Bohren, Noah; Hakimov, Rustamdjan; Lalive, Rafael

Working Paper

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

IZA Discussion Papers, No. 17302

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Bohren, Noah; Hakimov, Rustamdjan; Lalive, Rafael (2024) : Creative and Strategic Capabilities of Generative AI: Evidence from Large-Scale Experiments, IZA Discussion Papers, No. 17302, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/305744

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17302

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

Noah Bohren
Rustamdjan Hakimov
Rafael Lalive

SEPTEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17302

Creative and Strategic Capabilities 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

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 Capabilities of Generative AI: Evidence from Large-Scale Experiments*

Generative artificial intelligence (AI) has made substantial progress, but its full capabilities remain unclear, and we still lack a comprehensive understanding of how people augment productivity with AI and perceive AI-generated outputs. 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, creativity, strategic skill, experiment, algorithm-aversion

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 100018 207722), 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 creativity 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 predefined 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 HumanAgainsAI 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 the 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 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 63% of cases but are significantly less able to correctly identify chatbots' responses as AI-generated (61% for ChatGPT and 37% 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 cannot 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

[1]We 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 outperformance 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 discusses potential concerns regarding the insights our study, and Section 5 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 1250 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 [2]. We accessed the AIs through their standard chat interface without adjusting parameters like temperature.

Human participants were randomly assigned to one of three treatments [3]:

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.

[2]Note that we pre-registered 200 responses per AI. However, slightly more responses were generated by our RAs, and we decided to include all of them in the analysis.

[3]Note that the actual number of participants in each treatment group slightly differs from the pre-registered targets. We initially aimed for 700, 300, and 300 participants, recruited via Prolific. However, we received 702, 304, and 315 complete responses, respectively, due to the intentional invitation of a larger pool of participants to account for potential dropout before task completion.

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.[4] 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 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

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

correctly distinguished the origin of 15 or more texts were rewarded with an additional bonus of £0.5.[5]

Each text was rated by online raters at least 10 times in all three experimental conditions.[6]

Finally, a secondary group of five research assistants (two students from an art school in Lausanne and three 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.[7] 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:

[5]Note 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.

[6]Every 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.

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

- **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 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 round, we used the prompt "For round [n], I choose [*ramdom 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 Ratings by Online Raters

We start with 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	% Prompt A	Avg Time	% Male	Avg Age
ChatGPT4	224	589	50%	-	-	-
Bard	216	400	50%	-	-	-
HumanBaseline	688	716	44.3%	5.98	49%	44
HumanAgainstAI	253	739	45.8%	6.67	48%	43.9
HumanPlusAI	309	860	43%	5.55	50.2%	44.7

Avg Length = Average number of characters
Avg Time = Average time taken in minutes
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"

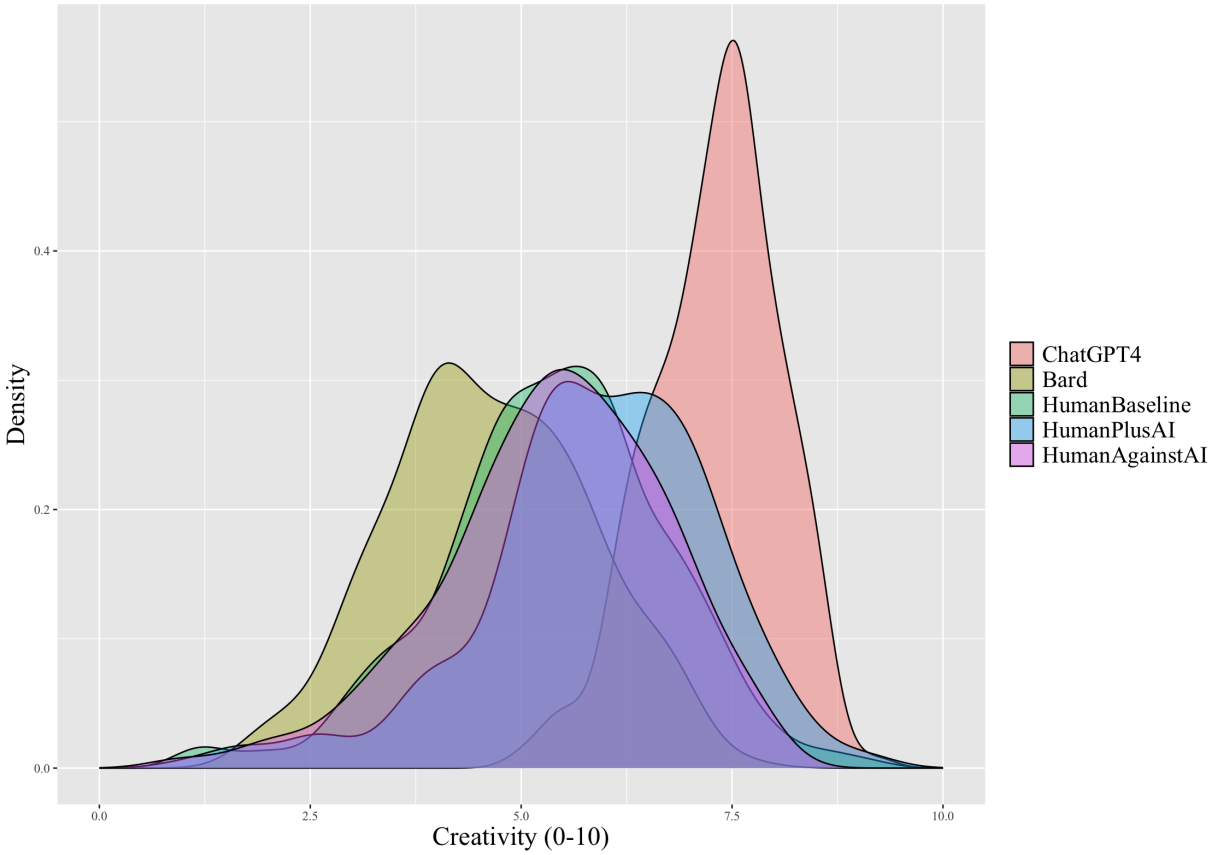
AI-generated texts are significantly shorter than those produced by the HumanBaseline group ($P < 0.001$),[8] whereas texts generated by the HumanPlusAI are significantly longer ($P < 0.001$). Furthermore, we observe statistically significant differences in the time required to produce texts: the HumanPlusAI 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 13).

Turning to the treatment comparison of creativity, Figure 1 illustrates the distribution of creativity ratings across sources, as evaluated by online raters.[9] 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.

[8]Unless 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.

[9]The 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 treatment differences are robust for 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 of heterogeneous treatment effects in Model (3) reveals a significant negative impact on female participants' creativity in the HumanAgainstAI group ($P < 0.05$), with no significant effects for male participants. These gender differences remain robust in Models (4) and (5), which control for individual characteristics. In the HumanPlusAI group (Models 6–8), female participants exhibit a larger creativity boost from AI augmentation than males, suggesting that AI may complement women's creative processes more effectively.

Table 2: Creativity ratings by online raters

Dependent Variable:	Creative Rating							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-0.924***	-0.936***	-0.936***			-0.936***		
	(0.051)	(0.050)	(0.050)			(0.050)		
ChatGPT4	1.879***	1.864***	1.864***			1.864***		
	(0.039)	(0.039)	(0.039)			(0.039)		
HumanPlusAI	0.500***	0.500***	0.500***	0.500***	0.474***			
	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)			
HumanPlusAI (Female)						0.565***	0.626***	0.581***
						(0.065)	(0.069)	(0.068)
HumanPlusAI (Male)						0.435***	0.394***	0.388***
						(0.073)	(0.077)	(0.076)
HumanPlusAI (Other)						0.495**	0.003	-0.060
						(0.242)	(0.290)	(0.288)
HumanAgainstAI	-0.032	-0.037				-0.037	-0.044	-0.047
	(0.054)	(0.054)				(0.054)	(0.054)	(0.053)
HumanAgainstAI (Female)			-0.152**	-0.147*	-0.133*			
			(0.073)	(0.075)	(0.073)			
HumanAgainstAI (Male)			0.075	0.079	0.054			
			(0.071)	(0.075)	(0.074)			
HumanAgainstAI (Other)			0.179	-0.248	-0.179			
			(0.213)	(0.267)	(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 rating 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 ratings by source and AI used

Source	Avg Creativity	N texts
Bard		
Bard	4.46	216
ChatGPT4		
ChatGPT (4.0)	7.24	224
HumanPlusAI		
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 14, 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 19.5% of participants completed their tasks in under 180 seconds. However, Figure 3 shows that those who spent less than 180 seconds were significantly less creative compared to the rest ($P < 0.001$).

In Table 15 in the appendix, we reproduced Table 2 after removing participants who spent less than 180 seconds 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 23%.

Figure 2: Histogram of time spent on writing task

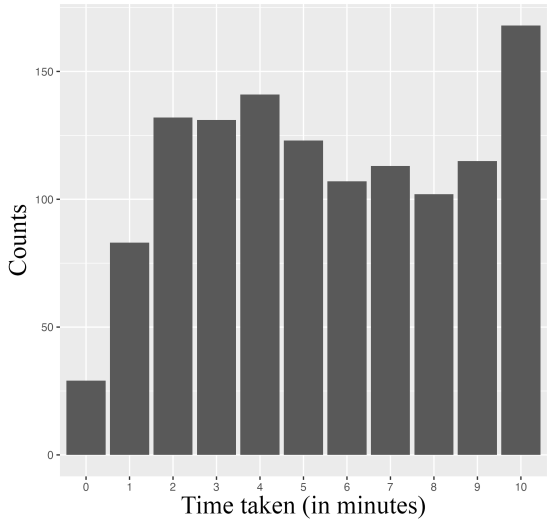
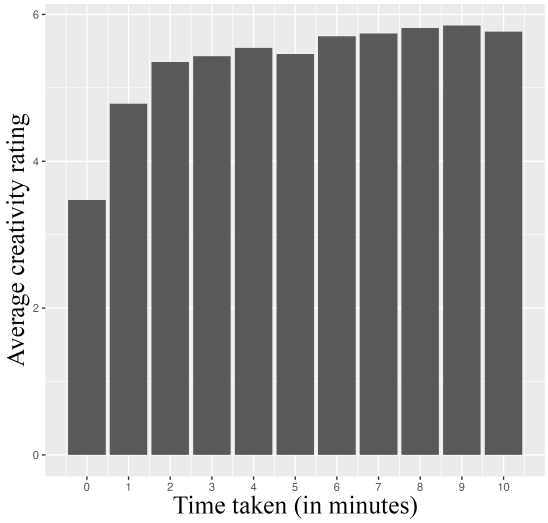


Figure 3: Average creativity rating by time taken



Notes: We convert seconds to minutes by rounding down to the nearest whole minute. e.g., both 121 seconds and 179 seconds will be converted to 2 minutes.

One might argue that average creativity is less important than top creativity, as the latter is more likely to spur innovation.[10] 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).

[10]While 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: Distribution of Responses in Top Creativity Percentiles by Source

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 (20.0%)	3 (17.6%)
HumanAgainstAI	11 (6.5%)	2 (2.3%)	0 (0%)

3.1.2 Ratings 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).[11] 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. Finally, Table 16 in the appendix reports the result of the regression analysis of the sub-dimensions of creativity.

[11]Note 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:	Creative Rating							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.176***	-1.177***	-1.176***			-1.178***		
	(0.064)	(0.065)	(0.065)			(0.065)		
ChatGPT4	0.551***	0.550***	0.551***			0.547***		
	(0.072)	(0.072)	(0.072)			(0.072)		
HumanPlusAI	0.142**	0.142**	0.142**	0.125*	0.118*			
	(0.067)	(0.067)	(0.067)	(0.068)	(0.067)			
HumanPlusAI (Female)						0.264***	0.224**	0.192**
						(0.082)	(0.091)	(0.091)
HumanPlusAI (Male)						0.008	0.034	0.052
						(0.091)	(0.103)	(0.101)
HumanPlusAI (Other)						0.281	-0.239	-0.196
						(0.435)	(0.526)	(0.491)
HumanAgainstAI	0.016	0.016						
	(0.073)	(0.073)						
HumanAgainstAI (Female)			-0.082	-0.137	-0.117	-0.082	-0.108	-0.095
			(0.097)	(0.101)	(0.098)	(0.097)	(0.103)	(0.100)
HumanAgainstAI (Male)			0.086	0.167	0.151	0.085	0.140	0.133
			(0.099)	(0.107)	(0.106)	(0.098)	(0.110)	(0.109)
HumanAgainstAI (Other)			0.680**	0.203	0.304	0.671**	0.035	0.160
			(0.297)	(0.420)	(0.440)	(0.299)	(0.440)	(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:	Creative Rating		
Model:	(1)	(2)	(3)
Age	0.000	0.002	0.002
	(0.002)	(0.002)	(0.002)
Gender Female	-0.063	-0.028	-0.017
	(0.048)	(0.047)	(0.047)
Gender Other	0.445***	0.424***	0.478***
	(0.149)	(0.143)	(0.148)
Emloyed	-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 rating 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 rating 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.

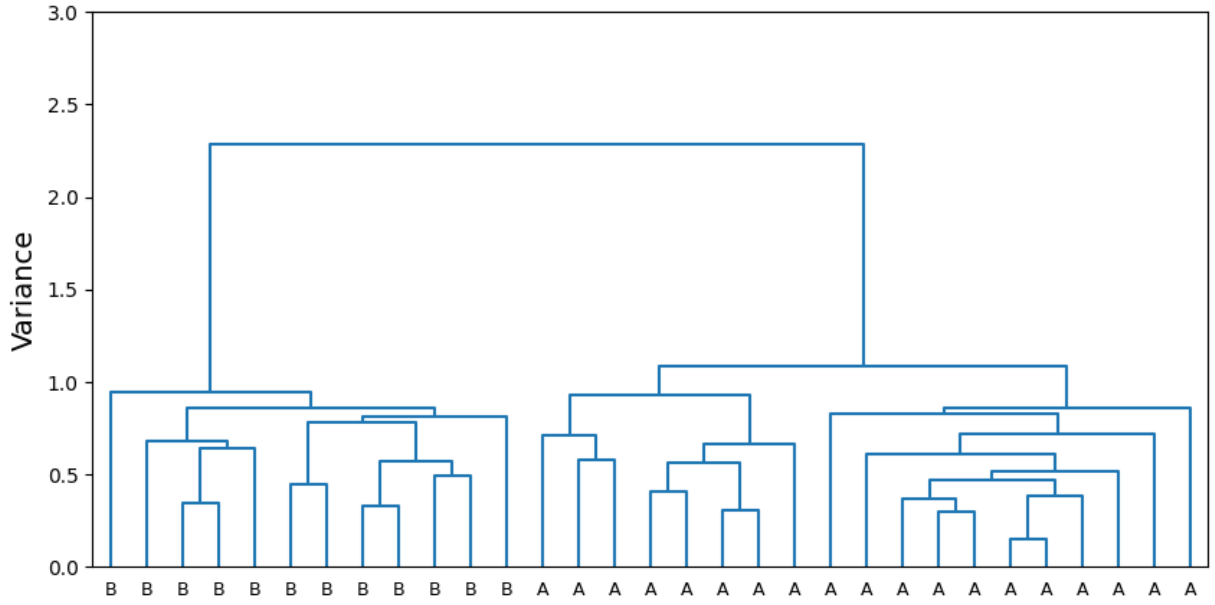
In this section, we use a combination of machine learning and deep learning approaches to measure the diversity of unique ideas that a source can generate. We start by converting each text into a numerical form using Sentence-BERT (sBERT) (Reimers and Gurevych, 2019).[12] This process transforms sentences 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

[12]Specifically, 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 address[13]. The clustering effectively separates texts by prompts, maintaining distinctions across different sources.

Figure 4: Example of Hierarchical Clustering Dendrogram using a selection of HumanBaseline 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 100 most creative texts for each source and (3) 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. Our analysis 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 a negligible difference in the variety of ideas produced across sources when considering all texts. However, when focusing on the most creative responses, humans demonstrate greater diversity, even at higher thresholds. This suggests that the most creative humans maintain an advantage over AI in producing a wide array of ideas on a given topic.

[13]A = "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"

Table 7: Comparison of Ratios of Unique Ideas Across Sources and Sample for Different Global Thresholds

Sample	ALL		Top 100 ‡		Top 10% †	
Threshold	(0.5)	(1)	(0.5)	(1)	(0.5)	(1)
ChatGPT-4	28.1%	6.7%	36.0%	7.0%	35.9%	7.8%
Bard	12.5%	5.1%	17.0%	6.0%	-	-
HumanBaseline	30.2%	6.7%	52.0%	10.0%	64.5%	9.7%
HumanAndAI	29.9%	7.3%	38.0%	6.0%	50.0%	12.5%
HumanAgainstAI	35.9%	6.9%	42.0%	8.0%	53.3%	20.0%

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

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 when compared to the best humans. 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

This section investigates whether our experimental treatments influence raters' evaluations of the creativity of generated responses. Table 8 presents the estimated treatment effects on each dimension of creativity. The first key finding is that responses that were grammatically corrected did not receive significantly different creativity ratings compared to uncorrected responses. This suggests that creativity assessments are independent of linguistic accuracy, providing reassurance that raters focus on the creative content rather than the technical quality of the text.

The second result is that informing raters that some responses might be AI-generated does not affect the overall creativity ratings. This indicates that the mere awareness of AI involvement does not alter the perceived creativity of the responses.

Table 8: Creativity ratings by online raters by raters treatments

Dependent Variable:	Creative Rating		Surprise Rating		Useful Rating		Original Rating	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	5.631***	4.124***	4.735***	3.603***	5.425***	4.481***	5.098***	3.617***
	(0.049)	(0.388)	(0.055)	(0.492)	(0.056)	(0.415)	(0.051)	(0.327)
CorrectedRater	-0.041	-0.024	-0.043	-0.026	0.042	0.043	-0.059	-0.040
	(0.069)	(0.068)	(0.076)	(0.075)	(0.076)	(0.075)	(0.071)	(0.070)
AIRaters	-0.062	-0.050	-0.101	-0.095	-0.044	-0.050	-0.057	-0.048
	(0.070)	(0.068)	(0.076)	(0.075)	(0.077)	(0.076)	(0.071)	(0.070)
Controls	None	Age Gender Risk	None	Age Gender Risk	None	Age Gender Risk	None	Age Gender Risk
Observations	63,950	63,627	63,950	63,627	63,950	63,627	63,950	63,627
R^2	8.09×10^{-5}	0.012	0.0002	0.012	0.0001	0.013	8.54×10^{-5}	0.013
Adjusted R^2	4.97×10^{-5}	0.012	0.0002	0.011	0.0001	0.013	5.41×10^{-5}	0.013

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

Table 9 further explores how raters' perceptions of the source of a response—whether they believe it was generated by AI or a human—affect their creativity ratings within the AIRaters treatment. This analysis aims to determine whether differences in ratings stem from raters' guesses about the response's origin, while controlling for the inherent creative qualities of each text. This distinction is crucial, as AI-generated texts have generally been rated more creative than human-generated texts, which could introduce bias if not properly accounted for.

To isolate the effect of the rater's guess, we leverage the fact that the texts are identical between the CorrectedRater and AIRaters groups. First, we compute the average creativity ratings (and sub-dimensions) for each text in the CorrectedRater group. These average ratings are then subtracted from the corresponding ratings in the AIRaters group, effectively demeaning the ratings to control for text-specific creative characteristics. This adjustment allows us to focus on the influence of the rater's guess on the demeaned ratings, ensuring that any observed differences are attributed to the raters' perceptions of the source of the text, rather than the inherent creative quality of the text.

We next regress these demeaned ratings on the raters' guesses (AI or human). The results show a notable pattern: for all creativity dimensions except for "surprise," texts that raters believed to be AI-generated received significantly lower ratings. This suggests a consistent negative bias against AI-generated content.

Table 9: Demeaned creativity ratings by rater guess of Human or AI source

Dependent Variable:	Demeaned Creative		Demeaned Surprise		Demeaned Original		Demeaned Useful	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.032	-1.276***	-0.030	-1.069**	0.047	-1.278***	-0.007	-1.093**
	(0.053)	(0.349)	(0.057)	(0.459)	(0.054)	(0.335)	(0.056)	(0.512)
Guess AI	-0.113**	-0.121**	-0.074	-0.084*	-0.106**	-0.115**	-0.208***	-0.210***
	(0.048)	(0.048)	(0.049)	(0.048)	(0.049)	(0.049)	(0.052)	(0.051)
Controls	None	Age Gender Risk	None	Age Gender Risk	None	Age Gender Risk	None	Age Gender Risk
Observations	21,851	21,839	21,851	21,839	21,851	21,839	21,851	21,839
R^2	0.0004	0.014	0.0002	0.015	0.0004	0.015	0.001	0.012
Adjusted R^2	0.0004	0.013	0.0001	0.014	0.0003	0.015	0.001	0.012

Notes: OLS regression of creativity ratings by online raters on treatment groups. 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 10). Raters classify the human-generated ideas as being human, i.e., correctly, in 63% of cases (no difference between HumanBaseline and HumanAgainstAI), 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 61% 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 (59%).

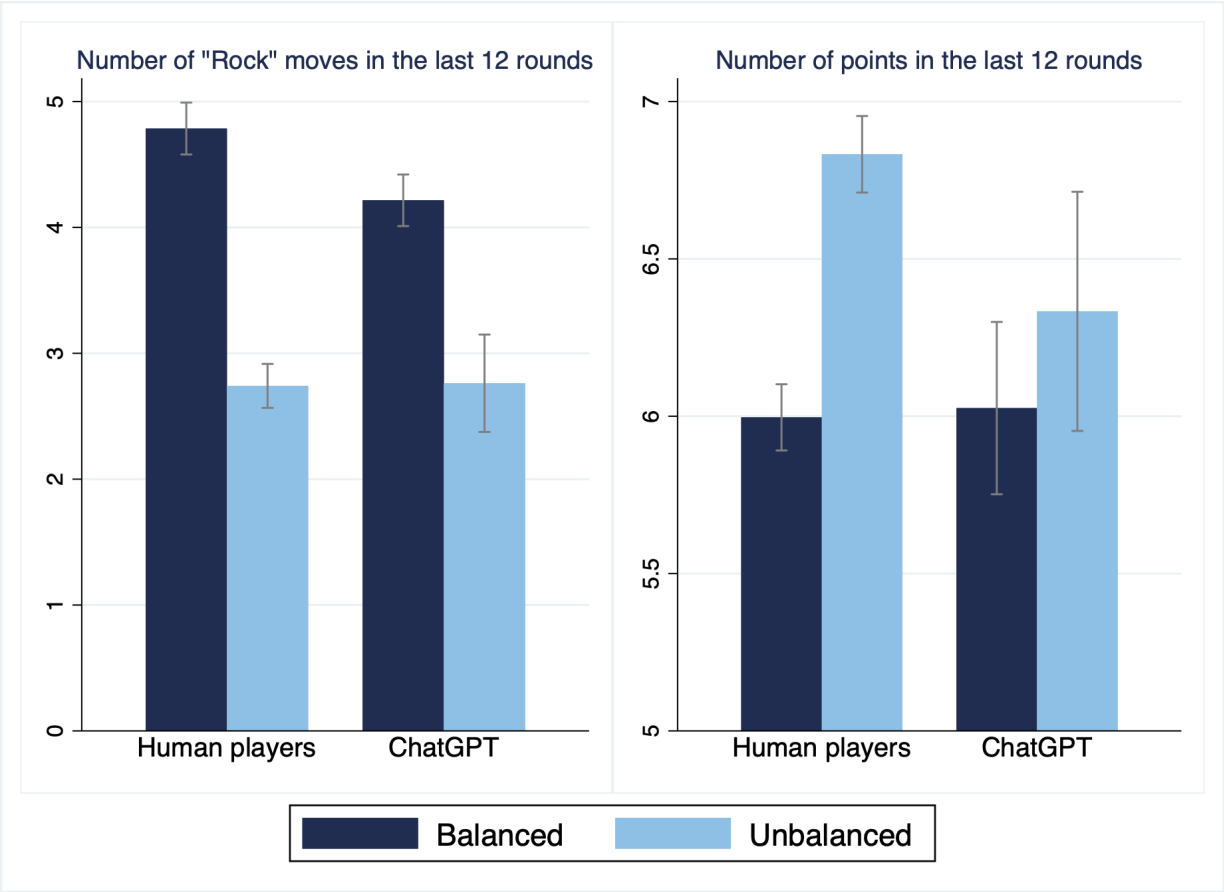
Table 10: Rater's average identification rates for responses as human or AI, by source

Source	Guess *Human*	Guess *AI*	Std. Error
Chatbots			
Bard	63.0%	37.0%	0.009
ChatGPT-4	39.3%	60.7%	0.009
Humans			
HumanBaseline	62.9%	37.1%	0.005
HumanAgainstAI	63.2%	36.8%	0.008
HumanPlusAI	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 and the first column of table 12 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 right panel of Figure 5 and Table 11 illustrates the points won across treat-

ments. 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 11: OLS for number of points in the last 12 rounds on source

Dependent Variable:	Points (Last 12)	
Model:	Balanced	Unbalanced
Constant	6.009***	6.304***
	(0.133)	(0.181)
Human	-0.012	0.529***
	(0.144)	(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 12 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 12: Regression analysis of move selection in final 12 rounds by source and treatment

Dependent Variable:	Rock (Last 12)	Paper (Last 12)	Scissors (Last 12)
Model:	(1)	(2)	(3)
Constant	4.786***	3.777***	3.437***
	(0.093)	(0.100)	(0.083)
ChatGPT4	-0.596**	-0.053	0.545**
	(0.247)	(0.265)	(0.219)
Unbalanced	-2.045***	2.046***	-0.001
	(0.132)	(0.142)	(0.117)
ChatGPT4 × Unbalanced	0.557	-1.425***	0.828**
	(0.376)	(0.404)	(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 Discussions

In this section, we address potential concerns regarding the insights our study provides and suggest directions for future research based on the limitations of the current study.

Tasks

One potential criticism is the specificity of the tasks chosen to measure creativity and strategic skills. While our creativity task selection was informed by a broad literature from psychology and economics, measuring creativity remains inherently challenging. We believe that the tasks capture key dimensions of creativity applicable across various domains, such as marketing slogan generation, storytelling, and scientific research. According to Charness and Grieco (2019), performance in our chosen task is robust to incentives, an important consideration when evaluating intrinsic creative potential.

For strategic skills, we deliberately selected a task that was straightforward and easy to understand, minimizing the need for extensive instructions. Our goal was to provide a task that allowed for real-time adaptation to the opponent's moves, making it an indicator of strategic intelligence. While ChatGPT-4 demonstrated some level of strategic adaptation, the task was limited to a set number of rounds in a relatively simple game of Rock-Paper-

Scissors. Its ability to adapt in more complex, long-term strategic environments remains an open area for future research.

Understanding the capabilities of AI in other creative and strategic tasks could be a fruitful area for future research, especially in contexts where interaction with AI is not solely reliant on text data. Exploring how AI performs in tasks involving visual, auditory, or mathematical inputs, for instance, could provide deeper insights into its capabilities.

Sample

Our study utilized a representative sample of the U.S. population, but future research could focus on samples of professionals in creative or strategic fields. While our broader sampling approach is valid for capturing everyday innovations in small to medium-sized enterprises, understanding how AI performs relative to top professionals in specific fields is critical for assessing AI's future role in the labor market. This line of inquiry, particularly in job-specific contexts, could yield important insights into the application of AI in specialized industries.

Novelty of technology

Another consideration is the relative novelty of AI technology, which may have influenced our findings, particularly regarding human reactions and perceptions of AI. As AI becomes more integrated into various sectors, it is likely that users will become more skilled at leveraging its capabilities. However, our findings, such as the lower creative performance among female participants and the harsher ratings for AI-generated responses, suggest that deeper psychological or societal biases may persist, even as AI technology matures. Addressing these biases will be crucial for ensuring the equitable adoption and effectiveness of AI in the future.

5 Conclusion

Our findings present a compelling case for the creative capabilities of ChatGPT-4. It significantly outperformed average human output in our open-ended creativity task and demonstrated the ability to adapt to biased opponents in a strategic setting.

In the creativity task, AI, such as ChatGPT-4, proved to be a valuable asset in generating novel ideas within established contexts. The potential benefits for organizations are evident, from streamlining brainstorming processes to improving the quality of idea generation. In the strategic task, AI showed emerging potential in decision-making, as ChatGPT-4 adapted its strategy over a 24-round series of interactions, suggesting its utility in providing real-time strategic advice.

However, while AI can augment human creativity, the effect size is relatively small, and human-AI collaboration does not yet outperform AI operating alone. This underscores the

importance of developing skills, such as effective prompting, to maximize the potential of AI-assisted creativity.

Our study also revealed a gender disparity in creative performance when competing with AI. As AI becomes more prevalent in the workplace, understanding how these social dynamics may exacerbate existing inequalities will be crucial. Targeted training or interventions may be needed to ensure that AI tools are accessible and beneficial across gender lines.

Finally, we uncovered a perceptual bias among human raters, who assigned lower scores to outputs they believed were AI-generated, reflecting public skepticism or resistance toward AI. As AI becomes more integrated into various sectors, addressing these biases will be essential to fully realize its benefits.

In conclusion, our study highlights the current capabilities of AI in creative and strategic tasks, while underscoring important implications for its adoption. Organizations can leverage AI to enhance creativity and decision-making, but success will depend on addressing social biases, optimizing human-AI collaboration, and ensuring equitable access to AI technologies.

References

Abada, I. and Lambin, X. (2023). Artificial intelligence: Can seemingly collusive outcomes be avoided? *Management Science*.

Arntz, M., Gregory, T., and Zierahn, U. (2016). The risk of automation for jobs in oecd countries: A comparative analysis.

Autor, D. H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, 29(3):3–30.

Autor, D. H. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American economic review*, 103(5):1553–1597.

Boden, M. (2001). Creativity and knowledge. *Creativity in education*, pages 95–102.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2):347–356.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative ai at work. Technical report, National Bureau of Economic Research.

Charness, G. and Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2):454–496.

Charness, G. and Grieco, D. (2024). Creativity and ai. *Available at SSRN 4686415*.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Chugunova, M. and Luhan, W. J. (2024). Ruled by robots: Preference for algorithmic decision makers and perceptions of their choices. *Public Choice*, pages 1–24.

Dargnies, M.-P., Hakimov, R., and Kübler, D. (2023). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). 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*, (24-013).

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 2156. Springer.

Doshi, A. R. and Hauser, O. (2023). Generative artificial intelligence enhances creativity. *Available at SSRN*.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Felten, E., Raj, M., and Seamans, R. (2023). How will language modelers like chatgpt affect occupations and industries? *arXiv preprint arXiv:2303.01157*.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Girotra, K., Meincke, L., Terwiesch, C., and Ulrich, K. T. (2023a). Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071*.

Girotra, K., Meincke, L., Terwiesch, C., and Ulrich, K. T. (2023b). Ideas are dimes a dozen: Large language models for idea generation in innovation. *The Wharton School Research Paper Forthcoming*. Available at SSRN: https://ssrn.com/abstract=4526071 or http://dx.doi.org/10.2139/ssrn.4526071.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics*, 118(3):1049–1074.

Greiner, B., Grünwald, P., Lindner, T., Lintner, G., and Wiernsperger, M. (2024). *Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study*. Vienna University of Economics and Business, Department of Strategy.

Guilford, J. P. (1975). Varieties of creative giftedness, their measurement and development. *Gifted child quarterly*, 19(2):107–121.

Huang, F., Kwak, H., and An, J. (2023). 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*, pages 294–297.

Kiron, D. and Schrage, M. (2019). Strategy for and with ai. *MIT Sloan Management Review*, 60(4).

Kuzman, T., Mozetic, I., and Ljubešic, N. (2023). Chatgpt: beginning of an end of manual linguistic data annotation. *Use Case of Automatic Genre Identification. ArXiv abs/2303.03953*.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annu. Rev. Econ.*, 3(1):601–630.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Saccardo, S., Pietrasz, A., and Gneezy, U. (2018). On the size of the gender difference in competitiveness. *Management Science*, 64(4):1541–1554.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267:664–681.

Shakked, N. and Whitney, Z. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Smith, A. and Anderson, J. (2014). Ai, robotics, and the future of jobs.

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

6 Appendix

Table 13: 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 14: Sub-dimensions of creativity by online raters

Dependent Variable:	original							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.039***	-1.056***	-1.056***			-1.056***		
	(0.051)	(0.051)	(0.051)			(0.051)		
ChatGPT4	1.962***	1.942***	1.942***			1.942***		
	(0.043)	(0.043)	(0.043)			(0.043)		
HumanPlusAI	0.335***	0.335***	0.336***	0.335***	0.304***			
	(0.052)	(0.052)	(0.052)	(0.052)	(0.051)			
HumanPlusAI (Female)						0.390***	0.433***	0.379***
						(0.068)	(0.072)	(0.070)
HumanPlusAI (Male)						0.283***	0.266***	0.256***
						(0.069)	(0.073)	(0.072)
HumanPlusAI (Other)						0.290	-0.347	-0.337
						(0.236)	(0.289)	(0.295)
HumanAgainstAI	-0.027	-0.034				-0.034	-0.044	-0.047
	(0.054)	(0.053)				(0.053)	(0.053)	(0.052)
HumanAgainstAI (Female)			-0.122*	-0.123*	-0.111			
			(0.071)	(0.074)	(0.071)			
HumanAgainstAI (Male)			0.044	0.056	0.034			
			(0.069)	(0.073)	(0.072)			
HumanAgainstAI (Other)			0.278	-0.243	-0.205			
			(0.234)	(0.292)	(0.306)			
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	0.393	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:	surprise							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.040***	-1.061***	-1.061***			-1.061***		
	(0.048)	(0.048)	(0.048)			(0.048)		
ChatGPT4	1.627***	1.603***	1.603***			1.603***		
	(0.043)	(0.042)	(0.042)			(0.042)		
HumanPlusAI	0.193***	0.193***	0.193***	0.193***	0.163***			
	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)			
HumanPlusAI (Female)						0.217***	0.253***	0.203***
						(0.064)	(0.067)	(0.067)
HumanPlusAI (Male)						0.172***	0.159**	0.150**
						(0.064)	(0.068)	(0.067)
HumanPlusAI (Other)						0.141	-0.396	-0.378
						(0.248)	(0.312)	(0.317)
HumanAgainstAI	0.085*	0.077				0.077	0.065	0.067
	(0.052)	(0.051)				(0.051)	(0.051)	(0.050)
HumanAgainstAI (Female)			0.013	0.016	0.026			
			(0.069)	(0.071)	(0.069)			
HumanAgainstAI (Male)			0.145**	0.147**	0.138**			
			(0.066)	(0.070)	(0.069)			
HumanAgainstAI (Other)			0.086	-0.433	-0.412			
			(0.241)	(0.305)	(0.309)			
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	0.397	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:	useful							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-0.144***	-0.179***	-0.179***			-0.178***		
	(0.052)	(0.049)	(0.049)			(0.049)		
ChatGPT4	1.024***	0.985***	0.985***			0.985***		
	(0.041)	(0.039)	(0.039)			(0.039)		
HumanPlusAI	0.362***	0.363***	0.363***	0.362***	0.365***			
	(0.046)	(0.046)	(0.046)	(0.046)	(0.047)			
HumanPlusAI (Female)						0.438***	0.421***	0.425***
						(0.055)	(0.059)	(0.060)
HumanPlusAI (Male)						0.284***	0.305***	0.309***
						(0.066)	(0.069)	(0.070)
HumanPlusAI (Other)						0.449***	0.292	0.220
						(0.125)	(0.190)	(0.192)
HumanAgainstAI	-0.091*	-0.104**				-0.104**	-0.105**	-0.121**
	(0.053)	(0.052)				(0.052)	(0.051)	(0.051)
HumanAgainstAI (Female)			-0.134*	-0.186**	-0.195***			
			(0.071)	(0.073)	(0.072)			
HumanAgainstAI (Male)			-0.082	-0.022	-0.047			
			(0.067)	(0.071)	(0.071)			
HumanAgainstAI (Other)			0.065	-0.020	-0.006			
			(0.223)	(0.251)	(0.260)			
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	0.372	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 attitudes, 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: Creativity ratings by online raters (Writing time > 180s)

Dependent Variable:	Creative Rating							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.122***	-1.150***	-1.150***			-1.151***		
	(0.050)	(0.050)	(0.050)			(0.050)		
ChatGPT4	1.682***	1.650***	1.650***			1.650***		
	(0.039)	(0.039)	(0.039)			(0.039)		
HumanPlusAI	0.458***	0.449***	0.449***	0.459***	0.424***			
	(0.054)	(0.054)	(0.054)	(0.054)	(0.053)			
HumanPlusAI (Female)						0.438***	0.520***	0.487***
						(0.071)	(0.074)	(0.072)
HumanPlusAI (Male)						0.451***	0.402***	0.366***
						(0.075)	(0.078)	(0.077)
HumanPlusAI (Other)						0.646**	0.199	0.119
						(0.253)	(0.312)	(0.314)
HumanAgainstAI	-0.117**	-0.127**				-0.127**	-0.137**	-0.126**
	(0.056)	(0.055)				(0.055)	(0.055)	(0.052)
HumanAgainstAI (Female)			-0.198***	-0.144*	-0.111			
			(0.075)	(0.078)	(0.073)			
HumanAgainstAI (Male)			-0.071	-0.111	-0.130*			
			(0.072)	(0.075)	(0.074)			
HumanAgainstAI (Other)			0.083	-0.447	-0.280			
			(0.215)	(0.279)	(0.306)			
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	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 16: Sub-dimensions of creativity ratings by RAs

Dependent Variable:	Original Rating							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.426***	-1.437***	-1.437***			-1.439***		
	(0.079)	(0.081)	(0.081)			(0.081)		
ChatGPT4	0.042	0.028	0.028			0.025		
	(0.094)	(0.093)	(0.093)			(0.093)		
HumanPlusAI	-0.416***	-0.412***	-0.412***	-0.439***	-0.457***			
	(0.088)	(0.086)	(0.086)	(0.089)	(0.088)			
HumanPlusAI (Female)						-0.330***	-0.403***	-0.467***
						(0.109)	(0.123)	(0.123)
HumanPlusAI (Male)						-0.497***	-0.465***	-0.437***
						(0.116)	(0.134)	(0.134)
HumanPlusAI (Other)						-0.424	-0.733	-0.671
						(0.493)	(0.725)	(0.722)
HumanAgainstAI	0.077	0.071						
	(0.093)	(0.091)						
HumanAgainstAI (Female)			0.044	-0.048	-0.010	0.043	-0.037	-0.013
			(0.118)	(0.124)	(0.122)	(0.118)	(0.127)	(0.126)
HumanAgainstAI (Male)			0.080	0.134	0.122	0.080	0.127	0.129
			(0.124)	(0.136)	(0.135)	(0.124)	(0.139)	(0.138)
HumanAgainstAI (Other)			0.448	0.240	0.310	0.442	0.107	0.215
			(0.437)	(0.560)	(0.546)	(0.438)	(0.671)	(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	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:	Surprise Rating							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.202***	-1.211***	-1.211***			-1.213***		
	(0.073)	(0.074)	(0.074)			(0.074)		
ChatGPT4	0.065	0.052	0.052			0.051		
	(0.088)	(0.087)	(0.087)			(0.087)		
HumanPlusAI	-0.338***	-0.334***	-0.334***	-0.344***	-0.348***			
	(0.081)	(0.080)	(0.080)	(0.083)	(0.082)			
HumanPlusAI (Female)						-0.252**	-0.318***	-0.365***
						(0.104)	(0.117)	(0.118)
HumanPlusAI (Male)						-0.398***	-0.337***	-0.293**
						(0.105)	(0.121)	(0.120)
HumanPlusAI (Other)						-0.801*	-1.275**	-1.203*
						(0.460)	(0.631)	(0.644)
HumanAgainstAI	0.090	0.085						
	(0.087)	(0.086)						
HumanAgainstAI (Female)			0.043	-0.037	-0.026	0.042	-0.030	-0.032
			(0.111)	(0.115)	(0.112)	(0.111)	(0.118)	(0.116)
HumanAgainstAI (Male)			0.111	0.186	0.178	0.111	0.190	0.197
			(0.119)	(0.131)	(0.131)	(0.119)	(0.134)	(0.133)
HumanAgainstAI (Other)			0.415	0.315	0.418	0.411	-0.104	0.036
			(0.374)	(0.501)	(0.489)	(0.375)	(0.576)	(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	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:	Useful Rating							
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bard	-1.041***	-1.047***	-1.046***			-1.049***		
	(0.068)	(0.069)	(0.069)			(0.069)		
ChatGPT4	0.322***	0.314***	0.315***			0.310***		
	(0.073)	(0.072)	(0.072)			(0.073)		
HumanPlusAI	0.006	0.009	0.010	-0.023	-0.025			
	(0.067)	(0.066)	(0.066)	(0.067)	(0.067)			
HumanPlusAI (Female)						0.158*	0.117	0.083
						(0.082)	(0.091)	(0.092)
HumanPlusAI (Male)						-0.154*	-0.170*	-0.141
						(0.089)	(0.102)	(0.100)
HumanPlusAI (Other)						0.175	-0.077	-0.072
						(0.389)	(0.441)	(0.422)
HumanAgainstAI	0.008	0.005						
	(0.072)	(0.072)						
HumanAgainstAI (Female)			-0.097	-0.187*	-0.180*	-0.098	-0.144	-0.147
			(0.097)	(0.103)	(0.101)	(0.097)	(0.104)	(0.102)
HumanAgainstAI (Male)			0.091	0.171*	0.145	0.090	0.126	0.111
			(0.095)	(0.104)	(0.105)	(0.094)	(0.107)	(0.107)
HumanAgainstAI (Other)			0.475	0.170	0.252	0.461	0.135	0.224
			(0.358)	(0.412)	(0.422)	(0.360)	(0.429)	(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	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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

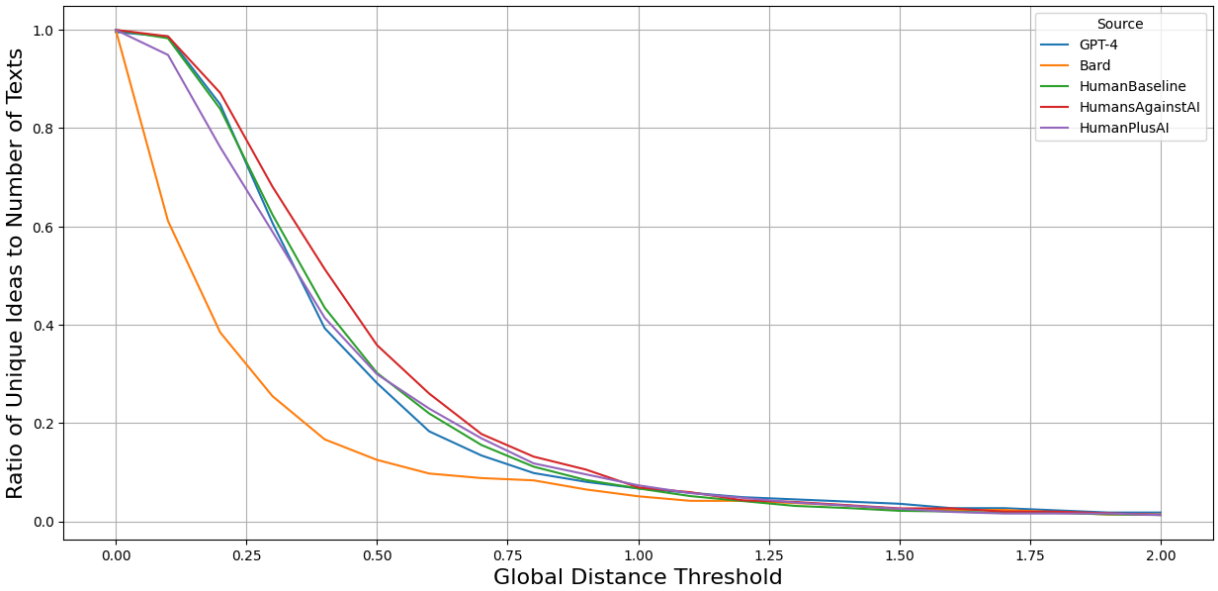


Figure 6: Percentage of unique ideas for ALL responses across global thresholds

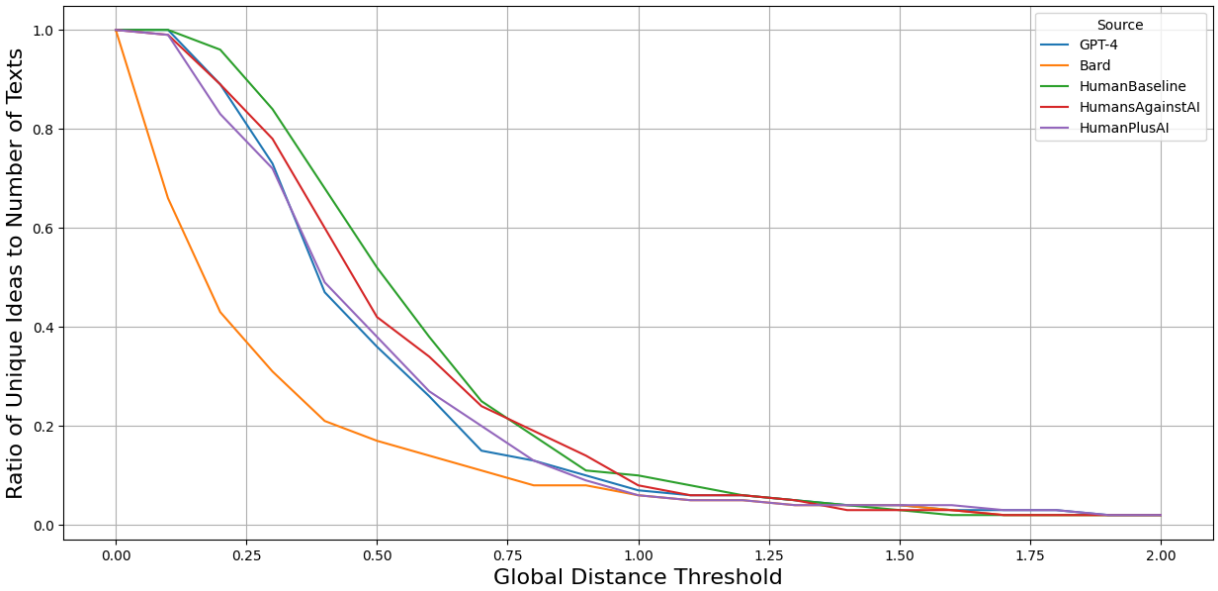


Figure 7: Percentage of unique ideas for TOP 100‡ responses across global thresholds

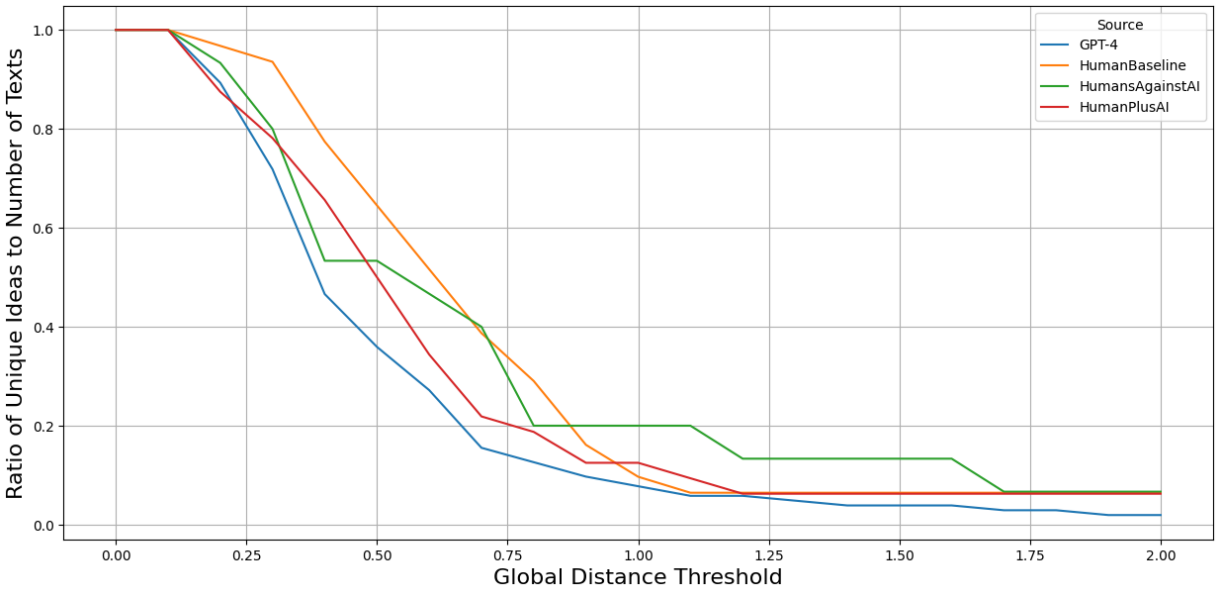


Figure 8: Percentage of unique ideas for TOP 10%† responses across global thresholds