submitted ideas in terms of originality and appropriateness, and these ratings were used to compute the creativity score.

Experiment 5

Prolific participants in the USA were randomly assigned to one condition in the two (technology: ChatGPT-assisted or Web-search-assisted)

by two (empathy: Baseline or High) between-participants design. The participants were blind to the hypotheses and conditions of the experiment. They engaged in an idea-generation contest in which they were asked to submit one idea to repurpose two unused items (an old flashlight and a hair spray; see Supplementary Note C for the survey materials). The participants were asked to submit one idea, and there was no time limit on the task. They entered their idea into an open-ended text box. We preregistered to collect 400 participants via Prolific on 21 June 2023 (https://aspredicted.org/TCP_W26). We requested 400 responses, but we received 394 complete responses due to some participants submitting the survey without finishing it entirely. We then applied the preregistered exclusion criteria, removing eight participants who provided non-sensical responses and three participants who failed the attention check. This resulted in a final sample of 383 participants (mean age, 40.24 years; 47% female) for analysis. No statistical methods were used to predetermine sample sizes, but our sample sizes are larger than those reported in previous research[13,22].

Creativity ratings. External judges (n = 310; mean age, 43.68 years; 59% female) rated the creativity of each idea using the same six-item scale used in Experiment 1 (α = 0.92). As a robustness check, three expert judges rated all ideas in terms of creativity, articulateness and verbosity, as in Experiment 1. See Supplementary Note G for the details.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data, including the secondary dataset and the experiments, can be found in the Open Science Framework at https://osf.io/rzn87/.

Code availability

The corresponding code can be found in the Open Science Framework at https://osf.io/rzn87/.

References

1. Hu, K. ChatGPT sets record for fastest-growing user base—analyst note. *Reuters* (3 February 2023); https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
2. Tong, A. Exclusive: ChatGPT traffic slips again for third month in a row. *Reuters* (7 September 2023); https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/
3. *GPT-4 Technical Report* (OpenAI, 2023); https://cdn.openai.com/papers/gpt-4.pdf
4. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
5. Stevenson, C., Smal, I., Baas, M., Grasman, R. & van der Maas, H. Putting GPT-3's creativity to the (alternative uses) test. In *Proc. 13th International Conference on Computational Creativity* (eds Hedblom, M. M. et al.) 164–168 (Association for Computational Creativity, 2022).
6. Góes, F., Volpe, M., Sawicki, P., Grzes, M. & Watson, J. Pushing GPT's creativity to its limits: alternative uses and Torrance Tests. In *Proc. 14th International Conference on Computational Creativity* (eds Pease, A. et al.) 342–346 (Association for Computational Creativity, 2023).
7. Summers-Stay, D., Voss, C. R. & Lukin, S. M. Brainstorm, then select: a generative language model improves its creativity score [Paper Presentation]. In *AAAI-23 Workshop on Creative AI Across Modalities* (2023).
8. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.12712 (2023).
9. Kaufman, J. C., Beghetto, R. A. & Watson, C. Creative metacognition and self-ratings of creative performance: a 4-C perspective. *Learn. Individ. Differ.* **51**, 394–399 (2016).
10. Plucker, J. A. & Renzulli, J. S. in *Handbook of Creativity* (eds Sternberg, R. J. & Renzulli, J. S.) 35–61 (Cambridge Univ. Press, 1999).
11. Sternberg, R. J. & Lubart, T. I. in *Handbook of Creativity* (eds Sternberg, R. J. & Renzulli, J. S.) 3–15 (Cambridge Univ. Press, 1999).
12. Burroughs, J., Moreau, C. P. & Mick, D. in *Handbook of Consumer Psychology* (eds Haugtvedt, C. et al.) 1011–1038 (Routledge, 2008).
13. Moreau, C. P. & Dahl, D. W. Designing the solution: the impact of constraints on consumers' creativity. *J. Consum. Res.* **32**, 13–22 (2005).
14. Runco, M. A., Illies, J. J. & Eisenman, R. Creativity, originality, and appropriateness: what do explicit instructions tell us about their relationships? *J. Creat. Behav.* **39**, 137–148 (2005).
15. Mednick, S. A. The associative basis of the creative process. *Psychol. Rev.* **69**, 220–232 (1962).
16. Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E. & Benedek, M. The roles of associative and executive processes in creative cognition. *Mem. Cogn.* **42**, 1186–1197 (2014).
17. Koestler, A. *The Act of Creation* (Hutchinson & Co., 1964).
18. Levin, I. Creativity and two modes of associative fluency: chains and stars. *J. Pers.* **46**, 426–437 (1978).
19. Benedek, M., Könen, T. & Neubauer, A. C. Associative abilities underlying creativity. *Psychol. Aesthet. Creat. Arts* **6**, 273–281 (2012).
20. Shneiderman, B. Creativity support tools. *Commun. ACM* **45**, 116–120 (2002).
21. Herring, S. R., Jones, B. R. & Bailey, B. P. Idea generation techniques among creative professionals. In *Proc. 2009 42nd Hawaii International Conference on System Sciences* (ed. Sprague, R. H.) 1–10 (IEEE, 2009).
22. Guzik, E. E., Byrge, C. & Gilde, C. The originality of machines: AI takes the Torrance Test. *J. Creat.* **33**, 100065 (2023).
23. Hubert, K. F., Awa, K. N. & Zabelina, D. L. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Sci. Rep.* **14**, 3440 (2024).
24. Haase, J. & Hanel, P. H. P. Artificial muses: generative artificial intelligence chatbots have risen to human-level creativity. *J. Creat.* **33**, 100066 (2023).
25. Mehta, R., Zhu, R. & Cheema, A. Is noise always bad? Exploring the effects of ambient noise on creative cognition. *J. Consum. Res.* **39**, 784–799 (2012).
26. Grosskopf, S. in *The Measurement of Productive Efficiency: Techniques and Applications* (eds Fried, H. O. et al.) 160–194 (Oxford University Press, 1993).
27. Coelli, T. J., Rao, D. S. P., O'Donnell, C. J. & Battese, G. E. *An Introduction to Efficiency and Productivity Analysis* (Springer Science & Business Media, 2005).
28. Jackson, F. Could ChatGPT REALLY slay Google? MailOnline looks at whether the trendy bot can put an end to internet giant's £120 billion dominance by revamping how we search the net. *Daily Mail* (11 May 2023); https://www.dailymail.co.uk/sciencetech/article-11723499/Could-ChatGPT-replace-Google-Experts-weigh-win-race-AI-search-engine.html
29. Heaven, W. D. Chatbots could one day replace search engines. Here's why that's a terrible idea. *MIT Technology Review* (29 March 2022); https://www.technologyreview.com/2022/03/29/1048439/chatbots-replace-search-engine-terrible-idea/

30. Lucas, B. J. & Nordgren, L. F. The creative cliff illusion. *Proc. Natl Acad. Sci. USA* **117**, 19830–19836 (2020).
31. Dewar, R. D. & Dutton, J. E. The adoption of radical and incremental innovations: an empirical analysis. *Manage. Sci.* **32**, 1422–1433 (1986).
32. Madjar, N., Greenberg, E. & Chen, Z. Factors for radical creativity, incremental creativity, and routine, noncreative performance. *J. Appl. Psychol.* **96**, 730–743 (2011).
33. Johnson, A. Here's what to know about OpenAI's ChatGPT—what it's disrupting and how to use it. *Forbes* (12 December 2022); https://www.forbes.com/sites/ariannajohnson/2022/12/07/heres-what-to-know-about-openais-chatgpt-what-its-disrupting-and-how-to-use-it/
34. Roose, K. How does ChatGPT really work? *New York Times* (28 March 2023); https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html
35. Lam, T. C. M., Kolomitro, K. & Alamparambil, F. C. Empathy training: methods, evaluation practices, and validity. *J. Multidiscip. Eval.* **7**, 162–200 (2011).
36. Kahriman, I. et al. The effect of empathy training on the empathic skills of nurses. *Iran. Red Crescent Med. J.* **18**, e24847 (2016).
37. Goldstein, A. P. & Michaels, G. Y. *Empathy: Development, Training, and Consequences* (Routledge, 2021).
38. Castelo, N., Bos, M. W. & Lehmann, D. R. Task-dependent algorithm aversion. *J. Mark. Res.* **56**, 809–825 (2019).
39. Longoni, C. & Cian, L. Artificial intelligence in utilitarian vs. hedonic contexts: the "word-of-machine" effect. *J. Mark.* **86**, 91–108 (2022).
40. Rosseel, Y. lavaan: an R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36 (2012).

Acknowledgements
The authors received no specific funding for this work.

Author contributions
B.C.L. and J.C. engaged in idea generation, designed and performed research, analysed data and wrote the paper together.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01953-1.

Correspondence and requests for materials should be addressed to Byung Cheol Lee.

Peer review information *Nature Human Behaviour* thanks Fabricio Góes, Max Kreminski and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Byung Cheol Lee & Jaeyeon (Jae) Chung

Last updated by author(s): Jun 13, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Studies were programmed in Qualtrics and administered using Mechanical Turk and Prolific for online samples.
Data analysis	Data were analyzed in R (4.2.1) using the following packages: lavaan (0.6.12), tidyverse (1.3.2), emmeans (1.7.5), irr (0.84.1), rstatix (0.7.0), kableExtra (1.3.4). All codes used to analyze data can be found in the Open Science Framework (OSF) at: https://osf.io/rzn87/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data can be found in the Open Science Framework (OSF) at: https://osf.io/rzn87/

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Participants were not specifically sampled based on sex or gender.
Reporting on race, ethnicity, or other socially relevant groupings	Participants were not specifically sampled based on race, ethnicity, or other groupings.
Population characteristics	Participants aged 18 or older were drawn from a general population research participant pool (MTurk and Prolific).
Recruitment	Participants were recruited from Mturk and Prolific via a listing that described the study as "Study on Social Science." It is possible that individuals with a strong interest in social science were more likely to participate. However, since participants were randomly assigned to different conditions, we believe that this potential bias did not significantly affect our results.
Ethics oversight	This research was approved by the Institutional Review Board at Rice University.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	All studies were experiments. Participants were randomly assigned to either use ChatGPT or not use ChatGPT while completing a task that demand creativity. Data included open-ended text responses and quantitative responses, including Likert scales and numerical responses.
Research sample	We used sample of participants with accounts on Mechanical Turk or Prolific. This sample was not representative of overall population (Experiment 1: Mage = 42.29, 52% female, Experiment 2A: Mage = 39.80, 44% female, Experiment 2B: Mage = 41.68, 52% female, Experiment 3: Mage = 43.11, 50% female, Experiment 4: Mage = 38.72, 34% female, Experiment 5: Mage = 40.24, 47% female). All samples were convenience samples.
Sampling strategy	All samples were convenience samples. Participants self-selected into online studies. Researchers had no control over selecting participants other than specifying data-quality restrictions.
Data collection	In Experiments 1, 2A, 3, 4, and 5, no statistical methods were used to pre-determine sample sizes but our sample sizes were chosen to be similar or larger than those reported in previous research. In Experiment 2A, the sample size was determined via a priori power analysis using G*Power to ensure 80% power to detect an effect size of d = .4 at the 5% alpha level. In Experiments 2A, 2B, 4, and 5, we pre-registered the data collection procedure. All studies were fully double-blinded. All instructions and responses were programmed in Qualtrics. No element of the study involved face-to-face interaction with the experimenter.
Timing	Data were collected in 2023 and 2024: Experiment 1 (Feb 2023), Experiment 2A (Jun 2023), Experiment 2B (Jan 2024), Experiment 3 (May 2023), Experiment 4 (Jun 2023), Experiment 5 (July 2023)
Data exclusions	Participants were excluded from analysis if they failed attention check or provided non-sensical responses. This resulted in removing 27 participants in Experiment 1, 7 in Experiment 2A, 14 in Experiment 3, 5 in Experiment 4, and 11 in Experiment 5.
Non-participation	Participants who declined consent or did not complete the study to the end were dropped from the dataset. In Experiment 1, 1021 individuals clicked the survey link and 260 finished the survey. In Experiment 2A, 395 individuals clicked the survey link and 298 finished the survey. In Experiment 2B, 356 clicked the survey link and 216 finished the survey. In Experiment 3, 479 clicked the survey link and 208 finished the survey. In Experiment 4, 542 clicked the survey link and 405 finished the survey. In Experiment 5, 554 clicked the survey link and 394 finished the survey.
Randomization	Participants were randomly assigned to conditions via Qualtrics randomizers. The researchers had no control over participant randomization.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.