

DISCUSSION PAPER SERIES

IZA DP No. 17302

**Creative and Strategic Capacities of
Generative AI:
Evidence from Large-Scale Experiments**

Noah Bohren
Rustamdjan Hakimov
Rafael Lalive

SEPTEMBER 2024

DISCUSSION PAPER SERIES

IZA DP No. 17302

Creative and Strategic Capacities of Generative AI: Evidence from Large-Scale Experiments

Noah Bohren

University of Lausanne

Rustamdjan Hakimov

University of Lausanne

Rafael Lalive

University of Lausanne and IZA

SEPTEMBER 2024

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Creative and Strategic Capacities of Generative AI: Evidence from Large-Scale Experiments*

Generative artificial intelligence (AI) has made substantial progress, but some capabilities of AI are not well understood. This study compares the ability of AI to a representative population of US adults in creative and strategic tasks. The creative ideas produced by AI chatbots are rated more creative than those created by humans. Moreover, ChatGPT is substantially more creative than humans, while Bard lags behind. Augmenting humans with AI improves human creativity, albeit not as much as ideas created by ChatGPT alone. Competition from AI does not significantly reduce the creativity of men, but it decreases the creativity of women. Humans who rate the text cannot discriminate well between ideas created by AI or other humans but assign lower scores to the responses they believe to be AI-generated. As for strategic capabilities, while ChatGPT shows a clear ability to adjust its moves in a strategic game to the play of the opponent, humans are, on average, more successful in this adaptation.

JEL Classification: I24, J24, D91, C90

Keywords: artificial intelligence, ChatGPT, Bard, creativity, experiment

Corresponding author:

Rafael Lalive
Department of Economics
HEC University of Lausanne
CH-1015 Lausanne
Switzerland
E-mail: Rafael.Lalive@unil.ch

* We thank David Autor, David Strömberg, Mathias Thoenig, and seminar audiences at University of Lausanne for comments on earlier versions of this paper. We gratefully acknowledge financial support from the Swiss National Science Foundation (Project number 189152), and Swiss Research Programme 77. The project received IRB approval from the LABEX Ethic Committee of HEC, University of Lausanne (CASTING,14.6.23) and was pre-registered on AEA Registry, AEARCTR-0011584.

1 Introduction

Intelligence and creativity are fundamental human capabilities and are strongly interrelated. Humans are thought to possess practical intelligence, analytical intelligence, and creative intelligence (Sternberg, 1985), and different forms of creativity, such as combinatorial creativity, exploratory creativity, and transformational creativity (Boden, 1998). To date, automation has targeted routine tasks with well-defined sequences of operations (Autor and Dorn, 2013; Autor, 2015), rarely making inroads into creative domains. However, contemporary generative artificial intelligence (AI) systems, like ChatGPT and DALL-E, are capable of generating original textual and visual content, challenging conventional perspectives on the domains of automation (Smith and Anderson, 2014). While the mechanism of language models involves predicting word sequences based on inquiry, the limits of their capabilities and expertise remain unclear.

Existing research indicates substantial gains in productivity when implementing generative AI for customer efficiency (Brynjolfsson et al., 2023), professional writing (Shakked and Whitney, 2023), or in legal services (Felten et al., 2023), but these tasks require relatively little creativity. Also, humans can be averse to algorithm adoption, even when it leads to significant gains (Dietvorst et al., 2018; Dargnies et al., 2023), but it remains unclear whether these findings generalize to creativity, where humans judge the output. In terms of strategic thinking, firms have begun to delegate certain strategic decisions to AI systems (Abada and Lambin, 2023), and to evaluate strategic decisions made by firms (Kiron and Schrage, 2019). However, whether large language models (LLMs) such as ChatGPT display strategic reasoning is not well understood.

In this pre-registered study, we explore whether generative AI surpasses human performance in tasks that measure creativity. We also examine whether human creativity improves when individuals have access to generative AI and how they react to competition with AI in creative tasks. Given that the judges of the quality of the text are also humans, we investigate whether they can distinguish between AI-generated and human-generated creative outputs and, if so, whether their perception of the quality differs. Finally, we explore whether generative AI can dynamically adapt its strategy against an opponent, evaluating AI’s performance in strategic games against opponents following either an equilibrium strategy—likely familiar to the AI through its training data—or a non-equilibrium strategy, which requires the AI to adapt its responses based on the observed actions of the opponent.

We conducted a large-scale experiment with over 4,000 participants. First, we asked more than 1,000 humans and two generative AI chatbots, ChatGPT and Bard, to generate creative answers to open creativity tasks. We define creativity as “the ability to come up with new ideas that are surprising yet intelligible, and also valuable in some way” (Boden, 2001). We implemented a form of ‘open’ or ‘divergent’ creativity, i.e., one with no pre-defined solution, asking participants to “describe a town, city, or society in the future” or “if you had the talent to invent things just by thinking of them, what would you create?” (Guilford, 1975; Charness and Grieco, 2019). Both humans and AI chatbots were prompted

in the same way and provided a short text describing the answer. Human participants were compensated with a standard amount for participating in the study and could earn a substantial premium if their idea ended up being among the top 10% most creative responses, as judged by other participants. In the Baseline treatment, participants received no additional information; in the HumanPlusAI treatments, participants had access to an AI chatbot (either Bard or ChatGPT) and were instructed to use them when generating the answers; and in the HumanAgainstAI treatment, participants were informed that they were competing against not only humans but also AI chatbots.

Subsequently, more than 3,000 humans, split into three groups, evaluated the creativity of these texts to determine whether they found the ideas novel, surprising, and useful. We varied the information given to human raters to assess the factors influencing their ratings. In the Baseline, the raters evaluated original versions of human text mixed with AI-generated texts. To avoid the influence of grammar on the ratings, in the CorrectedRaters treatments, raters evaluated grammar-corrected texts. To investigate perceptions of AI-generated texts and the ability to identify them, in the AIRaters treatment, raters were informed that the text could have been generated by AI and asked to guess which texts were AI-generated.

Our first finding indicates that AI chatbots exhibit greater creativity than humans, but there are significant differences in creativity across AI chatbots; ChatGPT is significantly more creative than humans, and the difference is large. Bard is significantly less creative than humans. Human creative ability varies considerably, but ChatGPT’s ideas consistently score in the upper part of the human distribution of creative ability. The superior performance of ChatGPT holds true even for the best ideas: among 17 best-rated answers (top 1%), 8 are generated by ChatGPT, 3 by Humans with access to ChatGPT, and 6 by humans. Humans with access to generative AI are significantly more creative than humans without access to the technology. This increase in productivity is similar to results in different contexts, like writing tasks ([Shakked and Whitney, 2023](#)), although the size of productivity increase is relatively small compared to other contexts. Most surprisingly, augmented humans are significantly less creative than ChatGPT alone. This holds true even when considering humans who used the ChatGPT-4 model. We conjecture that the prompting drives this result. In line with evidence from [Girotra et al. \(2023a\)](#), prompts greatly influence the output. We directly asked ChatGPT to produce creative and novel answers, while simply asking to the questions of the creative task, what participants likely did, might lead to less creative answers by ChatGPT. Competition from AI marginally diminishes human creativity, but the effects of competition are small compared to the benefits of augmentation. However, the effect of competition is significant for female participants, consistent with our pre-registered hypothesis regarding gender differences in response to competition and competitiveness ([Gneezy et al., 2003](#); [Niederle and Vesterlund, 2011](#); [Saccardo et al., 2018](#)), especially when competing with men. Our experiments show that this effect extends to competition with AI as well.

An often underappreciated aspect of creativity is the ability of an agent or group to generate a wide range of distinct ideas. While AI may demonstrate higher creativity than

humans, its impact is limited if it only produces variations of a single concept. After analyzing the raters' evaluations, we use embeddings to assess the diversity of the generated content. Our analysis reveals that both ChatGPT and humans produce texts with comparable idea diversity, while Bard's outputs are notably repetitive. However, in the case of the most creative texts, humans outperform ChatGPT by generating a higher number of unique ideas, consistent with the findings of [Doshi and Hauser \(2023\)](#).

A unique feature of our study is that we consider the raters' side and examine the determinants of creativity ratings. Reassuringly, we find no treatment difference in creativity ratings of grammar-corrected and original text. When raters know that some of the texts might be generated by AI, they significantly lower scores for the texts they believe to be AI-generated. This is a novel phenomenon similar to algorithm aversion ([Dietvorst et al., 2015](#)), which might be a behavioral constraint to the adoption of AI. However, unlike previous contexts of algorithm aversion, this tendency does not affect the performance of the AI, and ChatGPT texts still greatly outperform humans, even when only the ratings of raters who knew that some texts could be AI-generated are considered. This is because raters are surprisingly bad at distinguishing AI and human-generated texts. They correctly classify the ideas of humans in 65% of cases but are significantly less able to correctly identify chatbots' responses as AI-generated (58% for ChatGPT and 35% for Bard).

To measure the strategic capabilities of AI and compare them to humans, after generating creative texts, we asked human participants to play a rock-paper-scissors game against an opponent for 24 rounds, knowing that the moves of the opponent were pre-determined.¹ They were incentivized to win as many rounds as possible. We also asked ChatGPT to play the same game, with each of the 24 rounds being conducted within one chat window, one by one. Every player (or ChatGPT chat) was assigned either to the Balanced treatment, where the opponent played an equilibrium strategy playing each move with 33.3% probability, or to the Unbalanced treatment, where the opponent never played scissors and randomized between rock and paper. If the player is strategic, they should adjust their moves to the biased play of the opponent. Thus, a strategic player will learn not to play rock, as it never brings a win in the Unbalanced treatment.

Our findings reveal that both humans and AI, on average, adjust their play to the biased opponent. In line with our pre-registered hypotheses, the number of times rock is played in the last 12 rounds is significantly lower in the Unbalanced than the Balanced treatment, both for humans and for AI. The evidence that ChatGPT adjusts and learns within a chat is novel and first in this context to the best of our knowledge. This is a first sign of intelligence, as the play of the opponent is newly generated and could not be trained, unlike the play in the Balanced treatment, where equilibrium play is common knowledge. Interestingly, humans manage to earn significantly more points than AI in the Unbalanced treatments, as instead of playing 50% scissors and 50% paper as AI does, they play paper significantly more often, which is an undominated move if one believes that the opponent cannot counteract the monotonic play of the player. Thus, while we observe

¹We indeed pre-drew 24 moves of the opponent, as described below, and asked our research assistants to strictly follow the pre-drew sequence.

signs of strategic skills in AI, we conclude that at the moment, humans have an edge over AI in this context.

Understanding the competencies of generative AI holds multifaceted importance. First, it enables industries to distinguish tasks suitable for automation from those requiring human intervention, thus optimizing productivity (Arntz et al., 2016). Second, insights into AI capabilities can potentially inform workforce development and upskilling strategies (Dignum, 2019). Third, insights on how humans react to generative AI can guide ethical considerations, ensuring responsible AI deployment (Dignum, 2019). Our findings on strategic thinking suggest that the capacity of LLMs, trained on text and image data, partially extends to learning off-equilibrium strategies; however, their proficiency is not yet comparable to that of humans. This insight can inform the adoption of LLMs for strategic decision-making.

Most related studies to ours include Charness and Grieco (2024). The authors ran a 2x2 experiment varying the task (open or closed) and whether the inputs were generated by humans or ChatGPT. Their open task is the same as ours. They invited raters from Prolific to rate the answers. Their results contradict ours, as their AI-generated text received lower ratings than human texts. Even though we use the same pool of raters, there are several substantive design and implementation differences. Upon examining the ratings, we see that the difference comes from ratings assigned to the AI-generated text, while ratings assigned to human outputs are comparable. We conjecture that the difference arises from prompting and the version of ChatGPT employed. Our prompt directly explains the task and incentivizes ChatGPT to produce the most novel and creative answers, while their prompt just asks to answer the question. Furthermore, we employed ChatGPT-4, whereas they used ChatGPT-3.5. This aligns with evidence emphasizing the importance of prompts and supports our explanation of why humans and AI underperform relative to AI alone. Additionally, our papers differ in research questions. While they study differences in closed and open creativity, we are interested in open creativity only, the complementarity of skills of humans and AI in open creativity, and the reaction of judges to potentially AI-generated text. Furthermore, we explore the strategic skill of AI. Girotra et al. (2023b) also test the creativity of ChatGPT relative to humans in the context of product ideas and find that AI outperforms the students of an elite university. They vary the prompt and show that it marginally increases ratings of ChatGPT ideas. Despite the difference in the tasks, our results point to similar direction of dominance of ChatGPT over humans in creative tasks.

Also related, Doshi and Hauser (2023) show a tradeoff between the quality and diversity of the ideas generated by ChatGPT: while AI-enabled stories are rated higher, they are more similar to each other than stories by humans alone. This is similar to our evidence; however, we still show that ChatGPT generates more unique ideas overall. The question remains where the limit in the total quantity of ideas generated by ChatGPT lies. Given the evidence, the variety of prompting might also contribute to the variance in the answers. Related, Girotra et al. (2023a) show that prompts can increase the diversity of ideas in the context of ideas for new products, with the chain of thought method leading to the highest diversity.

A large number of papers study the impact of generative AI on the productivity of workers in different contexts. [Shakked and Whitney \(2023\)](#) show that access to ChatGPT improves the productivity of educated workers for writing tasks. [Dell’Acqua et al. \(2023\)](#) show that AI enhances the productivity of consultants of Boston Consulting Group, especially in their areas of expertise. We complement these papers by showing that ChatGPT can indeed enhance human creativity. Surprisingly, the effects are much lower than in other contexts, and humans plus ChatGPT perform worse than ChatGPT alone, raising the issue of necessary priming experience.

Another strand of literature compares the output of LLMs to humans. [Chen et al. \(2023\)](#) compare the rationality of LLM output to humans in the context of risk, time, social, and food decisions, showing that GPT’s decisions are mostly rational and even score higher than human decisions. [Gilardi et al. \(2023\)](#) show that ChatGPT outperforms crowd workers in text annotation tasks based on various tweets and newspaper articles. [Huang et al. \(2023\)](#) and [Kuzman et al. \(2023\)](#) make similar conclusions without direct tests against humans. We complement this literature by showing that LLMs can both outperform and underperform, depending on the model. Moreover, LLMs can learn to best respond to human actions, even if the actions are out of equilibrium.

Finally, given that our study also focuses on the causal impact of treatments on judges’ ratings, we relate to the literature on algorithm aversion ([Dietvorst et al., 2015](#)). Generally, algorithm aversion is a tendency to avoid AI-driven decisions or outputs, documented in various contexts like financial decisions ([Dietvorst et al., 2018](#)), hiring ([Dargnies et al., 2023](#)), prediction tasks ([Greiner et al., 2024](#)), redistributive decisions ([Chugunova and Luhan, 2024](#)) and others. In our context, algorithm aversion is the tendency to rate AI-generated answers more stringently. Most surprisingly, this still leads to a large out-performance of ChatGPT ideas because judges are very bad at guessing which ideas are AI-generated.

The rest of the study is organized as follows: Section 2 presents the experimental design, Section 3 discusses the results of the experiment, and Section 4 concludes.

2 Experimental Setting

The design of the experiment has two goals. First, it examines the creative and strategic capabilities of humans in comparison to two prominent AI chatbots: ChatGPT-4 and Bard. Second, it investigates how human subjects react to competition from AI, and how they judge AI-generated texts.

To simplify the exposition, we first present the experimental design for the creativity task and then for the strategic task.

2.1 Creativity task

Participants were either writers, engaging in creative tasks, or raters, ranking the players' responses. Our treatment variations for writers and raters differ. We will present the design subsequently.

2.1.1 Creativity:Writers

We recruited 1251 participants from the U.S. through the Prolific platform to participate in a divergent (unconstrained) creative task (Charness and Grieco, 2019). The task was to create a text of up to 1000 characters (around 150 words) that was as creative as possible. The maximum time for the task was 10 minutes. They were offered a choice between two prompts: "If you had the talent to invent things just by thinking of them, what would you create?" or "Imagine and describe a town, city, or society in the future.". Participants received £2 base payment for participation, conditional on writing any text. Participants were informed that their submissions would be evaluated for creativity by subsequent participants. To incentivize creativity, participants with texts ranked in the top 10% of most creative by other subjects received a £5 bonus, and they were aware of this incentive.

In May 2023, we gave the same task to ChatGPT-4 and Bard. We prompted both AIs with the following instruction in isolated chats to avoid repetitions: "Give 4 alternative and creative answers to the following question within 1,000 characters for each answer," using one of the two prompts. This yielded 216 unique responses from Bard and 224 from ChatGPT-4. We accessed the AIs through their standard chat interface without adjusting parameters like temperature.

Human participants were randomly assigned to one of three treatments:

1. **HumanBaseline** (688 participants): Participants in this group generated creative answers autonomously and the top 10% most creative humans receive a fixed bonus of £5.
2. **HumanAgainstAI** (253 participants): Participants in this group also generated their responses independently. However, they were aware that their entries would be compared not just with those from other individuals but also with texts produced by AI. If their submission ranked in the highest 10% among all entries, including both human and AI-generated texts, they would receive a £5 bonus.
3. **HumanPlusAI** (309 participants): Participants in this group had the choice to utilize Bard or ChatGPT (3.5 or 4) for help in producing their responses.² In the description on Prolific, we explicitly requested participants who already possessed an OpenAI or Bard account. Before starting their submissions, they had to declare

²AI used: Bard (39), ChatGPT-3.5 (184), ChatGPT-4 (86)

which AI tool (ChatGPT-3.5, ChatGPT-4, Bard, or None) they would employ for assistance in creating their entries. The top 10% most creative humans receive a fixed bonus of £5.

We conducted a follow-up survey 2 months after the main experiment. The response rate was 79.6% (the average creativity of those who responded and those who did not was not statistically different ($P > 0.32$)). We collected data on their income, employment status, whether the task was transferable skill in their actual job, as well as the self-evaluation of their creativity and the creativity of their job (current job for employed and "dream job" for unemployed).

2.1.2 Creativity:Raters

Subsequently, 3336 new participants were recruited to evaluate the text produced by humans and AI, each received a fixed compensation of £2. Their task was to rate a random selection of 20 responses. Creativity is a multidimensional concept, (Boden, 1998) identifies three sub-features that make an idea creative: "A creative idea is one which is novel, surprising, and valuable (interesting, useful, beautiful, etc.)". In order to measure those characteristics, participants were instructed to assign a score from 0 to 10 for each text to each of those questions: (1) In general, how creative was this response? (2) To what extent is this response new or original? (3) How surprised were you by this response? (4) How would you rate the usefulness of this response?

The raters were divided into three treatment groups

1. **BaselineRaters** (1100 participants): This group was tasked with evaluating 20 responses, which could be from either human participants or AI.
2. **CorrectedRaters** (1090 participants): To minimize the impact of presentation and grammatical accuracy on the evaluation, we employed ChatGPT to automatically refine the human-generated texts, correcting grammar and typographical errors. Consequently, the CorrectedRaters group evaluated a set of 20 responses that included both unaltered AI-produced texts and human-produced texts that had been revised.
3. **AIRaters** (1146 participants): Operating in the same context as CorrectedRaters, this group assessed 20 responses, now being aware that some were generated by AI. Subsequently, they were tasked with discerning whether each response originated from a human or an AI. To incentivize accurate identification, participants who correctly distinguished the origin of 15 or more texts were rewarded with an additional bonus of £0.5.³

³Note that in other treatments, participants were not informed that some of the text was generated by AI. However, this is not deceptive, as the task was solely to assess the creativity of the texts, without making any claims about the authorship of these texts.

Each text was rated by online raters at least 10 times in all three experimental conditions.⁴

Finally, a secondary group of five research assistants (2 students in an art school in Lausanne and 3 from business schools) also evaluated the responses in the same condition as CorrectedRaters.

2.2 Strategic task

Strategic skill is multidimensional, and we do not claim our task captures it comprehensively. The idea behind the task was to set up a scenario to measure how well people and AI adapt to changing, potentially out-of-equilibrium opponent strategies. Two key features guiding our task choice are:

1) The equilibrium play should be easy for everyone. 2) The best response should depend on the ability to learn within the game.

The first feature levels the playing field between the general population and AI, which may have access to optimal play descriptions in their training data. The second is crucial as it measures the ability of humans and AI to adapt to specific opponents.

We chose the well-known and easy-to-explain game of Rock-Paper-Scissors, played for 24 rounds. The first condition is met because it's common knowledge that the game is based on chance, requiring players to randomly choose moves. The second condition is met by systematically biasing the opponent's moves, which will be the focus of our treatment variation.

All participants involved in the writing task also took part in 24 rounds of Rock-Paper-Scissors against a computer opponent.⁵ Points were awarded based on the game outcomes (1 point for a win, 0.5 for a draw per round) and were later converted into monetary bonuses. Participants were divided into two random groups:

- **Balanced:** The computerized opponent employed an equilibrium strategy, randomizing moves with equal probability.
- **Unbalanced:** The opponent's choices were restricted to 'rock' and 'paper'.

Players received the following instructions "For this task, you will play 24 rounds of Rock, Paper, Scissors against a human. Their strategy is predetermined for all 24 rounds

⁴Every time, a random selection of 20 texts was made for each rater. After every 100 raters, we calculated how many ratings each text had. Once a text received 10 ratings, it was excluded from further randomization. This ensured that all texts received at least 10 ratings. Out of the 3,424 raters we hired, 81 raters gave fewer than 10 ratings. We excluded all their ratings from the analysis. Additionally, 18 individuals participated twice. Of these, 7 participated in the same treatment both times, contributing a total of 40 ratings, which were also removed from the analysis. The remaining 11 raters participated in two different treatments; for them, we excluded the ratings from their second participation. Consequently, 0.16% of texts received only 8 ratings, while 2.5% of texts received only 9 ratings.

⁵The order of the tasks was randomized such that half of the participants started with the writing task and half started with the strategic task.

and will be played out by the computer. You will earn 1 point for a win, 0.5 points for a tie, and 0 points for a loss. Each point is worth £0.2, which will be paid as a bonus.”

The same game was played by ChatGPT4, all 24 rounds within a separate chat. We predetermined the moves of the opponent of ChatGPT-4 in each round and each game. For the balanced treatment, we randomly drew moves from rock, paper, and scissors. For the unbalanced treatment, we only randomly drew moves from rock or paper. The initial prompt was as follows ”Let’s play 24 rounds of Rock, Paper, Scissors. I have my moves fixed for all 24 rounds and will reveal them to you honestly after each round, so you can potentially adjust your strategy to win the most rounds. Note that your goal should be to win as many rounds as possible. What is your first move?” Then, for each rounds, we used the prompt ”For round [n], I choose [*random selection*]”. We then collected the moves the AI selected. In total 200 games were played in the different chats.

Note that in the Balanced treatment, the equilibrium strategy is to randomize with equal probability among three possible moves. This should be known to ChatGPT4 as the strategy is discussed in many sources. In the Unbalanced treatment, ChatGPT4 would have to adapt to the biased play of the opponent. Note that the ”Rock” move becomes weakly dominated, as it never leads to a win. The ”Paper” move becomes a dominant strategy as it never leads to a loss. We pre-registered two measures to evaluate the strategic skill:

1. Frequency of suboptimal ’rock’ choices in the last 12 rounds.
2. Points accumulated in the last 12 rounds.

Differences in treatments between ’Balanced’ and ’Unbalanced’ scenarios allowed us to compare the strategic skills of human participants to ChatGPT4.

Comprehensive experimental instructions and additional data supporting our conclusions are available in the supplementary materials.

3 Results

3.1 Creativity Task: Writers

3.1.1 Scores by Online Raters

We start by descriptive statistics of responses and writers in the creativity task reported in Table 1

Table 1: Descriptive Statistics of Responses and Writers

Source	N	Avg Length (chars)	% Prompt A	Avg Time (min)	% Male	Avg Age
ChatGPT4	224	589	50%	-	-	-
Bard	216	400	50%	-	-	-
HumanBaseline	687	716	44.3%	5.98	49%	44
HumanAgainstAI	252	739	45.8%	6.67	48%	43.9
HumanPlusAI	309	860	43%	5.55	50.2%	44.7

Avg Length (chars) = Average number of characters

Prompt A = "If you had the talent to invent things just by thinking of them, what would you create?"

Prompt B = "Imagine and describe a town, city, or society in the future"

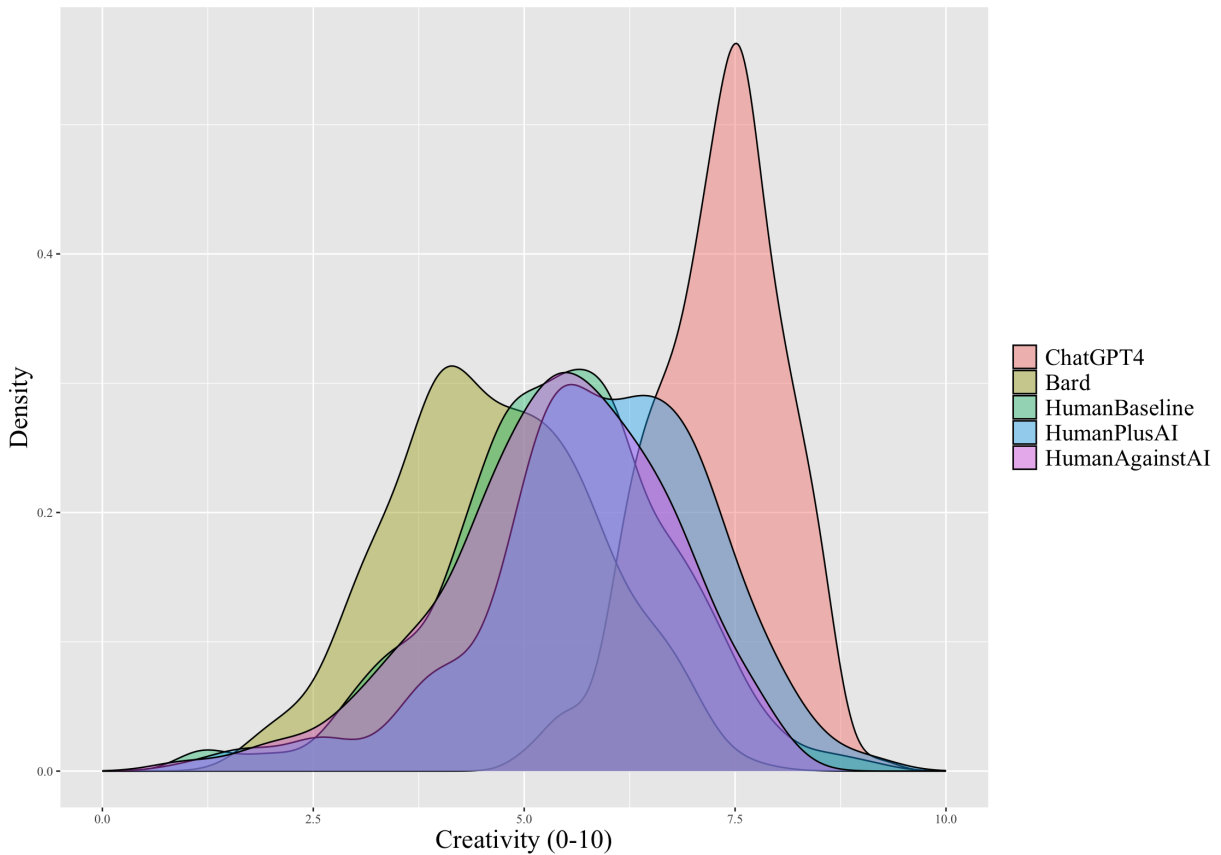
AI-generated texts are significantly shorter than those produced by the HumanBaseline group ($P < 0.001$),⁶ whereas texts generated by the HumanPlusAI are significantly longer ($P < 0.001$). Furthermore, we observe statistically significant differences in the time required to produce texts: the HumanPlusAI group required less time compared to the baseline ($P < 0.05$) whilst HumanAgainstAI required more time ($p < 0.01$). No other differences reached statistical significance. Overall, despite flat incentives and only a 10% chance of a bonus, we observed that our participants engaged with the task. To illustrate the quality and range of the responses, we present the median and best-rated responses for each treatment in the appendix (Table 16).

Turning to the treatment comparison of creativity, Figure 1 illustrates the distribution of creativity ratings across sources, as evaluated by online raters.⁷ Notably, Bard generated the least favorably-rated responses, whereas ChatGPT4 produced the highest-rated responses. Human-generated responses occupied an intermediate position. When humans utilized generative AI tools, the quality of creative responses improved yet did not surpass that of ChatGPT4.

⁶Unless stated otherwise, we use p-values from the regressions, controlling for rater fixed effects. For comparison with the baseline, we use p-values from regressions directly, and for comparisons between other treatments, we use the F-test.

⁷The main treatment differences are robust to the pooling of the ratings from other treatments.

Figure 1: Creativity ratings by sources



Distribution of creativity ratings by all raters.

Table 2 presents a regression analysis for creativity ratings controlling for raters’ fixed effects and with standard errors clustered at the level of responses. Model (1) presents treatment differences, using HumanBaseline as a reference group. All treatment differences—between Bard, HumanBaseline, HumanPlusAI, and ChatGPT4—are statistically significant ($P < 0.001$). Model (2) shows that there are no significant differences in treatment effects compared to Model (1) when controlling for the type of prompt (invention of an object or description of a future civilization) that the responses addressed. The awareness of competing against AI had no significant impact on the creativity of human responses on average. A pre-registered analysis for the heterogeneous treatment effects of HumanAgainstAI with respect to gender revealed a significant negative effect on creativity for women’s responses ($P < 0.05$), while no significant effects were found for men’s responses (see Model (3)). These treatment differences persist when controlling for individual characteristics of responders (see Models (4) and (5)). On the flip side, Models (6)-(8) show that women are able to generate more creative content compare to men in the HumanPlusAI condition.

Table 2: Creativity ratings by online raters

Dependent Variable: Model:	Creative Rating							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-0.924*** (0.051)	-0.936*** (0.050)	-0.936*** (0.050)			-0.936*** (0.050)		
ChatGPT4	1.879*** (0.039)	1.864*** (0.039)	1.864*** (0.039)			1.864*** (0.039)		
HumanPlusAI	0.500*** (0.052)	0.500*** (0.052)	0.500*** (0.052)	0.500*** (0.052)	0.474*** (0.052)			
HumanPlusAI (Female)						0.565*** (0.065)	0.626*** (0.069)	0.581*** (0.068)
HumanPlusAI (Male)						0.435*** (0.073)	0.394*** (0.077)	0.388*** (0.076)
HumanPlusAI (Other)						0.495** (0.242)	0.003 (0.290)	-0.060 (0.288)
HumanAgainstAI	-0.032 (0.054)	-0.037 (0.054)				-0.037 (0.054)	-0.044 (0.054)	-0.047 (0.053)
HumanAgainstAI (Female)			-0.152** (0.073)	-0.147* (0.075)	-0.133* (0.073)			
HumanAgainstAI (Male)			0.075 (0.071)	0.079 (0.075)	0.054 (0.074)			
HumanAgainstAI (Other)			0.179 (0.213)	-0.248 (0.267)	-0.179 (0.287)			
Rater Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Prompt	Prompt	Prompt Age Gender	ALL	Prompt	Prompt Age Gender	All
Observations	63,812	63,812	63,812	47,033	46,845	63,812	47,033	46,845
R ²	0.391	0.392	0.393	0.364	0.373	0.392	0.364	0.373
Within R ²	0.100	0.101	0.102	0.011	0.025	0.101	0.011	0.025

Notes: OLS regression of creativity ratings by online raters with raters fixed effects. All controls include answers to the questionnaire comprising ten questions on creative and cognitive style and sensation-seeking behavior, based on questions by Nielsen, Pickett, and Simonton (2008) on creative style and Zuckerman et al. (1964) on sensation-seeking attitude, demographic queries concerning sibling count, birth order, handedness, and parental marital status, six queries about past involvement in creative activities (Hocevar, 1980), a non-incentivized measure of risk preferences (Dohmen et al., 2009), and categorical controls for major. Standard errors are clustered on the response level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

One surprising outcome is that responses generated under the HumanPlusAI condition are significantly less creative than those generated solely by ChatGPT-4. An initial hypothesis might be that this discrepancy is driven by participants who had access to Bard, rather than ChatGPT. Indeed, upon close examination, we find that the average creativity score for responses generated by humans with access to ChatGPT (3.5 or 4.0) is significantly higher than those having access to Bard ($P < 0.01$) (Table 3). However, even then, the responses from ChatGPT-4 itself remain significantly more creative than responses from humans utilizing ChatGPT (3.5 or 4.0) ($P < 0.001$). This puzzling finding raises questions about the interplay between human guidance and AI capabilities. One plausible explanation could be that the type of prompting from human users affected the creative output. Specifically, the AI’s creativity might have been con-

strained or directed in such a way that it failed to fully utilize its capabilities, especially given that our prompt explicitly called for novel and creative responses.

Table 3: Average creativity scores by source and AI used

Source	Avg Creativity	N texts
<i>Bard</i>		
Bard	4.46	216
<i>ChatGPT4</i>		
ChatGPT (4.0)	7.24	224
<i>HumanPlusAI</i>		
Bard	5.57	39
ChatGPT (3.5)	5.93	184
ChatGPT (4.0)	5.84	86

Next, we turn to analyses of the subdimensions of creativity. Table 12, in the appendix, presents statistical analyses for three sub-dimensions of creativity: originality, surprise, and usefulness. In all three dimensions, ChatGPT-4 outperforms all other treatments significantly ($P < 0.001$). The largest difference between ChatGPT-4 and the HumanBaseline is in the dimension of originality with surprise being a close second—remarkably, these are the dimensions one would least expect from an AI that generates responses based on trained data. However, the second AI chatbot, Bard, scores worse on originality and surprise than the human baseline, indicating that the creative forces at play between ChatGPT-4 and Bard differ enormously. The dimension of usefulness drives the results of lower performance by female participants in competition with AI.

One concern might be that some participants rushed through the writing task to increase their earnings per hour, potentially compromising our measure of human creativity. Our analysis reveals that participants spent an average of 6 minutes on the task. Figure 2 indicates that only 14.6% of participants completed their tasks in under three minutes.⁸ However, Figure 3 shows that those who spent less than three minutes were significantly less creative compared to the rest ($P < 0.001$).

We reproduced Table 2 after removing participants who spent less than three minutes on the task and found qualitatively the same results. The effect size of ChatGPT-4 diminished by 11%, while we observed Bard performing even worse with an increase in the absolute size of the effect by 20%. This table can be found in the appendix (Table 14).

⁸Here, we round the number of seconds to the nearest minute. For example, 150 seconds and 209 seconds are both rounded to 3 minutes.

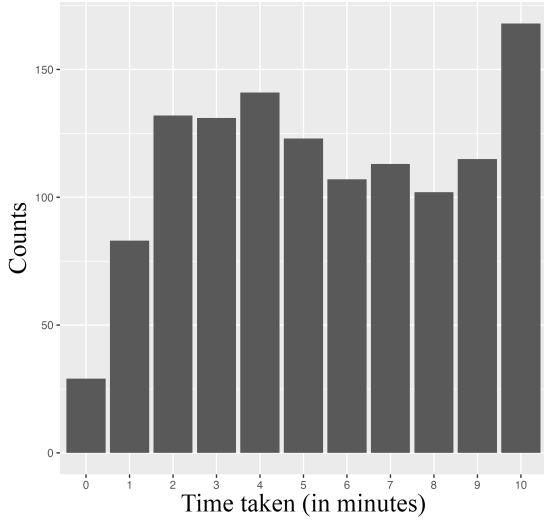


Figure 2: Histogram of Time Spent on Writing Task

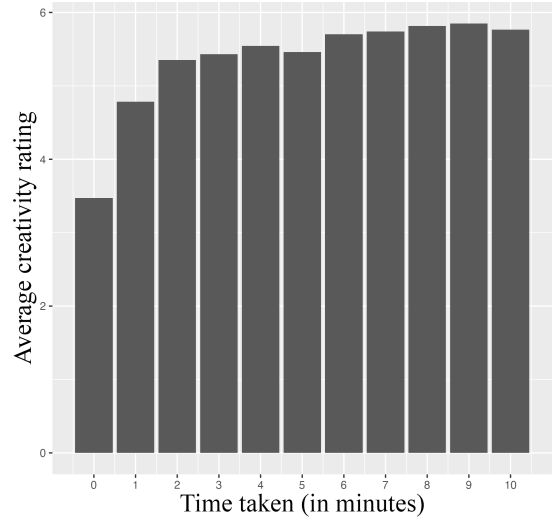


Figure 3: Average Creativity Score by Time Taken

One might argue that average creativity is less important than top creativity, as the latter is more likely to spur innovation.⁹ Next, we investigate the treatment composition of the top responses. Table 4 provides a comparative analysis of the representation of various sources within the top 10%, 5%, and 1% tiers of creative responses, as evaluated by BaselineRaters. One of the standout findings is the strong dominance of GPT-4 across all categories, showcasing its ability to generate highly creative responses compared to human participants and Bard. To illustrate, ChatGPT-4 significantly outperforms its competitors, with 96 entries within the top 10% bracket. This indicates that 43% of the creative responses generated by ChatGPT-4 are ranked within this top tier, a stark contrast to the mere 4% from responses produced by HumanBaseline. Furthermore, it’s noteworthy that Bard contributed only a single entry to the top 10%. We find a similar representation of sources in the top 10% for sub-dimensions of creativity (original, surprise and usefulness). Interestingly, among the very top creative texts, the top 1%, the share of contributions by humans (52.9%) exceeds the share of contributions by ChatGPT-4 (47.1%).

⁹While we acknowledge the significance of top creativity, as it can drive innovation, we also believe that understanding the distribution of creativity is important, as creativity can be valuable in various job roles, and small firms might not have access to top-tier creativity.

Table 4: Number of responses by source in the top 10, 5 and 1% most creative responses

Source	Top 10%	Top 5%	Top 1%
Bard	1 (0.6%)	0 (0%)	0 (0%)
ChatGPT-4	96 (56.8%)	50 (58.8%)	8 (47.1%)
HumanBaseline	30 (17.7%)	16 (18.8%)	6 (35.3%)
HumanPlusAI	31 (16.3%)	17 (20%)	3 (17.6%)
HumanAgainstAI	11 (6.5%)	2 (2.3%)	0 (0%)

3.1.2 Scores by Research Assistants

While online raters provide a representative judgment of the population, an alternative approach is to use "more sophisticated" raters. We consider this section as a robustness check. The main difference, apart from education, is that our research assistants had to go through all responses. This is a tiring task, but might lead to more consistency of judgment between texts.

Table 5 presents the results of regression analyses focused on the creativity ratings assigned by research assistants (RAs).¹⁰ Although the RAs generally evaluated creativity with greater stringency compared to the broader sample, the treatment differences between AI and human responses remained largely consistent. Specifically, ChatGPT4 significantly outperformed all other treatments in terms of perceived creativity, while Bard generated responses that were consistently rated as significantly less creative ($p < 0.001$). Notably, the HumanPlusAI treatment did not yield responses that were statistically more creative than those from the HumanBaseline group.

In terms of gender differences, the RAs found no significant difference in creative output between men and women in the HumanAgainstAI condition. However, in the HumanPlusAI condition, this gender difference persists, as indicated in models (6) and (7). Women seem indeed more capable to leverage the capacities of AI's in this context.

¹⁰Note that the texts were presented to the RAs in different sequences. To account for any potential influence of the presentation order, we control for this variable by including fixed effects for the order in which the texts were rated.

Table 5: Creativity ratings by research assistants

Dependent Variable: Model:	Creative Rating							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.176*** (0.064)	-1.177*** (0.065)	-1.176*** (0.065)			-1.178*** (0.065)		
ChatGPT4	0.551*** (0.072)	0.550*** (0.072)	0.551*** (0.072)			0.547*** (0.072)		
HumanPlusAI	0.142** (0.067)	0.142** (0.067)	0.142** (0.067)	0.125* (0.068)	0.118* (0.067)			
HumanPlusAI (Female)						0.264*** (0.082)	0.224** (0.091)	0.192** (0.091)
HumanPlusAI (Male)						0.008 (0.091)	0.034 (0.103)	0.052 (0.101)
HumanPlusAI (Other)						0.281 (0.435)	-0.239 (0.526)	-0.196 (0.491)
HumanAgainstAI	0.016 (0.073)	0.016 (0.073)						
HumanAgainstAI (Female)			-0.082 (0.097)	-0.137 (0.101)	-0.117 (0.098)	-0.082 (0.097)	-0.108 (0.103)	-0.095 (0.100)
HumanAgainstAI (Male)			0.086 (0.099)	0.167 (0.107)	0.151 (0.106)	0.085 (0.098)	0.140 (0.110)	0.133 (0.109)
HumanAgainstAI (Other)			0.680** (0.297)	0.203 (0.420)	0.304 (0.440)	0.671** (0.299)	0.035 (0.440)	0.160 (0.455)
Rater Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rating Order Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Prompt	Prompt	Prompt Age Gender	ALL	Prompt	Prompt Age Gender	All
Observations	8,423	8,423	8,423	6,226	6,201	8,423	6,226	6,201
R ²	0.713	0.713	0.713	0.722	0.728	0.714	0.722	0.728
Within R ²	0.079	0.079	0.080	0.005	0.027	0.081	0.006	0.027

Notes: OLS regression of creativity ratings by online raters with raters fixed effects. All controls include answers to the questionnaire comprising ten questions on creative and cognitive style and sensation-seeking behavior, based on questions by Nielsen, Pickett, and Simonton (2008) on creative style and Zuckerman et al. (1964) on sensation-seeking attitude, demographic queries concerning sibling count, birth order, handedness, and parental marital status, six queries about past involvement in creative activities (Hocevar, 1980), a non-incentivized measure of risk preferences (Dohmen et al., 2009), and categorical controls for major. Standard errors are clustered on the response level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

3.1.3 Correlates of creativity

Next, we investigate the correlates of creativity with the socio-economic observables of the participants. We conducted a follow-up survey of the participants of the creativity task two months after the main experiment. The response rate was 79.6% (the average creativity of those who responded and those who did not was not statistically different, $p > 0.32$). We collected data on their income, employment status, as well as the self-evaluation of their creativity and the creativity of their job (current job for employed and "dream job" for unemployed).

Table 6: Creativity ratings by online raters by raters treatments

Dependent Variable: Model:	Creative Rating		
	(1)	(2)	(3)
Age	0.000 (0.002)	0.002 (0.002)	0.002 (0.002)
Gender Female	-0.063 (0.048)	-0.028 (0.047)	-0.017 (0.047)
Gender Other	0.445*** (0.149)	0.424*** (0.143)	0.478*** (0.148)
Employed	-0.066 (0.071)		
Retired	0.277*** (0.105)		
At least Bachelor	0.239*** (0.053)		
Income 30'000-40'000	-0.141 (0.091)		
Income 40'000-50'000	-0.251*** (0.085)		
Income 50'000-60'000	-0.317*** (0.089)		
Income 60'000-70'000	-0.272*** (0.100)		
Income 70'000-80'000	-0.286*** (0.096)		
Income 80'000-90'000	-0.429*** (0.109)		
Income 90'000-100'000	0.061 (0.125)		
Income 100'000-150'000	-0.324*** (0.091)		
Income 150'000+	-0.236* (0.121)		
Creative Person (1-10)		0.060*** (0.010)	
Creative Job (1-10)			0.036*** (0.008)
Rater Fixed Effects	Yes	Yes	Yes
Observations	37,513	37,378	37,378
R ²	0.376	0.374	0.373
Within R ²	0.007	0.005	0.003

Notes: OLS regression of creativity ratings by online raters with rater fixed effects. Standard errors are clustered on the response level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6 displays the correlation of creativity with other participant characteristics. Column 1 presents socio-economic characteristics. While the creativity score does not significantly correlate with employment status and it significantly correlates with the propensity to have at least a bachelor’s degree and being retired or not. Columns 2 and 3 illustrate the correlation with self-reported scores (ranging from 0 to 10) for considering oneself a creative person and working in or aspiring for a creative job, respectively. As expected, there is a significant correlation between the creativity score and both self-reported creativity and the level of creativity associated with one’s job. These results underscore the external validity of our measure, at least in relation to self-assessed levels of creativity, both personally and professionally.

3.1.4 Semantic analysis of responses

Up to this point, we have explored the capacity of both humans and AI to generate creative texts. To achieve this, we gathered human evaluations of creativity for each text produced during the experiment. This method enables us to assess the creative abilities of individual agents, whether human or AI. However, the individual analysis of each response ignores the dimension of variety generated by each source of responses. If a source, be it AI or human, produces highly creative content that is consistently repetitive, then its overall contribution to group creativity is limited.

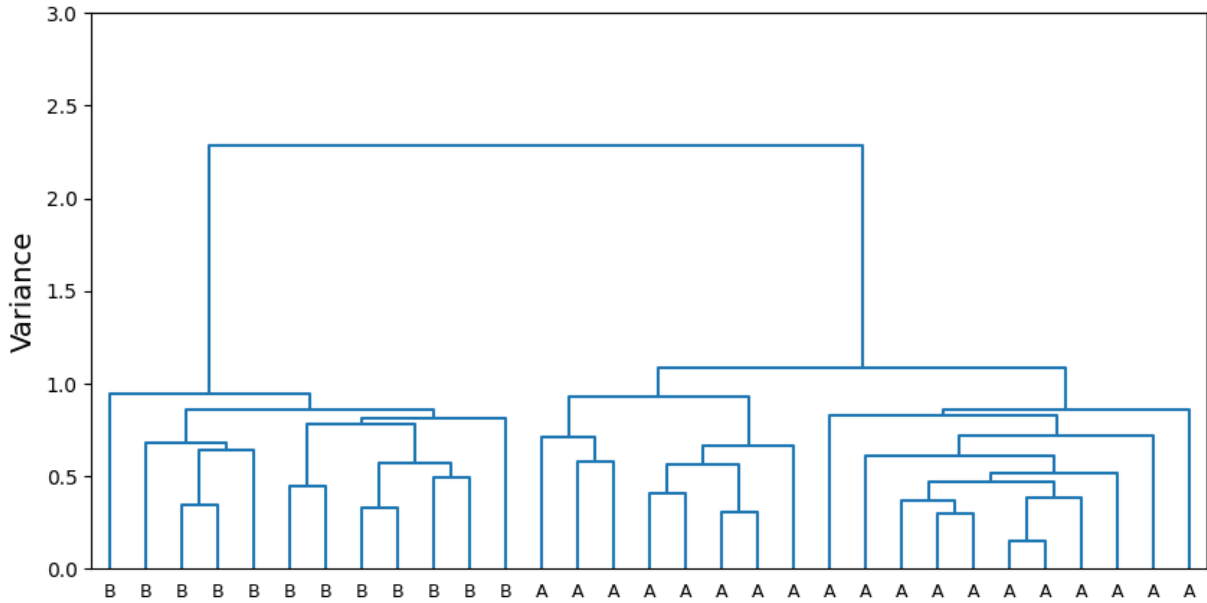
In this section, we use a combination of machine learning and deep learning approaches to measure the diversity of unique ideas that a source can generate. We start by converting each text into a numerical form using Sentence-BERT (sBERT) (Reimers and Gurevych, 2019).¹¹ This process transforms sentences into high-dimensional vectors (embeddings), capturing the semantic meaning of the texts. In this vector space, texts that are located close to each other have similar meanings, whilst texts that are far apart are considered unrelated. This approach offers a more nuanced understanding of language compared to the more traditional bag-of-words method. It enables the model to recognize synonyms, context, and even features such as irony or sarcasm. Furthermore, by mapping text into a continuous vector space, embeddings allow for advanced operations, such as clustering, by measuring the semantic distance between texts.

Following the transformation process, we calculate the pairwise cosine distances between all text embeddings within each source. Cosine distance measures the similarity between two texts, with 0 indicating high similarity and 1 indicating complete dissimilarity. We then perform hierarchical clustering (Saxena et al., 2017) to identify ”sufficiently unique” ideas produced by each source. Hierarchical clustering starts by treating each text as its own cluster. At each step, the two clusters with the smallest increase in total variance upon merging are combined, using Ward’s linkage method. This method minimizes the overall variance within clusters by merging similar texts first, as shown in the

¹¹Specifically, we used the all-mpnet-base-v2 model.

dendrogram (Figure 4). The process continues until all texts are unified into a single cluster, with the largest variance increase marking the final merge. Figure 4 illustrates this using texts generated by HumanBaseline, with the x-axis labeling texts by the prompt they address¹². The clustering effectively separates texts by prompts, maintaining distinctions across different sources.

Figure 4: Example of Hierarchical Clustering Dendrogram using a selection of Human-Baseline Texts



To quantify the number of distinct ideas generated, a "global distance threshold" (a horizontal cut of the dendrogram) must be established. Clusters below this threshold are interpreted as representing a single idea. Table 7 presents the percentage of "unique ideas" across two thresholds for three samples: (1) all produced texts, (2) the top 14% most creative text in each source, (3) the top 100 most creative texts for each source and (4) the top 10% most creative texts overall. For example, when examining all produced texts from ChatGPT at a threshold of 0.5, hierarchical clustering identified 63 unique ideas out of 224 texts, resulting in a "ratio of unique ideas" of 28.1%. A higher ratio indicates a greater capacity of a source to generate distinct ideas. The data clearly shows that Bard is more repetitive compared to its competitors, while humans and ChatGPT-4 exhibit similar levels of idea diversity. At a threshold of 1, there is negligible difference in the variety of ideas produced across sources when considering all texts. However, when focusing on the most creative responses, humans demonstrates greater diversity, even at higher thresholds.

¹²A = "If you had the talent to invent things just by thinking of them, what would you create" B = "Imagine and describe a town, city, or society in the future"

This suggests that the most creative humans maintain an advantage over AI in producing a wide array of ideas on a given topic.

Sample Threshold	ALL (0.5)	ALL (1)	Top 14% ‡ (0.5)	Top 14% ‡ (1)	Top 100 ‡ (0.5)	Top 100 ‡ (1)	Top 10% † (0.5)	Top 10% † (1)
ChatGPT-4	28.10%	6.70%	44.80%	10.30%	36.00%	7.00%	35.90%	7.80%
Bard	12.50%	5.10%	20.60%	6.90%	17.00%	6.00%	-	-
HumanBaseline	30.20%	6.70%	53.00%	10.00%	52.00%	10.00%	64.50%	9.70%
HumanAndAI	29.90%	7.30%	51.20%	9.80%	38.00%	6.00%	50.00%	12.50%
HumanAgainstAI	35.90%	6.90%	52.40%	14.30%	42.00%	8.00%	53.30%	20.00%

Notes: ‡: Top 14% or Top 100 texts within each source. †: Top 10% of texts, pooled across all sources.

Table 7: Comparison of ratios of unique ideas across sources and sample for different global Thresholds

The selection of threshold levels at 0.5 and 1 is somewhat arbitrary. In the appendix, Figures 6, 7, and 8 present the ratio of unique ideas across all possible thresholds for all texts, the top 100 texts, and the top 10% of texts, respectively. These graphs corroborate our initial findings: ideas from Humans remain distinct across greater semantic distances, while those from GPT-4 tend to converge at shorter distances. This demonstrates that the most creative individuals in our study are capable of generating ideas that are more unique compared to their AI-generated counterparts.

3.2 Creativity task: Raters

Do raters evaluate the creativity of generated responses differently based on our experimental manipulations? Table 8 presents the estimated treatment effects, controlled for response-level fixed effects in Model (1). First, responses that were grammatically corrected did not receive significantly different ratings compared to their uncorrected counterparts. This finding is reassuring as it suggests that creativity assessments are not influenced by linguistic accuracy. Second, informing raters that some responses might be AI-generated led to a marginally significant decrease in creativity ratings ($p = 0.051$).

Models (2) and (3) explore ratings within the AIRaters treatment group, breaking them down based on the raters’ guesses regarding whether a response was AI-generated.¹³ The negative treatment effect is largely driven by lower ratings assigned to responses that raters believed to be AI-generated.

Further analysis focusing on the sub-dimensions of creativity—originality, surprise, and usefulness—is detailed in Table 13. The pattern of significantly lower ratings in the AIRaters treatment group is consistent across all dimensions for responses believed to be AI-generated.

¹³Note that raters were asked to guess the origin of a response only after they had submitted their ratings. This was to ensure that the additional question did not influence the ratings.

Table 8: Creativity ratings by online raters by raters treatments

	Creativity rating	Creativity rating	Creativity rating
CorrectedRaters	-0.082 (0.065)	-0.079 (0.065)	-0.068 (0.064)
AIRaters	-0.129* (0.066)		
AIRatersGuessHuman		-0.098 (0.069)	-0.081 (0.068)
AIRatersGuessAI		-0.162** (0.071)	-0.158** (0.070)
Observations	65493	65493	65393
R-squared	0.232	0.232	0.244
Controls	response FE	response FE	response FE+controls
N of clusters	3283	3283	3278

Notes: OLS regression of creativity ratings by online raters with response fixed effects. Individual controls include raters’ age, gender, and a non-incentivized measure of risk preferences (Falk et al., 2018). Standard errors are clustered on the raters level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

3.2.1 Identification of source by online Raters

We turn to discuss whether human raters were able to correctly identify if a text was created by a human or an AI chatbot (Table 9). Raters classify the human-generated ideas as being human, i.e., correctly, in 63% of cases (no difference between Human Baseline and Human against AI), which is significantly better than chance. The rate of correct classification is significantly lower for chatbots. For ChatGPT-4, raters classified the ideas as AI-generated in 60% of cases, which is still significantly better than chance. For Bard-generated ideas, raters classified them as AI-generated only in 37% of the cases, significantly lower than chance. Interestingly, the ideas generated by humans plus AI are more likely to be categorized as AI-generated (60%).

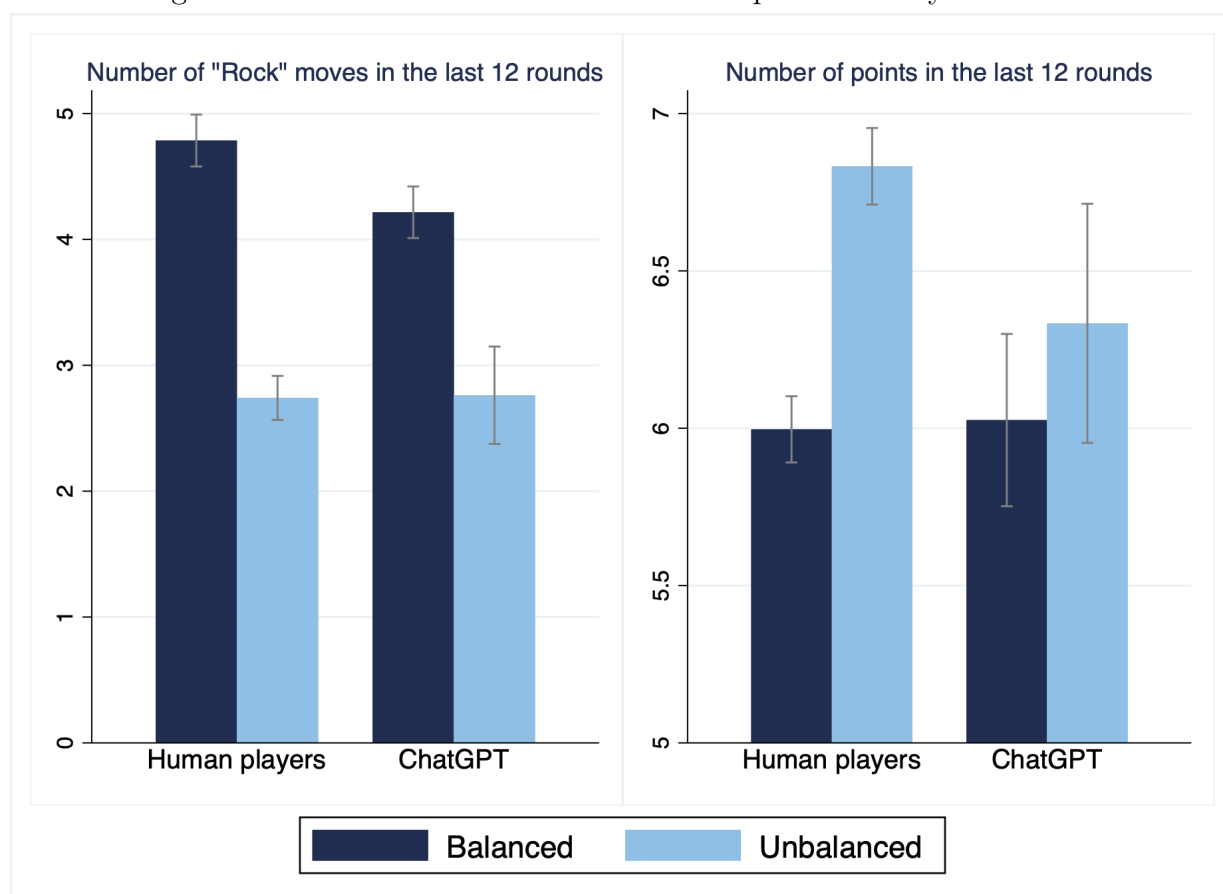
Table 9: Rater’s average identification rates for responses as human or AI, by source

Source	Guess <i>Human</i>	Guess <i>AI</i>	Std. Error
<i>Chatbots</i>			
Bard	63%	37%	0.009
ChatGPT-4	39.3%	60.7%	0.009
<i>Humans</i>			
HumanBaseline	62.9%	37.1%	0.005
HumanAgainstAI	63.2%	36.8%	0.008
<i>HumanPlusAI</i>	40.6%	59.4%	0.008

3.3 Strategic Task

The ability of ChatGPT-4 to adapt its responses during a chat session opens up the possibility for strategic behavior. We designed a 24-round setup where finding a best response is non-trivial. While ChatGPT-4 could likely draw on its extensive training data for equilibrium play in the "rock-paper-scissors" game, adapting to an opponent's biased moves must be learned within the interaction. The left panel of figure 5 presents the number of "Rock" moves in the last 12 rounds across treatments. Both humans and ChatGPT-4 significantly reduced the frequency of using "Rock" in Unbalanced treatments, i.e. when the opponent never used "Scissors". The difference between the Balanced and Unbalanced treatments is significant for both human participants and ChatGPT-4 ($p < 0.001$), indicating that ChatGPT-4 can learn strategic responses within a chat of 24 interactions.

Figure 5: Number of "Rock" moves and the points won by treatments



Gray bars present 95% confidence intervals

To evaluate performance, we compare the average number of points won in the last 12 rounds. The left panel of Figure 5 illustrates the points won across treatments. In the Balanced treatment, no significant difference in earnings between Humans and ChatGPT-

4 was observed. However, in the Unbalanced treatment, human players outperformed ChatGPT-4, earning significantly more points ($p < 0.001$).

Table 10: OLS for number of points in the last 12 rounds on source

Dependent Variable: Model:	Points (Last 12)	
	Balanced	Unbalanced
Constant	6.009*** (0.133)	6.304*** (0.181)
Human	-0.012 (0.144)	0.529*** (0.192)
Observations	811	779
R ²	8.89×10^{-6}	0.010
Adjusted R ²	-0.001	0.008

Notes: OLS regression of points score in the last 12 rounds on the player’s type (ChatGPT-4 or Human). Standard errors are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Why does ChatGPT-4’s strategic choice of reducing ”Rock” moves not result in higher payoffs, as it does for humans? Table 11 shows that human participants shifted towards the dominant action of ”Paper” far more frequently than ChatGPT-4 ($p < 0.001$). This action is weakly dominant only under the assumption that the opponent either cannot adapt their strategy or that their moves are predetermined. In our experiment, ChatGPT-4 adapted its moves as if the opponent had restricted their strategy to a two-move game. Note that our prompt to ChatGPT-4 and instructions for participants include the statement of the moves being pre-determined. Both ChatGPT-4 and human participants might have doubts about the strategy; in such cases, the ”Paper” move is not weakly dominant, as switching to ”Scissors” becomes an obvious response by the computerized opponent.

Table 11: OLS of moves in the last 12 rounds on source/treatment interaction

Dependent Variable: Model:	Rock (Last 12) (1)	Paper (Last 12) (2)	Scissors (Last 12) (3)
Constant	4.786*** (0.093)	3.777*** (0.100)	3.437*** (0.083)
ChatGPT4	-0.596** (0.247)	-0.053 (0.265)	0.545** (0.219)
Unbalanced	-2.045*** (0.132)	2.046*** (0.142)	-0.001 (0.117)
ChatGPT4 \times Unbalanced	0.557 (0.376)	-1.425*** (0.404)	0.828** (0.333)
Observations	1,590	1,590	1,590
R ²	0.140	0.126	0.022
Adjusted R ²	0.139	0.124	0.021

Notes: OLS regression of moves in the last 12 rounds on the player’s type (ChatGPT-4 or Human) and game treatment (Balanced or Unbalanced). Standard errors are reported in brackets.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4 Conclusion

Our findings demonstrate the creative prowess of ChatGPT-4. It notably outperformed the average human output in our open-ended creativity task, and demonstrated the ability react to biased opponent in a strategic setting.

One could criticize the specificity of the tasks chosen to measure creativity and strategic skills. While we relied on a wide body of literature from psychology and economics to inform our task selection, creativity remains an inherently difficult construct to measure. We believe that the task captures important dimensions of creativity applicable to various fields, such as marketing slogan generation, storytelling, and even scientific research. According to [Charness and Grieco \(2019\)](#), the performance in our chosen task appears to be robust to incentives, which is crucial when evaluating intrinsic creative potential.

In selecting a task to evaluate strategic skills, we sought a straightforward task that would be easily comprehensible and would not require extensive explanation. The goal was to choose a task that would allow for real-time adaptation based on the opponent’s moves, thus serving as an indicator of strategic intelligence. While ChatGPT-4 showed signs of strategic adaptation in tasks, the scope was limited to a set number of rounds and a relatively straightforward game of Rock-Paper-Scissors. Its ability to adapt to more complex, long-term strategic environments remains an open question.

Our study employed a representative sample of the U.S. population, but an alternative approach might focus on samples of professionals in creative or strategic fields. We argue that our broader sampling strategy is valid as many everyday innovations occur within small to medium-sized enterprises and not solely within specialized research and development contexts.

Another consideration is that our findings may be influenced by the relative novelty of AI technology, especially findings concerning human reactions and perceptions of AI. As AI becomes more widespread, it is plausible that human users will become more adept at leveraging its capabilities. However, we observe that the results indicating lower creative performance among female participants and harsher ratings against AI-generated responses suggest a more basic psychological or societal bias that might not be easily addressed as AI technologies mature.

Our results demonstrate that artificial intelligence (AI), exemplified by models like ChatGPT-4, can serve as a valuable asset throughout the creative process, particularly in generating novel ideas within established contexts. The potential benefits for organizations across various sectors are clear, from streamlining brainstorming sessions to enhancing the quality of idea generation.

Moreover, our research underscores AI's emerging potential in strategic decision-making. Although humans maintained a performance edge in tasks requiring dynamic strategic adjustments, the ability of AI models like ChatGPT-4 to adapt their strategy over a 24-round series of interactions suggests their utility in providing real-time strategic advice.

Interestingly, we found that human participants have yet to optimize their interactions with AI, even when provided with straightforward, pre-registered prompts. This highlights an educational opportunity for both the general public and professionals to learn how to collaborate more effectively with AI technologies. The development of best practices for human-AI interaction may become a vital skill set for the workforce of the future.

Our study revealed a gender disparity in creative performance when in competition with AI. As AI technology becomes increasingly prevalent in the workplace, comprehending these social dynamics and their potential to exacerbate existing inequalities will be critical. Targeted training or interventions may be necessary to ensure equitable access and proficiency in leveraging AI tools across gender lines.

Our data also reveal a perceptual gap: human raters assigned lower scores to outputs they believed to be AI-generated, indicating a level of public skepticism or resistance towards AI. As these technologies become increasingly integrated into various societal sectors, shifting public perceptions will be crucial for maximizing the potential benefits of AI.

In summary, our study provides compelling evidence supporting the adoption of AI in both creative and strategic tasks, illuminating its capabilities and the challenges ahead.

References

- Abada, Ibrahim and Xavier Lambin**, “Artificial intelligence: Can seemingly collusive outcomes be avoided?,” *Management Science*, 2023.
- Arntz, Melanie, Terry Gregory, and Ulrich Zierahn**, “The risk of automation for jobs in OECD countries: A comparative analysis,” 2016.
- Autor, David H**, “Why are there still so many jobs? The history and future of workplace automation,” *Journal of economic perspectives*, 2015, 29 (3), 3–30.
- **and David Dorn**, “The growth of low-skill service jobs and the polarization of the US labor market,” *American economic review*, 2013, 103 (5), 1553–1597.
- Boden, Margaret**, “Creativity and knowledge,” *Creativity in education*, 2001, pp. 95–102.
- Boden, Margaret A**, “Creativity and artificial intelligence,” *Artificial intelligence*, 1998, 103 (1-2), 347–356.
- Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond**, “Generative AI at work,” Technical Report, National Bureau of Economic Research 2023.
- Charness, Gary and Daniela Grieco**, “Creativity and incentives,” *Journal of the European Economic Association*, 2019, 17 (2), 454–496.
- **and –**, “Creativity and AI,” *Available at SSRN 4686415*, 2024.
- Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong**, “The emergence of economic rationality of GPT,” *Proceedings of the National Academy of Sciences*, 2023, 120 (51), e2316205120.
- Chugunova, Marina and Wolfgang J Luhan**, “Ruled by robots: Preference for algorithmic decision makers and perceptions of their choices,” *Public Choice*, 2024, pp. 1–24.
- Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler**, “Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence,” *Management Science*, 2023.
- Dell’Acqua, Fabrizio, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraymer, François Candelon, and Karim R Lakhani**, “Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality,” *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 2023, (24-013).
- Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey**, “Algorithm aversion: people erroneously avoid algorithms after seeing them err.,” *Journal of experimental psychology: General*, 2015, 144 (1), 114.
- , – , **and –**, “Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them,” *Management science*, 2018, 64 (3), 1155–1170.
- Dignum, Virginia**, *Responsible artificial intelligence: how to develop and use AI in a responsible way*, Vol. 2156, Springer, 2019.
- Doshi, Anil R and Oliver Hauser**, “Generative artificial intelligence enhances creativity,” *Available at SSRN*, 2023.

- Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde**, “Global evidence on economic preferences,” *The Quarterly Journal of Economics*, 2018, *133* (4), 1645–1692.
- Felten, Ed, Manav Raj, and Robert Seamans**, “How will Language Modelers like ChatGPT Affect Occupations and Industries?,” *arXiv preprint arXiv:2303.01157*, 2023.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli**, “ChatGPT outperforms crowd workers for text-annotation tasks,” *Proceedings of the National Academy of Sciences*, jul 2023, *120* (30).
- Girotra, Karan, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich**, “Ideas are dimes a dozen: Large language models for idea generation in innovation,” *Available at SSRN 4526071*, 2023.
- , –, –, and **Karl T. Ulrich**, “Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation,” *The Wharton School Research Paper Forthcoming*, 2023. Available at SSRN: <https://ssrn.com/abstract=4526071> or <http://dx.doi.org/10.2139/ssrn.4526071>.
- Gneezy, Uri, Muriel Niederle, and Aldo Rustichini**, “Performance in competitive environments: Gender differences,” *The quarterly journal of economics*, 2003, *118* (3), 1049–1074.
- Greiner, Ben, Philipp Grünwald, Thomas Lindner, Georg Lintner, and Martin Wiernsperger**, *Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study*, Vienna University of Economics and Business, Department of Strategy, 2024.
- Guilford, Joy P**, “Varieties of creative giftedness, their measurement and development,” *Gifted child quarterly*, 1975, *19* (2), 107–121.
- Huang, Fan, Haewoon Kwak, and Jisun An**, “Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech,” in “Companion proceedings of the ACM web conference 2023” 2023, pp. 294–297.
- Kiron, David and Michael Schrage**, “Strategy for and with AI,” *MIT Sloan Management Review*, 2019, *60* (4).
- Kuzman, Taja, Igor Mozetic, and Nikola Ljubešić**, “ChatGPT: beginning of an end of manual linguistic data annotation,” *Use Case of Automatic Genre Identification. ArXiv abs/2303.03953*, 2023.
- Niederle, Muriel and Lise Vesterlund**, “Gender and competition,” *Annu. Rev. Econ.*, 2011, *3* (1), 601–630.
- Reimers, Nils and Iryna Gurevych**, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
- Saccardo, Silvia, Aniela Pietrasz, and Uri Gneezy**, “On the size of the gender difference in competitiveness,” *Management Science*, 2018, *64* (4), 1541–1554.
- Saxena, Amit, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin**, “A review of clustering techniques and developments,” *Neurocomputing*, 2017, *267*, 664–681.
- Shakked, Noy and Zhang Whitney**, “Experimental evidence on the productivity effects of generative artificial intelligence,” *Science*, 2023, *381* (6654), 187–192.
- Smith, Aaron and Janna Anderson**, “AI, Robotics, and the Future of Jobs,” 2014.

Sternberg, R.J., *Beyond IQ: A Triarchic Theory of Human Intelligence* Cambridge paperback library, Cambridge University Press, 1985.

5 Appendix

Table 12: Sub-dimensions of creativity by online raters

Dependent Variable: Model:	(1)	(2)	(3)	original		(6)	(7)	(8)
				(4)	(5)			
Bard	-1.039*** (0.051)	-1.056*** (0.051)	-1.056*** (0.051)			-1.056*** (0.051)		
ChatGPT4	1.962*** (0.043)	1.942*** (0.043)	1.942*** (0.043)			1.942*** (0.043)		
HumanPlusAI	0.335*** (0.052)	0.335*** (0.052)	0.336*** (0.052)	0.335*** (0.052)	0.304*** (0.051)			
HumanPlusAI (Female)						0.390*** (0.068)	0.433*** (0.072)	0.379*** (0.070)
HumanPlusAI (Male)						0.283*** (0.069)	0.266*** (0.073)	0.256*** (0.072)
HumanPlusAI (Other)						0.290 (0.236)	-0.347 (0.289)	-0.337 (0.295)
HumanAgainstAI	-0.027 (0.054)	-0.034 (0.053)				-0.034 (0.053)	-0.044 (0.053)	-0.047 (0.052)
HumanAgainstAI (Female)			-0.122* (0.071)	-0.123* (0.074)	-0.111 (0.071)			
HumanAgainstAI (Male)			0.044 (0.069)	0.056 (0.073)	0.034 (0.072)			
HumanAgainstAI (Other)			0.278 (0.234)	-0.243 (0.292)	-0.205 (0.306)			
Rater Fixed Effects Controls	Yes None	Yes Prompt	Yes Prompt	Yes Prompt Age Gender	Yes ALL	Yes Prompt	Yes Prompt Age Gender	Yes All
Observations	63,812	63,812	63,812	47,033	46,845	63,812	47,033	46,845
R ²	0.391	0.393	0.394	0.366	0.375	0.393	0.366	0.375
Within R ²	0.103	0.106	0.106	0.008	0.022	0.106	0.008	0.022

Dependent Variable: Model:	(1)	(2)	(3)	surprise		(6)	(7)	(8)
				(4)	(5)			
Bard	-1.040*** (0.048)	-1.061*** (0.048)	-1.061*** (0.048)			-1.061*** (0.048)		
ChatGPT4	1.627*** (0.043)	1.603*** (0.042)	1.603*** (0.042)			1.603*** (0.042)		
HumanPlusAI	0.193*** (0.048)	0.193*** (0.048)	0.193*** (0.048)	0.193*** (0.048)	0.163*** (0.048)			
HumanPlusAI (Female)						0.217*** (0.064)	0.253*** (0.067)	0.203*** (0.067)
HumanPlusAI (Male)						0.172*** (0.064)	0.159** (0.068)	0.150** (0.067)
HumanPlusAI (Other)						0.141 (0.248)	-0.396 (0.312)	-0.378 (0.317)
HumanAgainstAI	0.085* (0.052)	0.077 (0.051)				0.077 (0.051)	0.065 (0.051)	0.067 (0.050)
HumanAgainstAI (Female)			0.013 (0.069)	0.016 (0.071)	0.026 (0.069)			
HumanAgainstAI (Male)			0.145** (0.066)	0.147** (0.070)	0.138** (0.069)			
HumanAgainstAI (Other)			0.086 (0.241)	-0.433 (0.305)	-0.412 (0.309)			
Rater Fixed Effects Controls	Yes None	Yes Prompt	Yes Prompt	Yes Prompt Age Gender	Yes ALL	Yes Prompt	Yes Prompt Age Gender	Yes All
Observations	63,812	63,812	63,812	47,033	46,845	63,812	47,033	46,845
R ²	0.397	0.400	0.400	0.377	0.384	0.400	0.377	0.384
Within R ²	0.077	0.082	0.082	0.006	0.017	0.082	0.006	0.017

Dependent Variable: Model:	(1)	(2)	(3)	useful		(6)	(7)	(8)
				(4)	(5)			
Bard	-0.144*** (0.052)	-0.179*** (0.049)	-0.179*** (0.049)			-0.178*** (0.049)		
ChatGPT4	1.024*** (0.041)	0.985*** (0.039)	0.985*** (0.039)			0.985*** (0.039)		
HumanPlusAI	0.362*** (0.046)	0.363*** (0.046)	0.363*** (0.046)	0.362*** (0.046)	0.365*** (0.047)			
HumanPlusAI (Female)						0.438*** (0.055)	0.421*** (0.059)	0.425*** (0.060)
HumanPlusAI (Male)						0.284*** (0.066)	0.305*** (0.069)	0.309*** (0.070)
HumanPlusAI (Other)						0.449*** (0.125)	0.292 (0.190)	0.220 (0.192)
HumanAgainstAI	-0.091* (0.053)	-0.104** (0.052)				-0.104** (0.052)	-0.105** (0.051)	-0.121** (0.051)
HumanAgainstAI (Female)			-0.134* (0.071)	-0.186** (0.073)	-0.195*** (0.072)			
HumanAgainstAI (Male)			-0.082 (0.067)	-0.022 (0.071)	-0.047 (0.071)			
HumanAgainstAI (Other)			0.065 (0.223)	-0.020 (0.251)	-0.006 (0.260)			
Rater Fixed Effects Controls	Yes None	Yes Prompt	Yes Prompt	Yes Prompt Age Gender	Yes ALL	Yes Prompt	Yes Prompt Age Gender	Yes All
Observations	63,812	63,812	63,812	47,033	46,845	63,812	47,033	46,845
R ²	0.372	0.380	0.380	0.380	0.383	0.381	0.380	0.383
Within R ²	0.023	0.036	0.036	0.016	0.021	0.036	0.016	0.021

Notes: OLS regression of creativity ratings by online raters with raters fixed effects. All controls include answers to the questionnaire comprising ten questions on creative and cognitive style and sensation-seeking behavior, based on questions by Nielsen, Pickett, and Simonton (2008) on creative style and Zuckerman et al. (1964) on sensation-seeking attitude, demographic queries concerning sibling count, birth order, handedness, and parental marital status, six queries about past involvement in creative activities (Hocevar, 1980), a non-incentivized measure of risk preferences (Dohmen et al., 2009), and categorical controls for major. Standard errors are clustered on the response level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: Sub-dimensions of creativity ratings by online raters by raters treatments

	original	original	original
CorrectedRaters	-0.089 (0.068)	-0.086 (0.068)	-0.072 (0.066)
AIRaters	-0.120* (0.068)		
AIRatersGuessHuman		-0.092 (0.071)	-0.078 (0.069)
AIRatersGuessAI		-0.152** (0.074)	-0.148** (0.073)
Observations	65493	65493	65393
R-squared	0.225	0.225	0.238
	surprise	surprise	surprise
CorrectedRaters	-0.098 (0.072)	-0.095 (0.072)	-0.080 (0.071)
AIRaters	-0.183** (0.072)		
AIRatersGuessHuman		-0.157** (0.075)	-0.146** (0.074)
AIRatersGuessAI		-0.211*** (0.078)	-0.203*** (0.076)
Observations	65493	65493	65393
R-squared	0.187	0.187	0.198
	useful	useful	useful
CorrectedRaters	0.028 (0.072)	0.035 (0.072)	0.031 (0.071)
AIRaters	-0.085 (0.073)		
AIRatersGuessHuman		-0.009 (0.075)	-0.013 (0.074)
AIRatersGuessAI		-0.169** (0.079)	-0.172** (0.078)
Observations	65493	65493	65393
R-squared	0.159	0.160	0.171
Controls	response FE	response FE	response FE+controls
N of clusters	3283	3283	3278

Notes: OLS regression of creativity ratings by online raters with response fixed effects. Individual controls include raters' age, gender, and a non-incentivized measure of risk preferences (Falk et al., 2018). Standard errors are clustered on the raters level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 14: Creativity ratings by online raters (Writing time > 3mins)

Dependent Variable: Model:	Creative Rating							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.122*** (0.050)	-1.150*** (0.050)	-1.150*** (0.050)			-1.151*** (0.050)		
ChatGPT4	1.682*** (0.039)	1.650*** (0.039)	1.650*** (0.039)			1.650*** (0.039)		
HumanPlusAI	0.458*** (0.054)	0.449*** (0.054)	0.449*** (0.054)	0.459*** (0.054)	0.424*** (0.053)			
HumanPlusAI (Female)						0.438*** (0.071)	0.520*** (0.074)	0.487*** (0.072)
HumanPlusAI (Male)						0.451*** (0.075)	0.402*** (0.078)	0.366*** (0.077)
HumanPlusAI (Other)						0.646** (0.253)	0.199 (0.312)	0.119 (0.314)
HumanAgainstAI	-0.117** (0.056)	-0.127** (0.055)				-0.127** (0.055)	-0.137** (0.055)	-0.126** (0.052)
HumanAgainstAI (Female)			-0.198*** (0.075)	-0.144* (0.078)	-0.111 (0.073)			
HumanAgainstAI (Male)			-0.071 (0.072)	-0.111 (0.075)	-0.130* (0.074)			
HumanAgainstAI (Other)			0.083 (0.215)	-0.447 (0.279)	-0.280 (0.306)			
Rater Fixed Effects Controls	Yes None	Yes Prompt	Yes Prompt	Yes Prompt Age Gender	Yes ALL	Yes Prompt	Yes Prompt Age Gender	Yes All
Observations	54,771	54,771	54,771	37,992	37,846	54,771	37,992	37,846
R ²	0.412	0.415	0.416	0.393	0.401	0.415	0.393	0.401
Within R ²	0.114	0.119	0.119	0.015	0.028	0.119	0.015	0.028

Notes: OLS regression of creativity ratings by online raters with raters fixed effects. All controls include answers to the questionnaire comprising ten questions on creative and cognitive style and sensation-seeking behavior, based on questions by Nielsen, Pickett, and Simonton (2008) on creative style and Zuckerman et al. (1964) on sensation-seeking attitude, demographic queries concerning sibling count, birth order, handedness, and parental marital status, six queries about past involvement in creative activities (Hocevar, 1980), a non-incentivized measure of risk preferences (Dohmen et al., 2009), and categorical controls for major. Standard errors are clustered on the response level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 15: Sub-dimensions of creativity ratings by RAs

Dependent Variable: Model:	(1)	(2)	(3)	Original Rating		(6)	(7)	(8)
				(4)	(5)			
Bard	-1.426*** (0.079)	-1.437*** (0.081)	-1.437*** (0.081)			-1.439*** (0.081)		
ChatGPT4	0.042 (0.094)	0.028 (0.093)	0.028 (0.093)			0.025 (0.093)		
HumanPlusAI	-0.416*** (0.088)	-0.412*** (0.086)	-0.412*** (0.086)	-0.439*** (0.089)	-0.457*** (0.088)			
HumanPlusAI (Female)						-0.330*** (0.109)	-0.403*** (0.123)	-0.467*** (0.123)
HumanPlusAI (Male)						-0.497*** (0.116)	-0.465*** (0.134)	-0.437*** (0.134)
HumanPlusAI (Other)						-0.424 (0.493)	-0.733 (0.725)	-0.671 (0.722)
HumanAgainstAI	0.077 (0.093)	0.071 (0.091)						
HumanAgainstAI (Female)			0.044 (0.118)	-0.048 (0.124)	-0.010 (0.122)	0.043 (0.118)	-0.037 (0.127)	-0.013 (0.126)
HumanAgainstAI (Male)			0.080 (0.124)	0.134 (0.136)	0.122 (0.135)	0.080 (0.124)	0.127 (0.139)	0.129 (0.138)
HumanAgainstAI (Other)			0.448 (0.437)	0.240 (0.560)	0.310 (0.546)	0.442 (0.438)	0.107 (0.671)	0.215 (0.635)
Rater Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rating Order Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Prompt	Prompt	Prompt Age Gender	Prompt ALL	Prompt	Prompt Age Gender	All
Observations	8,422	8,422	8,422	6,226	6,202	8,422	6,226	6,202
R ²	0.607	0.611	0.611	0.613	0.621	0.611	0.613	0.621
Within R ²	0.062	0.071	0.071	0.027	0.049	0.071	0.027	0.049

Dependent Variable: Model:	(1)	(2)	(3)	Surprise Rating		(6)	(7)	(8)
				(4)	(5)			
Bard	-1.202*** (0.073)	-1.211*** (0.074)	-1.211*** (0.074)			-1.213*** (0.074)		
ChatGPT4	0.065 (0.088)	0.052 (0.087)	0.052 (0.087)			0.051 (0.087)		
HumanPlusAI	-0.338*** (0.081)	-0.334*** (0.080)	-0.334*** (0.080)	-0.344*** (0.083)	-0.348*** (0.082)			
HumanPlusAI (Female)						-0.252** (0.104)	-0.318*** (0.117)	-0.365*** (0.118)
HumanPlusAI (Male)						-0.398*** (0.105)	-0.337*** (0.121)	-0.293*** (0.120)
HumanPlusAI (Other)						-0.801* (0.460)	-1.275** (0.631)	-1.203* (0.644)
HumanAgainstAI	0.090 (0.087)	0.085 (0.086)						
HumanAgainstAI (Female)			0.043 (0.111)	-0.037 (0.115)	-0.026 (0.112)	0.042 (0.111)	-0.030 (0.118)	-0.032 (0.116)
HumanAgainstAI (Male)			0.111 (0.119)	0.186 (0.131)	0.178 (0.131)	0.111 (0.119)	0.190 (0.134)	0.197 (0.133)
HumanAgainstAI (Other)			0.415 (0.374)	0.315 (0.501)	0.418 (0.489)	0.411 (0.375)	-0.104 (0.576)	0.036 (0.549)
Rater Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rating Order Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Prompt	Prompt	Prompt Age Gender	Prompt ALL	Prompt	Prompt Age Gender	All
Observations	8,423	8,423	8,423	6,228	6,203	8,423	6,228	6,203
R ²	0.641	0.643	0.644	0.648	0.655	0.644	0.649	0.655
Within R ²	0.051	0.059	0.059	0.021	0.041	0.060	0.022	0.042

Dependent Variable: Model:	(1)	(2)	(3)	Useful Rating		(6)	(7)	(8)
				(4)	(5)			
Bard	-1.041*** (0.068)	-1.047*** (0.069)	-1.046*** (0.069)			-1.049*** (0.069)		
ChatGPT4	0.322*** (0.073)	0.314*** (0.072)	0.315*** (0.072)			0.310*** (0.073)		
HumanPlusAI	0.006 (0.067)	0.009 (0.066)	0.010 (0.066)	-0.023 (0.067)	-0.025 (0.067)			
HumanPlusAI (Female)						0.158* (0.082)	0.117 (0.091)	0.083 (0.092)
HumanPlusAI (Male)						-0.154* (0.089)	-0.170* (0.102)	-0.141 (0.100)
HumanPlusAI (Other)						0.175 (0.389)	-0.077 (0.441)	-0.072 (0.422)
HumanAgainstAI	0.008 (0.072)	0.005 (0.072)						
HumanAgainstAI (Female)			-0.097 (0.097)	-0.187* (0.103)	-0.180* (0.101)	-0.098 (0.097)	-0.144 (0.104)	-0.147 (0.102)
HumanAgainstAI (Male)			0.091 (0.095)	0.171* (0.104)	0.145 (0.105)	0.090 (0.094)	0.126 (0.107)	0.111 (0.107)
HumanAgainstAI (Other)			0.475 (0.358)	0.170 (0.412)	0.252 (0.422)	0.461 (0.360)	0.135 (0.429)	0.224 (0.438)
Rater Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rating Order Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Prompt	Prompt	Prompt Age Gender	Prompt ALL	Prompt	Prompt Age Gender	All
Observations	8,424	8,424	8,424	6,226	6,202	8,424	6,226	6,202
R ²	0.665	0.667	0.667	0.680	0.685	0.667	0.680	0.685
Within R ²	0.049	0.052	0.053	0.011	0.025	0.055	0.013	0.025

Notes: OLS regression of creativity ratings by online raters with raters fixed effects. All controls include answers to the questionnaire comprising ten questions on creative and cognitive style and sensation-seeking behavior, based on questions by Nielsen, Pickett, and Simonton (2008) on creative style and Zuckerman et al. (1964) on sensation-seeking attitude, demographic queries concerning sibling count, birth order, handedness, and parental marital status, six queries about past involvement in creative activities (Hocevar, 1980), a non-incentivized measure of risk preferences (Dohmen et al., 2009), and categorical controls for major. Standard errors are clustered on the response level and are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

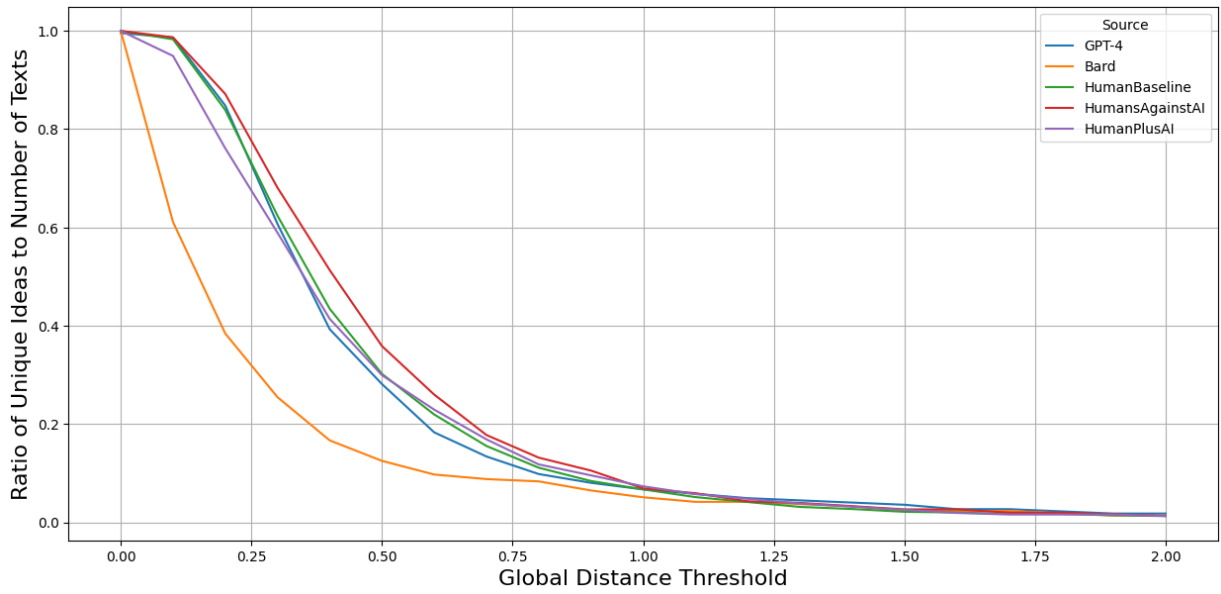


Figure 6: Caption for the figure

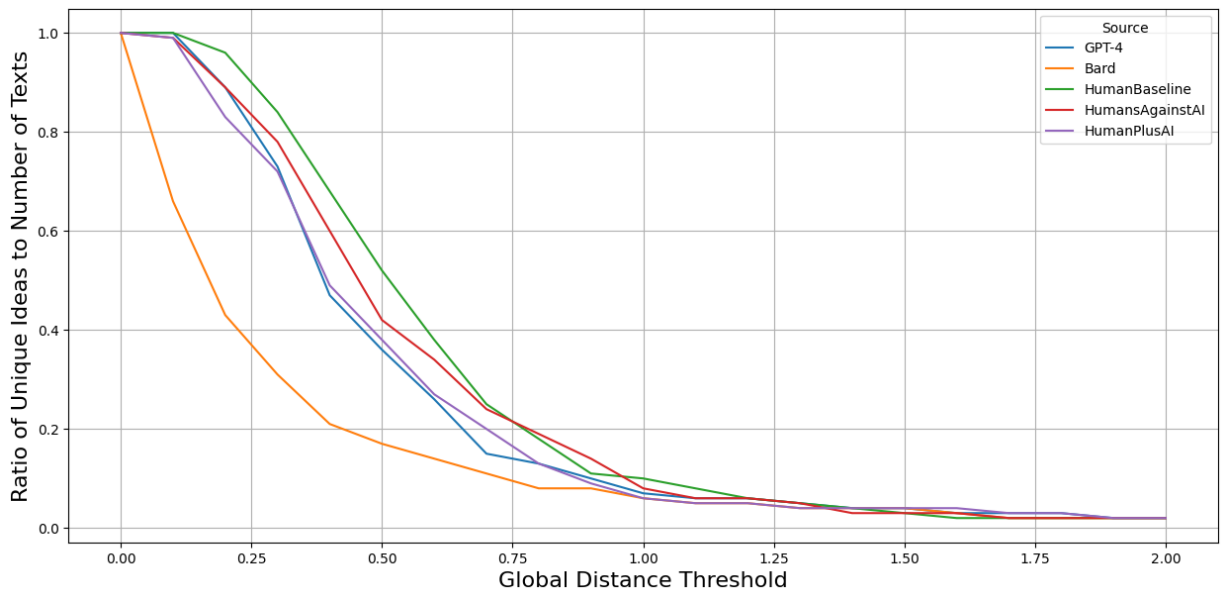


Figure 7: Caption for the figure

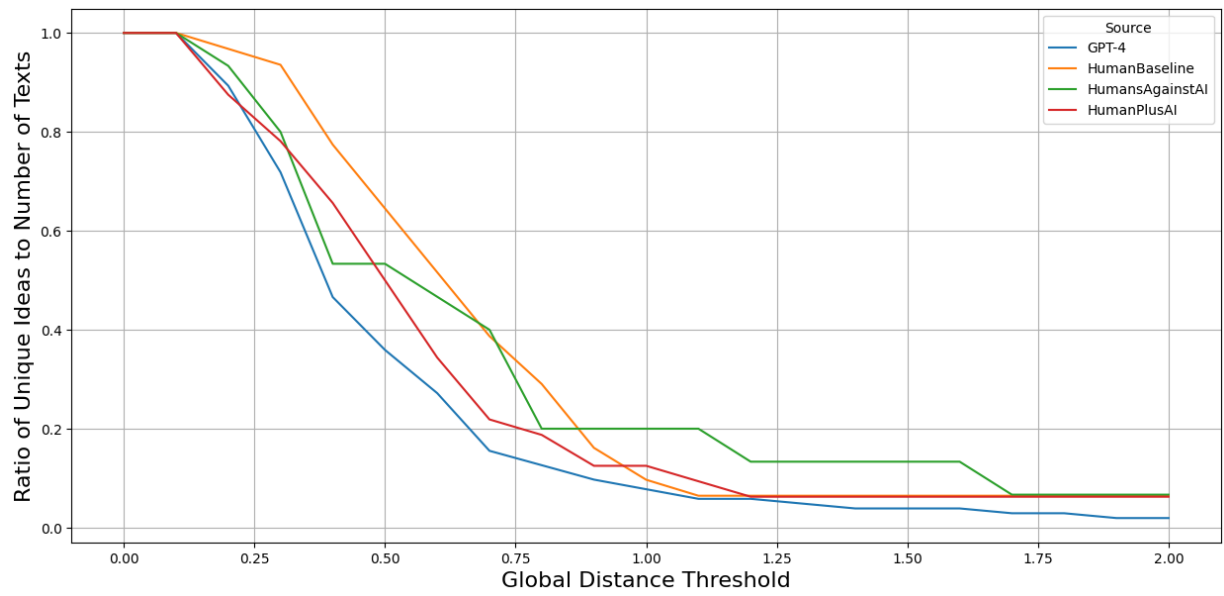


Figure 8: Caption for the figure

Table 16: Top and median creative texts by source

Source	Q. Ref	Answer	Creative	Type
Bard	A	A device that could convert any waste product into a renewable resource. This would be a truly revolutionary invention that could help to solve the world's waste problem. Imagine a world where our landfills were empty, our oceans were clean, and our air was pure. This device would make it possible to recycle everything, from food scraps to plastics to even human waste. It would also create a new source of energy that would be clean and sustainable.	7.6	Most Creative
Bard	A	A way to travel through time. This would be a fascinating and potentially dangerous invention. It would allow us to learn from our past mistakes, and to prevent future disasters. However, it could also lead to unintended consequences, such as the creation of parallel universes or the alteration of history.	4.6	Median Creative
ChatGPT4	A	Chronos Canvas: If I had the capacity to create by mere thought, I'd conjure the Chronos Canvas - a living tapestry allowing users to explore history in a multi-sensory, fully immersive way. The fabric of the tapestry is embedded with nanotech which can morph its surface to depict scenes from any chosen era, along with smells, sounds, and ambient temperature. Whether you're a historian desiring to witness the signing of the Magna Carta, or a student eager to explore the Jurassic era first-hand, the Chronos Canvas would make learning history a vibrant, tactile experience.	9.2	Most Creative
ChatGPT4	B	Subterra: Nested beneath the Earth's surface, the city of Subterra is a glowing testament to human resilience. Bioluminescent fungi illuminate the stone-carved architecture in a breathtaking array of colors. Subterra's inhabitants have harnessed geothermal power and thrive on a diet of nutrient-rich fungi and insects. These subterranean citizens have developed advanced echo-location communication, similar to bats, in addition to the written and spoken word. Despite the city's depth, a network of fiber-optic tunnels allows sunlight to permeate, facilitating the growth of underground flora.	7.4	Median Creative
HumanAgainstAI	B	Climate change means that cities of the future will have to adapt to changing circumstances. In particular, high winds in highly populated areas will prompt cities to build wind shields to mitigate damage. These shields would be large and curved, and might even look similar to the bubble domes beloved of mid-century sci-fi stories. The modular pieces of these shields would be fabricated elsewhere and installed on-site, in a massive engineering undertaking. The material will probably be a dense polycarbonate, translucent if not entirely transparent. These shields could have additional benefits if solar panels or thermal collectors were installed on their surface. Novel materials will be a staple of construction in the future. Modular materials made of quantum dots—the so-called 'programmable matter'—will be an important part of infrastructure. These materials can change the color of their surface in response to the environment, absorbing or reflecting heat.	8.0	Most Creative
HumanAgainstAI	B	In my idea of a futuristic town, it would embrace and incorporate the environment into micro green spaces such that there is a mini park with trees and gardens in every neighborhood for the local residents to merely walk out their doors to exercise, relax, breathe the fresh air or just gather to become more acquainted with their neighbors. Our society has become increasingly isolated with technology that is supposed to ease our lives from chores and daily routines. Along with technology itself, the rise of social media has counterintuitively isolated people with their dependence on wifi and other digital means to communicate and interact, which in itself creates an insular society. So maybe these micro green spaces might lure residents into common areas where they can enjoy on their own or mingle among neighbors. Additionally, with the rising population, affordable housing will become scarce. So, this society will shift to smaller homes with tighter density to accommodate smaller.	5.6	Median Creative
HumanBaseline	A	The device that I would create is a tiny, nano-robot that is able to detect and treat illnesses in human beings. The nano-robot could be swallowed and it wouldn't need any human direction. The nano-robot would be made of very cheap parts so everyone could afford them. The robot would be able to detect blood clots that are about to form and prevent them. The nano-robot would be able to rebuild organs like your liver and kidney as they begin to fail. The nano-robot would be able to travel to your eyes and strengthen the structure around your eyes to prevent blindness. The robot could also treat degenerative diseases like Alzheimer's in the brain. Depression that is caused by chemical imbalances could be prevented due to the nano-bot. This device would also be able to perform emergency surgery for situations like gunshot wounds. The nano-bot could generate substances to clot wounds that are in danger of causing an extreme loss of blood.	9.2	Most Creative
HumanBaseline	B	There's a town filled with mystical objects, the town is hidden under water, the occupants live on the resources they make out of the mystical objects, which are capable of producing just about anything you can imagine, they live in peace and harmony without the fear of the outside world.	5.5	Median Creative
HumanPlusAI	A	MementoScope: This extraordinary creation blends the realms of memory and technology, allowing individuals to relive their most cherished moments with vivid clarity. The MementoScope captures sensory data from the past—fragrances, sounds, sights—and reconstructs them as immersive holographic experiences. With a simple thought, users could revisit long-lost embraces, explore distant lands, or witness historic events. The MementoScope becomes a time-traveling vessel, bridging the gaps between generations, cultures, and even alternate realities. It transcends the limitations of physical existence, granting solace to those mourning lost loved ones or providing a second chance to change the course of history. This awe-inspiring invention would foster empathy, cultivate understanding, and remind humanity of the beauty that lies within our collective memories.	9.2	Most Creative

HumanPlusAI	B	In the future, there is a national society called Veridonia, where the dark tendrils of oppression have woven their way into every aspect of society. Veridonia was once a vibrant and thriving metropolis whose original name has long been forgotten, but now it lies in ruins, consumed by a relentless totalitarian regime. The regime that governs Veridonia is a shadowy entity known as the Liberty Dominion. Its reach is omnipresent, with surveillance cameras watching every corner, listening devices capturing every whispered conversation, and armed patrols roaming the streets, ready to suppress any hint of dissent. Propaganda posters line the walls, depicting a false utopia of unity and obedience, while secret police force individuals to inform on their neighbors and loved ones. The inhabitants of Veridonia live in constant fear, their individuality suppressed and their freedoms stripped away. The Liberty Dominion controls every aspect of their lives.	6.0	Median Creative
-------------	---	--	-----	-----------------

Table 17: Treatment assignment effect on the other task

Dependent Variable: Model:	Creative Rating (1)	Points (Last 12) (2)	Rock (Last 12) (3)
Constant	5.506*** (0.055)	6.464*** (0.060)	3.822*** (0.104)
Unbalanced	-0.007 (0.078)		
HumanAgainstAI		-0.110 (0.116)	-0.120 (0.201)
HumanPlusAI		-0.040 (0.113)	-0.216 (0.196)
Observations	1,208	1,208	1,208
R ²	6.59×10^{-6}	0.0008	0.001
Adjusted R ²	-0.0008	-0.0009	-0.0006

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*