CREATIVITY AND AI*

Gary Charness and Daniela Grieco

January 21, 2026

Abstract: *We investigate whether AI systems outperform humans in creative tasks that vary in their degree of "openness." To this end, we generated creative responses using three versions of ChatGPT and recruited 738 participants to blindly evaluate six creative answers randomly drawn from three pools—comprising 160 responses each—generated by both humans and AI. This process yielded 4,428 individual evaluations. Our results show that, regardless of the GPT version employed, human-generated responses achieve significantly and substantially higher average scores than machine-generated responses in open tasks. Conversely, AI-generated responses outperform human ones in closed tasks. Furthermore, we estimate that human imagination accounts for between 22% and 45% of the creative score in open tasks.*

Keywords: creativity; Artificial Intelligence; experiment; text analysis; structural estimation.

JEL codes: C91, O33, O39

* Corresponding author: Daniela Grieco, University of Milano, Department of Law C. Beccaria, via Festa del Perdono 3, 20122, Milano, Italy. Email: daniela.grieco@unimi.it Gary Charness, deceased on May 17, 2024, beloved friend, coauthor and mentor. A permission to submit the paper has been received from the co-author's estate on November 23, 2024.
We thank the University of Milano for financial support. The study received ethics approval from the University of Milano Ethics Committee (Parere 70/2023). The experiment was pre-registered in the AsPredicted database on June 26, 2023 under trial number #136955 (https://aspredicted.org/HMT_X2P); the follow-up experiments using an updated version of GPT (GPT-4o and GPT-4.5) have been pre-registered under trial number #194851 (https://aspredicted.org/2fgf-9kj5.pdf) on October 19, 2024 and under trial number #225528 (https://aspredicted.org/9wj8-3nsz.pdf) on April 29, 2025. For extremely valuable comments, we thank the Editor, three anonymous referees, Ted Bergstrom, Francesco Bripi, Nir Chemaya, David Cooper, John List, Daniel Martin, Telmo Pievani, Cheng-Zhong Qin, Ravi Vora, and the participants to the 2024 ESADE Workshop on Behavioral and Experimental Economics in Barcelona, the 2024 Experimental and Behavioral Economics Workshop in Santa Barbara, the 2025 Workshop "Advancing Behavioral Insights: A Tribute to Gary Charness" in Tucson, the 11th Annual CBESS Workshop on Behavioural Game Theory, the 2025 CIMEO Summer School in Experimetrics & Behavioral Economics, and seminars at the University of Milano, University of Strasbourg, University of the Balearic Islands.

1. Background

Creativity is a fundamental feature of human intelligence and represents a vital skill not only for successful job performance, but also for everyday challenges requiring new solutions. Creativity is grounded in a complex concurrence of capacities like the "association of ideas, reminding, perception, analogical thinking, searching a structured problem-space [...]" (Boden, 1998, p. 347). Creative achievements require developing ideas about what is valuable or interesting that moves beyond what is considered a "standard" (Jennings, 2010).

Artificial Intelligence (AI) is one of the most significant recent innovations in technology. It is that branch of computer science that simulates human intelligence processes like problem-solving and learning, so that computer programs can perform these processes instead of humans. AI language models such as OpenAI's Generative Pre-trained Transformer (GPT) are increasingly used to help perform tasks previously conducted by people (OECD, 2023) and have become a powerful tool for *creative endeavors* in several domains (Bubeck et al., 2023). GPT, and its conversational agent ChatGPT, exploit huge quantities of data, algorithms capable of learning from these data, and the computational power to do it. GPT-3.5 was released in November 2022, impressing people and attracting media attention. Relying on a huge amount of data and "machine learning" it was able to solve tasks, produce drawings or animations, answer questions, have a conversation, write essays about any possible topic, classify objects from previously unseen classes, review research papers (Dai et al., 2023). As a result, many people started testing it. In the moment we write, GPT-4.5 is the latest and most advanced model, launched in February 2025. According to their statement releases, both GPT-4.o and GPT-4.5 are reported to have improved their ability to recognize patterns, draw connections, generate creative insights, and significantly reduce hallucinations, i.e. produce content that is nonsensical or untruthful in relation to certain sources, relative to previous models[1].

There is a great deal of concern that AI will soon take over from humans in all or nearly all tasks. While the LLMs have already been recognized as successful in substituting humans in many endeavors like the ones mentioned (see for instance Gilardi et al., 2023), their ability to deal with creative tasks is still questioned, as emerges from the growing literature on creativity and AI (Bohren et al., 2024; Boussioux et al., 2024; Doshi and Hauser, 2023; Fu et al., 2024; Girotra et al., 2023; Joosten et al., 2024; Magni et al., 2024; Zhou and Lee, 2024), whose contributions will be discussed below in relation with the design and results of this paper. AI is programmed to process information in a certain way and achieve a particular result, and cannot deviate from the instructions received:

[1] https://openai.com/index/hello-gpt-4o/ and https://openai.com/index/introducing-gpt-4-5/

LLM is *de facto* a statistical exercise where the words of the query made to ChatGPT are converted into numbers, grouped with other words with similar meaning, and connected with other words according to the learned frequency of these associations. As such, they cannot be spontaneous or unpredictable like human creativity. Since the association relies on a certain probability, the LLM's output is not entirely predictable. The randomness of the text generated depends on the setting of parameters, in particular of temperature, which controls the "creativity" or randomness of the text generated.[2]

Creativity is about finding new solutions to problems that others may not have considered yet, subverting the rules, looking at a question from different perspectives and "thinking outside the box". Cognitive psychology identifies imagination as the source of creativity—a unique prerogative of humans that arises from social understanding and communication. According to the anthropologist Ian Tattersal, "Only human beings are able arbitrarily to combine and recombine mental symbols and to ask themselves questions such as "What if?" And it is the ability to do this, above everything else, that forms the foundation of our vaunted creativity" (Tattersall, 2001, p. 60). What distinguishes the human species is our capacity to imagine possible alternative worlds (Pievani, 2026).

In contrast, computer programs that use AI are coded to reach the (exact) results they are told to achieve relying on information that already exists. This is why they are typically used in tasks where accuracy is needed. Creativity is instead hard to reduce to a set of instructions or a mathematical formula and represents a human process open to many interpretations and viewpoints.

Our study tests whether AI outperforms humans in creative tasks with a different degree of "openness". As defined in Charness and Grieco (2019), "a true closed problem is one that is presented to the participant, when the method for solving the problem is known […]", like in what Collins and Amabile (1999) call "algorithmic" tasks. In contrast, open problems occur when the participant is required to find, invent, or discover the problems" (Unsworth, 2001). To be clear, we do not claim to have exhaustively examined all measures of creativity in all environments. Yet, we have results that serve as clear initial evidence on this topic. We use the same creativity methodology as in our previous articles (Charness and Grieco, 2019, 2022, 2023). In our case, the new subjects score the creativity of items previously produced by an unknown AI or human.

Our results show that humans' average performance in the Open task is significantly better than that of AI, regardless of the version of GPT used. Strikingly, this is entirely reversed in the closed task. The difference-in-difference across treatments is large, ranging between eight and ten times the

[2] A higher temperature results in more diverse output, while a lower temperature makes the output more deterministic: specifically, temperature affects the probability distribution over the possible tokens at each step of the generation process, with a temperature of 0 making the model completely deterministic, always choosing the most likely token.

standard errors depending on the ChatGPT version. To shed light on the drivers of creative output for humans and machines, we sketch a simple model that captures human versus AI preferences depending on the degree of task openness and we structurally estimate the two key preferences parameters: the extent to which humans and machines are affected by the openness of the creative task, and the relative weight of human imagination with respect to what is prescribed by instructions.

Our structural estimates show that imagination contributes between 22% and 45% of the creative score that humans reach in the open task, with human ability of producing novel, unusual ideas representing a key ingredient of their success with respect to AI machines, and AI improvement over time consisting of creative answers that have become better elaborated and less nonsensical, but that are not more novel and original. Human imagination has instead no role in sufficiently closed tasks, like the one we present in this experiment.

As a final test, we ask AI to evaluate the same answers and found that there was no systematic improvement in the precision of AI evaluations over time. However, when given more detailed prompts that restricted the scope of evaluation to very specific aspects of creativity, AI evaluations reflected those of humans.

Section 2 of this paper presents our experimental designs and results and in section 3 we put forward a formal model and perform our structural estimation. Section 4 describes the results when it is AI, and not human subjects, that evaluates creative answers. In section 5, we offer some discussion and conclude.

2. The experiment

2.1. Experimental design

The experiment has a 2x2 design, with two treatments ("Closed" vs. "Open"), differing in the degree of openness of the creative task, and two agents (humans vs. AI machines) producing answers to the task, with four conditions in total: AI-Closed (AI-C), AI-Open (AI-O), Humans-Closed (H-C), Humans-Open (H-O). A total amount of 738 subjects were recruited on Prolific and took part in one of the three sessions of the experiment: in Session I, the AI answers were produced with GPT-3.5 (this session was run in July 2023); in Session II, the AI answers were produced with GPT-4o (October 2024); in Session III, the AI answers were produced with GPT-4.5 (April 2025). Each session consisted of two sub-sessions: one for the Closed treatment, and one for the Open treatment. In each sub-session, subjects evaluated[3] six answers (three produced by humans and three produced by AI) drawn at random from a set of 80 answers (40 produced by human beings, 40 by AI) and

[3] We follow the literature that defines creativity as something whose novelty is recognized by others (Amabile et al., 1990; Mumford, 2003) and use creativity scores assigned by external judges as the main outcome variable.

presented in a random order, without knowing whether the answer was produced by a human being or by AI[4]. In total, 4428 evaluations (1506 in the first session, 1482 in the second, and 1440 in the second) were provided, with each answer receiving (at least[5]) nine independent evaluations. Appendix A provides a few illustrative examples of AI answers (generated by GPT-3.5, GPT-4o and GPT-4.5, varying the temperature level in the same way across tasks and GPT versions) and human answers in the different treatments. The experimental protocols are in the Supplementary Material, but we report the key aspects of the experimental design below.

Closed Treatment. In the Closed treatment, subjects evaluate the answers that 40 participants gave in Charness and Grieco (2019)'s laboratory experiment[6] to the following task: *"Write a creative story using compulsorily the following words: house, zero, forgive, curve, relevance, cow, tree, planet, ring, send."* This task is classified as "closed" (in relative terms, as opposed to the open task described below) because it provides requirements in the form of the ten words to be used mandatorily, making the task a problem to "solve", but keeping it comparable to the open task below.

The same number of answers (40) to this task were produced using AI (specifically, GPT-3.5 in the first session, GPT-4o in the second session, and GPT-4.5 in the third session), varying temperature from 0 to 2 (maximum possible value) but setting the same values across tasks and GPT versions, and leaving the other parameters at their default value (with the exception of the length parameter).[7] In the first and second session, GPT received as prompt exactly the same sentence used in instructions given to humans (Closed task: *"Write a creative story using compulsorily the following words: house, zero, forgive, curve, relevance, cow, tree, planet, ring, send.*), while in the third session an explicit request to be creative was added (*"Please be as creative as possible.").* It should be noted that human subjects received tournament incentives. Subjects made their evaluation, assigning a "creativity score" in a 1-10 scale.

[4] Subjects were not told that there was the chance that the answer(s) could have been produced by AI, since there is evidence showing that people may ascribe lower creativity to a product when they are told that the producer is an AI rather than a human (Magni et al., 2024).

[5] We collected 4428 total evaluations instead of the planned 4320 (corresponding to 9 evaluations for each of the 160 answers per session) because we ended up recruiting 738 respondents instead of the 720 initially planned because of simultaneous enrollments of respondents exceeding the requested amount.

[6] The 40 answers taken from Charness and Grieco (2019) correspond to the 40 answers in the verbal closed task produced by subjects receiving the Tournament treatment. In the Closed Task, subjects could also select a math sub-task, but only four subjects decided to complete it in the presence of monetary incentives. The average score is not significantly different from that in the verbal sub-task (5.250 vs. 5.975, $Z = 0.370$, $p = .711$, two-sample Wilcoxon rank-sum test). The instructions used for these experiments are provided in Appendix B1.

[7] Default length was set at 256 words, but we raised it to 400 in both treatments since we realized that answers initially produced in the Closed treatment were not complete; we raised it up to 400 after computing the average length in human answers and referencing Charness and Grieco (2019), which shows that response length has a minimal, linear effect on creative scores.

Open Treatment. In the Open treatment, subjects evaluate the answers that 40 participants gave in Charness and Grieco (2019)'s experiment[8] to the following task: *"If you had the talent to invent things just by thinking of them, what would you create and why?"*. This task is categorized as "open" because subjects are completely free to invent new things with no constraints. Again, human subjects received tournament incentives, and the same number of answers (40) were produced using GPT-3.5, GPT-4o and GPT-4.5, and subjects made their evaluation without knowing who produced each answer. As above, in the first and second session, GPT received as prompt exactly the same sentence used in instructions given to humans (task: *"If you had the talent to invent things just by thinking of them, what would you create and why?"*), while in the third session an explicit request to be creative was added (task: *"If you had the talent to invent things just by thinking of them, what would you create and why?* Please be as creative as possible*."*). The setting of the parameters was the same as in the Closed treatment.

The experimental protocol was identical across the three sessions, with one difference only: while in Session I subjects had to motivate the assigned creativity score by writing some free text, in Sessions II and III, on top of assigning a creativity score, they were asked to evaluate three specific aspects of each answer in a 1-10 scale: how unusual, well-elaborated, and nonsensical was the answer. These three items were selected relying on text analysis of free text motivations collected in Session I.

2.2. Experimental results

Table 1 provides an overview of the creative scores across the different conditions when AI's answers are generated using GPT-3.5 (Session I), GPT-4o (Session II) and GPT-4.5 (Session III). No matter the GPT model, in closed tasks, when considering each subject's average scores attributed to AI's and humans' answers, the average AI's score is significantly higher than those of humans. In the open task, the average human creativity score is always significantly higher than those of AI machines. In all cases, the difference-in-difference (1.963, 1.563 and 1.874 respectively) is highly significant, being between ten and eight times the standard errors. The distribution of human and AI creative scores is shown in Figure C1.

[8] The 40 answers taken from Charness and Grieco (2019) correspond to those from the Tournament treatment, where subjects chose this task over the alternative: "Imagine and describe a town, city, or society in the future." These answers are representative of the student pool in Charness and Grieco (2019), with average scores across the two sub-tasks not significantly differing (5.050 vs. 5.041, Z = -0.042, p = .966, two-sample Wilcoxon rank-sum test). Five answers were excluded due to illegible scans, but their exclusion did not affect the sample's representativeness (5.100 vs. 5.041, Z = -0.165, p = .869, two-sample Wilcoxon rank-sum test). The instructions used for these experiments are provided in Appendix B2.

Table 1: Creativity scores (humans' evaluation)

Session	Treatment	Type	Average	Std. Err.	Min	Max	Subj.
I	Closed	Human	5.331	0.161	2.025	8.950	128
I	Closed	AI	6.136	0.169	2.600	9.125	128
I	Open	Human	6.266	0.173	2.475	9.600	123
I	Open	AI	5.108	0.182	1.975	9.050	123
II	Closed	Human	5.936	0.156	2.600	9.175	119
II	Closed	AI	6.982	0.187	3.050	9.775	119
II	Open	Human	6.451	0.159	2.650	9.350	119
II	Open	AI	5.933	0.190	2.350	9.550	119
III	Closed	Human	6.021	0.165	2.550	9.000	122
III	Closed	AI	7.417	0.166	4.175	9.775	122
III	Open	Human	6.927	0.153	3.250	9.450	119
III	Open	AI	6.451	0.192	2.875	9.400	119

The table reports subjects' creative scores ranging from 1 to 10. Session can assume value equal to I (first session, run in July 2023 using GPT-3.5), II (second session, run in October 2024 using GPT-4o), or III (third session, run in April 2025 using GPT-4.5 and an explicit request to be "as creative as possible" in the prompt). Open refers to the open task, Closed to the closed task. AI refers to answers generated by AI, while Human refers to answers produced by human subjects. Min (Max) refers to the minimum (maximum) average score assigned to the answers of each treatment. Subj. refers to the number of subjects who have evaluated the answers in each treatment.

The better performance of AI in the closed task, and of humans in the open task, holds independently from the specific GPT model. However, the results show that the improved versions of GPT appear to have amplified AI's advantage in the closed task (the difference between AI and human scores rises from 0.805 to 1.046 to 1.396) and reduced the human advantage in the open task (the difference between human and AI scores reduces from 1.158 to 0.517 to 0.478)[9]. Further analysis presented below attempts to understand this result more in depth.

Table 2 reports a set of OLS regressions that confirm the results reported above, showing that AI exhibits a significantly lower performance in the open task (columns 1, 4, and 7). This result holds with subject fixed effects (columns 2, 5 and 8) and correcting for Multiple Hypotheses Testing[10] (columns 3, 6 and 9). In all regressions, we control for temperature, which has no significant effect on the creativity score.

[9] In Session 3, the prompt differs with the addition of the sentence "Be as creative as possible", which may partly explain GPT's improvement. Additionally, both human and GPT scores are higher, suggesting a possible subject-pool effect, with evaluators being generally more generous.

[10] Multiple Hypothesis Testing was performed using the *rwolf* command, which calculates Romano and Wolf's (2005a,b) stepdown adjusted p-values robust to multiple-hypothesis testing. This program follows the resampling algorithm described in Romano and Wolf (2016).

Table 2: Determinants of creativity scores

	Session I (GPT-3.5)			Session II (GPT-4o)			Session III (GPT-4.5)		
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	interactions	subject FE	MHT	interactions	subject FE	MHT	interactions	subject FE	MHT
AI	0.794***	0.814***	0.933***	1.046***	1.041***	0.806***	1.397***	1.396***	1.275***
	[0.241]	[0.205]	[0.282]	[0.246]	[0.206]	[0.303]	[0.240]	[0.189]	[0.363]
Open	0.915***	3.414**	2.659**	0.515**	-2.888	-6.224***	0.916***	2.510*	2.484***
	[0.244]	[1.440]	[1.147]	[0.247]	[5.719]	[2.071]	[0.243]	[1.281]	[0.499]
AI_open	-1.943***	-1.995***	-1.326***	-1.563***	-1.558***	-1.588***	-1.875***	-1.895***	-1.659***
	[0.344]	[0.294]	[0.476]	[0.348]	[0.292]	[0.417]	[0.342]	[0.269]	[0.436]
Subject Fixed Effect	no	yes	yes	no	yes	yes	no	yes	yes
Constant	5.189***	5.843*	5.976*	6.008***	-1.771	-6.260**	6.198***	6.615*	5.650***
	[0.293]	[3.448]	[3.442]	[0.306]	[5.663]	[3.166]	[0.294]	[3.610]	[0.436]
Observations	494	494	492	473	473	472	474	474	475
R-squared	0.064	0.664	0.810	0.051	0.667	0.838	0.076	0.718	0.828

OLS (errors clustered at the subject-response level in parentheses). The dependent variable is the average score assigned to creative answers, ranging from 1 to 10. AI is a dummy variable assuming value equal to 1 when the answer is produced by an AI GPT-3.5 (Session I) or GPT-4o (Session II), or GPT-4.5 machine (Session III), and 0 when produced by a human. Open is a dummy variable assuming value equal to 1 when the task is open, and 0 when the task is closed. AI_Open is the interacted variable between ai and Open. Temperature corresponds to the ChatGPT parameter temperature, ranging from 0 to 2.
*** significant at 1%; ** significant at 5%; * significant at 10%.

Table C1 in Appendix C shows the role of evaluators' demographic features in determining creativity scores, but no relevant results or consistent patterns were found across sessions.

2.3. Text analysis and decomposition of the creativity score

To better understand why AI is ineffective in the open task, we ran a textual analysis of the content of subjects' written motivations that accompanied the scores assigned in Session I. We conducted a topic modelling analysis using Latent Dirichlet Allocation procedure[11] on the full text of motivations to identify the major arguments the respondents brought. We ran the procedure multiple times, pre-setting different numbers of topics. We found that assuming three or more topics always resulted in overlapping sets of characterizing words, making it difficult to infer an underlying argument. The main keywords in each are different enough to identify two themes for each task, that

[11] Latent Dirichlet Allocation is one of the most popular machine-learning topic models, developed by Blei et al. (2003). It allows for the automatic clustering of any kind of text documents into a user-chosen number of clusters of similar content, usually referred to as "topics", representing each document as a probability distribution over topics and each topic as a probability distribution over words (Schwarz, 2018).

we summarize as related to the answer being "original" or "nonsensical" in the open task, and "original" or "well-elaborated" in the closed task. A summary statistics illustrating the five highest-loading words in each topic is provided in Table C2 in Appendix C. Figure C2 shows that humans produce answers that are judged as "not original enough" more frequently in the closed task, possibly suffering the burden of constraints, while outperform AI in originality in the open task; AI always produces more well-elaborated and less nonsensical answers.

Table 3: Creativity items scores (humans' evaluation)

Session	Treatment	Type	Originality	Elaboration	Nonsense
II	Closed	Human	6.114 (.180)	8.896 (.013)	5.837 (.191)
II	Closed	AI	6.453 (.177)	9.587 (.014)	5.159 (.212)
II	Open	Human	5.946 (.176)	9.120 (.013)	4.989 (.199)
II	Open	AI	4.751 (.202)	9.264 (.016)	3.731 (.186)
III	Closed	Human	5.839 (.175)	8.765 (.014)	5.038 (.196)
III	Closed	AI	6.600 (.192)	9.589 (.012)	4.788 (.212)
III	Open	Human	6.020 (.186)	9.271 (.011)	4.083 (.011)
III	Open	AI	4.970 (.194)	9.155 (.014)	3.188 (.178)

The table reports subjects' average scores for the three items (originality, elaboration and nonsense) in Session II and III. Scores range from 1 to 10; standard errors are in parenthesis. Open refers to the open task, Closed to the closed task. AI refers to answers generated by AI, while Human refers to answers produced by human subjects. These data refer to Session II, run in October 2024 using GPT-4o, and Session III, run in April 2025 using GPT-4.5.

Table 3 shows that, no matter the session, the frequency of nonsensical answers produced by AI is lower than that for humans in both tasks, except for the closed task in session III. This confirms that the efforts to reduce hallucination claimed in OpenAI's statement releases of GPT-4o and 4.5 have been successful. The level of elaboration tends to be higher for AI than for humans no matter the type of task. Nonetheless, humans still outperform AI in terms of originality in the open task. This result is in line with Girotra et al. (2023), where AI-generated new product ideas are perceived as less novel than those produced by students, indicating that AI relies on a "less diverse solution landscape" (p.1). In a similar flavour, but in designs where humans can rely on AI, Boussioux et al. (2024) find that human crowds are able to produce business solutions with a higher degree of novelty than human-AI solutions. Doshi and Hauser (2024) show that generative AI–enabled stories are more similar to each other than stories by humans alone, and Zhou and Lee (2023) show that AI-assisted artists produce artworks that their peers evaluate more favourably. On the contrary, our results are in contrast with Bohren et al. (2024): although using the same open task we use, in their experiment humans exhibit a worse performance than ChatGPT, with this result probably depending on the different sample of human subjects (Prolific subjects versus standard lab subjects, with only top 10% receiving a bonus

and a flat payment for the others[12]). In sum, this ability to generate novelty, that is a product of human *imagination*, seems to be the essence of the "human advantage" in open tasks, and its role will be quantified through the structural estimation in Section 3.

3. Structural model and estimation

To further understand the reduced-form results, we present a simple model capturing the utility function of an agent (human or machine) in a creative task with a certain degree of perceived openness $\varphi \geq 1$, with $\varphi = 1$ in case of fully-closed tasks, and the perceived openness increasing in φ. The agent choses the optimal creative output y in order to maximize the following utility:

$$U(y) = -\frac{\left(\frac{\bar{y}}{\varphi} - y + t\right)^2}{2} + \beta y \qquad (1)$$

where $\bar{y}$ is the output level corresponding to the "algorithmic" solution to the task, i.e., the solution that can be obtained by exactly following the instructions. In general, this function suggests that the agent has a quadratic disutility when the output y departs from the algorithmic solution of the task $\bar{y}$. Agents choose y, while $\bar{y}$ is a parameter that refers to the output that would be generated if the agent follows the instructions perfectly. The more open the task is, the smaller the role of instructions. For machines, a higher temperature t amplifies this distance from instructions, since a higher temperature t increases the variability in the output y. We assume $t = 0$ for humans but present a robustness check where we use self-reported creative style to "proxy" temperature and capture human heterogeneity in approaching creative tasks. In a fully-closed task, one could get the maximum score by following the instructions; with fully-open tasks ($\varphi \to \infty$), y* = *t*, so that the creative performance is determined by this parameter.

While AI cannot *imagine* anything that goes beyond the instructions, humans can. So, for humans we include a second determinant of creative output β, which captures the role of human imagination. Individuals derive pleasure from bringing imaginative ideas to life: this produces a multiplicative effect that increases their creative output[13]. For humans, the creative output chosen is thus a weighted average between the output produced by simply following the instruction, and the output that results from imagination.

[12] Evaluators in our experiment assign lower scores to GPT than those assigned in Bohren et al. (2024). We argue that this might be due to the comparatively better performance of humans, since "people are spontaneously inclined […] to evaluate others and themselves in a comparative manner" (Goffin and Olson, 1997, p. 48) also when assigning absolute scores.

[13] This assumption reflects evidence on intrinsic motivation to engage in creative endeavors (Amabile, 1989; 1996).

The F.O.C. reads as follows:

$$y^* = \frac{\bar{y}}{\varphi} + t + \beta \qquad (2)$$

and suggests that, given the output level associated with the algorithmic solution $\bar{y}$, the agents' utility may differ according to three drivers. First, the attitude towards the openness of the task φ; second, how manipulating the temperature t affects the creative score for machines; third, the relative weight of human imagination β with respect to instructions.

Identification of the above parameters $(\bar{y}, \varphi, t, \beta)$ is achieved by exploiting the features and the variation from the experimental design. We can define the following moment conditions:

$$\mathrm{E}(y_{AI-C}) = \frac{\bar{y}_C}{\varphi_C} + t$$

$$\mathrm{E}(y_{AI-O}) = \frac{\bar{y}_O}{\varphi_O} + t$$

$$\mathrm{E}(y_{H-C}) = \frac{\bar{y}_C}{\varphi_C} + \beta_C$$

$$\mathrm{E}(y_{H-O}) = \frac{\bar{y}_O}{\varphi_O} + \beta_O$$

The system is overidentified and estimation of the model parameters is achieved via GMM. This implies minimizing the weighted distance between the empirical value of the creative score in each condition and that predicted by the model.

The sources of identification for the algorithmic solutions in the closed task ($\bar{y}_C$) and in the open task ($\bar{y}_O$) are $\mathrm{E}(y_{AI-C})$ and $\mathrm{E}(y_{AI-O})$ with $t = 0$. We proxy those values with the average score attributed to deterministic answers, out of a 1-10 scale. The perceived openness of the two tasks (φ_C and φ_O) is identified by comparing the two treatments, Open and Closed. The identification of the weight attached to human imagination relative to instructions β relies on the difference in the scores between humans and machines. Finally, identification of t relies on the variation of temperature across AI-generated answers.

Our baseline estimation strategy uses the diagonal of the inverse variance-covariance matrix as a weighting matrix: the use of an identity matrix gives identical results. Both approaches overcome the issue arising from the poor small sample properties of the optimal weighting matrix (see Altonji and Segal, 1996). Table 4 reports the estimated preference parameters for the model using data from the three sessions. Columns 2, 4 and 6 present the estimation results for the restricted model in which we set $t > 1$ (i.e., restricting to the case of high randomness in the AI-generated answers). No matter

the session, all simulated moments fall within the 95% confidence interval of the empirical moments, except for β_C. This indicates that the theoretical model does a remarkably good job in replicating AI's and humans' performance in the experimental data.

Table 4: Estimated preference parameters

	Session I (GPT-3.5)		Session II (GPT-4o)		Session III (GPT-4.5)	
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
φ_C	1.958***	2.020***	1.024***	1.103***	0.953***	1.018***
	(0.064)	(0.102)	(0.064)	(0.056)	(0.026)	(0.044)
φ_O	2.449***	2.922***	1.326***	1.424***	1.201****	1.297***
	(0.114)	(0.245)	(0.053)	(0.084)	(0.043)	(0.067)
β_C	0.225	0.443	-0.023	0.540	-0.381	0.034
	(0.232)	(0.338)	(0.246)	(0.360)	(0.024)	(0.034)
β_O	2.186***	2.841***	1.549***	1.870***	1.516***	1.705***
	(0.256)	(0.376)	(0.251)	(0.352)	(0.246)	(0.336)
Observations	496	231	474	242	477	230
GMM Crit.	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001

The Table reports the set of preference parameters estimated using GMM and AI answers generated by GPT-3.5 (columns 1-2), GPT-4o (columns 3-4) and GPT-4.5 (columns 5-6). Columns (1), (3) and (5) report the estimation results for the full model, while columns (2), (4) and (6) report the estimation results for the restricted model (setting t > 1), respectively. Standard errors in parentheses.
*** significant at 1%; ** significant at 5%; * significant at 10%.

The estimation confirms that φ increases with the openness of the task, with φ_O= 2.449> φ_C = 1.958 in Session I, φ_O= 1.326> φ_C = 1.024 in Session II, and φ_O= 1.200> φ_C = 1.018 in Session III. Human imagination is significantly relevant only in open tasks, not in closed tasks (β_C is not significantly different from 0). This means that, for sufficiently closed tasks, following the instructions is enough, and no further boost from imagination is needed. In the open task, the relative role of imagination with respect to instructions accounts for about 35% of humans' average creative score in Session I, 24% in Session II and 22% in Session III. This relevance increases when restricting the analysis to AI answers generated with high temperature (as these are, by design, more peculiar), and comparing them with human performance (β_O = 2.841 > 2.186 in Session I, β_O = 1.870 > 1.549 in Session II, and β_O = 1.706 > 1.515 in Session III): when AI's randomness increases, human imagination accounts for the 45% (Session I), 30% (Session II) and 25% (Session III) of the creative score. Increasing randomness in AI generated answers determines an increase in the diversity of answers, but for diversity to translate in creativity and not just oddity, (human) imagination is needed, especially with GPT-3.5. In GPT-4o and GPT-4.5 this trade-off appears less severe.

Table C3 presents a robustness check that accounts for human heterogeneity in approaching creative tasks. Humans may differ in their creative styles (Galenson, 2004; 2010): individuals exhibiting an experimental creative style perceive task completion as a process of exploration, seeking to make discoveries during the course of their work. In contrast, individuals with a cognitive creative style tend to plan their work and execute it systematically. We assume that subjects who self-report to have an "experimental" creative style are intrinsically more likely to depart from instructions (Nielsen et al., 2008; Charness and Grieco, 2019) and thus exhibit a higher "temperature"[14]. We repeated the regressions from Table 4, adding self-reported creative style as a proxy for "temperature" in human subjects. We find that controlling for creative style reduces the significance of imagination in the open task. This suggests that creative style accounts for some of the impact that imagination initially appeared to have on creative performance. Furthermore, imagination turns out to be detrimental in the closed task, probably due to the presence of constraints that limit its expression.

We now focus on differences across GPT versions. Although we find that human imagination accounts for a significant portion of the creative score in open tasks, we observe a decline in the human advantage—from 35% to 22% over time—as AI models advance from GPT-3.5 to GPT-4.5. To gain a deeper understanding of how AI is improving, we now aim to "unpack" the evaluation of creativity by analyzing the scores participants assigned to the three elements that contribute to the overall creativity score: originality, elaboration, and nonsense.

The model is thus enriched by adding the following equation:

$$y = x + \alpha v + \gamma z + \varepsilon \qquad (3)$$

where the creativity score y is assumed to be the weighted sum[15] of originality x, elaboration v and nonsense z, with α and γ being the relative weight of elaboration and nonsense with respect to originality, plus an error term ε including other unobserved factors which might still also contribute to the creativity score.

These estimations (columns 1 and 3) allow quantifying how much of human advantage, relative to AI, is due to each item as compared to the others. We observe that, in closed tasks, elaboration counts for between 40-51% (depending on the GPT version) if compared to originality, while nonsense weights for 63-85% in reducing the overall creative score. In open tasks, the relative weight of originality is

[14] Questions 1–7 in Section 3 of the final questionnaire (Appendix B3) are used to construct the indicator of experimental creative style.

[15] The prevailing way to form scores from multiple-item scales in psychological research is to sum or average responses of all items (McNeish and Wolf, 2020). Our approach refers to "congeneric models" where every item is differentially related to the construct of interest by weighting its loading (Graham, 2006).

higher, with elaboration that counts between 33% and 40% and nonsense for 50-67%. When temperature rises (columns 2 and 4), in both tasks the role of elaboration relatively increases for GPT-4o; the same holds for nonsense. However, these two items lose significance with GPT-4.5. In sum, the human advantage appears to reside mainly in the ability to generate new, unusual ideas. Producing an answer which is well-elaborated matters to be evaluated as creative, but not as much as producing an original idea.

Table 5: Creativity score items: Estimated parameters

	Session II (GPT-4o)		Session III (GPT-4.5)	
VARIABLES	(1)	(2)	(3)	(4)
α_C	0.397***	0.439***	0.511**	0.538**
	(0.139)	(0.140)	(0.216)	(0.250)
γ_C	-0.635***	-0.735***	-0.853**	-0.895*
	(0.232)	(0.235)	(0.398)	(0.465)
α_O	0.330***	0.429***	0.389***	0.356*
	(0.094)	(0.109)	(0.106)	(0.297)
γ_O	-0.501***	-0.720***	-0.653**	-0.525
	(0.192)	(0.240)	(0.266)	(0.496)
Observations	474	242	477	247
GMM Crit.	0.000001	0.000001	0.000001	0.000001

The Table reports the set of preference parameters estimated using GMM and AI answers generated by GPT-4o. Column (1) reports the estimation results for the full model using an identity matrix (the diagonal of the inverse variance-covariance matrix as weighting matrix gives identical results), while column (3) reports the estimation results for the restricted model (setting t > 1), respectively. Standard errors in parentheses.
*** significant at 1%; ** significant at 5%; * significant at 10%.

Limiting nonsense matters too, but again to a lower extent than novelty. If human imagination accounts for 35–22% of the superior creative performance observed in human subjects, this advantage is primarily attributable to originality. In the latest version of GPT, the weight assigned to originality is approximately twice that of elaboration, and about 1.5 times greater than that assigned to nonsense.

While AI demonstrates improvements across two dimensions relevant to the formal presentation of creative output, these gains do not extend to the domain of creative content, i.e., originality.

4. AI's evaluation of creativity

A further question concerns AI's capability of *evaluating* creative output. AI is already used for selection practices such as screening applicants' resumes, costumers' credit worthiness, pricing of goods (Yarger et al., 2020). The question is whether AI is really able to evaluate the performance in a task, especially if AI itself is not excelling in that task, as happens in the case of our open endeavors.

To address this issue, we replicate the experiment by asking GPT to evaluate the same three tanks of 160 answers previously evaluated by human subjects, see Table C4 in Appendix C. With GPT-3.5 and GPT-4o, the creative scores assigned by AI to humans and AI in the two tasks do not differ significantly between each other, while GPT-4.5 assigns higher scores to AI in both tasks. It is interesting to note that AI scores tend to be more generous than humans' ones in GPT-3.5 and 4o sessions (but not in 4.5), and that the standard errors in AI's evaluations are always considerably lower than the one of humans (see Table 1 for a comparison). Tables C5 and C6 show there is no systematic improvement in the precision of AI assessments across sessions, with the exception of AI's ability to evaluate human answers.

As we did with human evaluators, we asked GPT-4o and 4.5 to evaluate the same three specific items of each creative answer as above, assigning a score in a 1-10 range about how much the answer was (1) original, (2) elaborated, and (3) nonsensical. When asked to provide more specific evaluations, GPT-4o and 4.5 assign scores that are more similar to humans' ones, showing that the ability in replicating humans' evaluation processes improves when the prompts are more precise (see Table C7). Finally, according to GPT, humans outperform AI in terms of originality in the open task. In summary, GPT concurs that the ability to generate novelty appears to be the essence of the "human advantage."

5. Discussion and conclusions

Artificial Intelligence advancements are occurring so fast that workers expect their displacement to be massive and long-lasting (Acemoglu and Restrepo, 2018). However, at the moment of writing it is far too early to have a clear sense of the impact of the last generation of AI's developments on workers, productivity and tasks. Our results focus on creative endeavours and show that humans (still) outperform AI in sufficiently open creative tasks, namely tasks where subjects are required to find, invent, or discover new solutions.

We present the results of three sessions of experiments conducted using three different versions of GPT and show that, also with the more advanced one, humans still outperform AI in highly creative tasks thanks to their ability to produce novel, unusual ideas. Our findings are even more striking since the experimental subjects are standard laboratory subjects (students) and were not made aware in advance of the topic of the experiment, so there was no self-selection of highly creative subjects, as might happen in case of real-life creative endeavors.

Our structural model illustrates the mechanism of creative production in AI versus humans: the creative output can be modelled as the weighted average between the output produced by simply following the instruction, and human imagination. The ability of human agents to "go beyond" what

is prescribed by instructions is estimated to account for a percentage between 22% and 45% of humans' creative score in the open task, while has no relevance in the closed task. Human imagination resulted to be more relevant when GPT-3.5 was prompted to increase the level of randomness of their associations, since it appeared not to be able to distinguish between creativity and randomness and, as such, incapable to detect nonsense output. GPT-4.o and 4.5 have improved in this respect but still lack the capability to produce unusual ideas. If we cannot expect AI to learn to be as creative as humans in the next future, training humans in understanding how AI works and reacts to prompts could possibly improve AI's (creative) output.

This finding offers at least a ray of hope for creative individuals who fear obsolescence at the hands of artificial intelligence. Although AI is improving in its ability to resemble humans, generate more elaborate answers, avoid hallucinations, and provide sound evaluations of creativity when prompted with sufficient precision, the way AI models are built—namely, by relying on existing information—contrasts with the need to depart from what is common and create something truly "unique," which is a key requirement of creative endeavours.

Our experiment exploits three GPT versions, released in November 2022, May 2024, and February 2025, respectively, thus providing three points of observation along the AI trajectory of technology progress. According to our estimations, from the former version to the most recent, AI has succeeded in reducing the gap between its capabilities and human ones of 11% (from 35% to 24%) in 18 months, but only of 2% (from 24% to 22%) in the last nine months. AI progresses are showing diminishing returns"[16]: AI models currently appear to be reaching a ceiling in their capabilities, indicating that using more data and computational power when pre-training them is not enough to turn machines into "all-knowing digital gods"[17]. In addition, when disentangling the components of the creativity score, it becomes clear that AI improvement consists of creative answers that have become better elaborated and less nonsensical, but that *are not* more novel and original.

If a substitution between AI and humans is unlikely, our result points more in the direction of synergies, providing evidence in favor of the tremendous possibilities of successfully exploiting the complementarities between humans' imagination, and AI's accuracy and computational capabilities. Furthermore, it stresses the importance to invest in "creative capital" and incentivize creative thinking, in line with what recently found by Albanesi et al. (2023) on 16 European Countries: on average employment shares have increased in jobs more exposed to AI, particularly in the case of occupations with a relatively higher proportion of younger and skilled workers. Similarly, Acemoglu

[16] https://techcrunch.com/2024/11/20/ai-scaling-laws-are-showing-diminishing-returns-forcing-ai-labs-to-change-course/
[17] *Ibidem.*

et al. (2023) show that US firms adopting AI are those (already) growing faster. However, how to successfully integrate human and AI is a critical and fascinating question. To answer this question, considerably more research is needed. We invite others to join us in this quest.

References

Acemoglu, D., Anderson, G., Beede, D., Buffington, C., Childress, E., Dinlersoz, E., ... Zolas, N. (2023). 'Advanced technology adoption: Selection or causal effects?', *AEA Papers and Proceedings*, vol. 113, pp. 210-14.

Acemoglu, D., Restrepo, P. (2018). 'Artificial intelligence, automation, and work', *The Economics of Artificial Intelligence: An Agenda*, pp. 197-236, Chicago: University of Chicago Press.

Albanesi, S., Dias da Silva, A., Jimeno, J.F., Lamo A., Wabitsch, A. (2023). 'New technologies and jobs in Europe', *CEPR Discussion Paper,* No. 18220.

Altonji, J. G., Segal, L. M. (1996). 'Small-sample bias in GMM estimation of covariance structures', *Journal of Business & Economic Statistics*, vol. 14(3), pp. 353-366.

Amabile, T. M., Goldfarb, P., Brackfleld, S. C. (1990). 'Social influences on creativity: Evaluation, coaction, and surveillance', *Creativity Research Journal*, vol. 3(1), pp. 6-21.

Biancotti, C., Camassa, C. (2023). 'Loquacity and visible emotion: ChatGPT as a policy advisor', *Mimeo*, Bank of Italy.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). 'Latent dirichlet allocation', *Journal of Machine Learning Research*, vol. 3 (Jan.), pp. 993-1022.

Boden, M.A. (1998). 'Creativity and artificial intelligence', *Artificial Intelligence*, vol. 103, pp. 347-356.

Bohren, N., Hakimov, R., Lalive, R. (2024). 'Creative and Strategic Capabilities of Generative AI: Evidence from Large-Scale Experiments', *IZA Discussion Papers*, No. 17302.

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., Lakhani, K. R. (2024). 'The crowdless future? Generative AI and creative problem-solving', *Organization Science*, vol. 35(5), pp. 1589-1607.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Zhang, Y. (2023). 'Sparks of artificial general intelligence: Early experiments with gpt-4', *arXiv preprint* arXiv:2303, No.12712.

Charness, G., Grieco, D. (2019). 'Creativity and incentives', *Journal of the European Economic Association*, vol. 17(2), pp. 454-496.

Charness, G., Grieco, D. (2022). 'Creativity and ambiguity tolerance', *Economics Letters*, vol. 218(110720), pp.1-3.

Charness, G., Grieco, D. (2023). 'Creativity and corporate culture', *The Economic Journal*, vol. 133(653), pp. 1846-1870.

Collins, M. A., Amabile, T. M. (1999). 'Motivation and creativity', in R. J. Sternberg (Ed.), *Handbook of Creativity*, pp. 297-312, New York: Cambridge University Press.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., Chen, G. (2023). 'Can large language models provide feedback to students? A case study on ChatGPT', in 2023 *IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 323-325, IEEE.

Doshi, A. R., Hauser, O. P. (2024). 'Generative AI enhances individual creativity but reduces the collective diversity of novel content', *Science Advances*, vol. 10(28), pp. 1-9.

Fu, Y., Bin, H., Zhou, T., Wang, M., Chen, Y., Lai, Z. G. D. C., ... Hiniker, A. (2024). 'Creativity in the age of AI: Evaluating the impact of Generative AI on design outputs and designers' creative thinking', *arXiv preprint arXiv*:2411.00168.

Galenson, D. (2004). 'A portrait of the artist as a very young or very old innovator: Creativity at the extremes of the life cycle', *NBER* 10515.

Galenson, D. W. (2010). 'Understanding creativity', Journal of Applied Economics, vol. 13(2), pp. 351-362.

Gilardi, F., Alizadeh, M., Kubli, M. (2023). 'Chatgpt outperforms crowd-workers for text-annotation tasks', *arXiv preprint arXiv:*2303.15056.

Girotra, K., Meincke, L., Terwiesch, C., Ulrich, K. T. (2023). 'Ideas are dimes a dozen: Large language models for idea generation in innovation', *SSRN* 4526071.

Goffin, R. D., Olson, J. M. (2011). 'Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others', *Perspectives on Psychological Science*, vol. 6(1), pp. 48-60.

Jennings, K. E. (2010). 'Developing creativity: Artificial barriers in artificial intelligence', *Minds and Machines*, vol. 20, pp. 489-501.

Joosten, J., Bilgram, V., Hahn, A., Totzek, D. (2024). 'Comparing the ideation quality of humans with generative artificial intelligence', *IEEE Engineering Management Review*, vol. 52(2), pp. 153-164.

Koivisto, M., Grassini, S. (2023). 'Best humans still outperform artificial intelligence in a creative divergent thinking task', *Scientific Reports*, vol. 13(1), pp. 13601.

Magni, F., Park, J., Chao, M. M. (2024). 'Humans as creativity gatekeepers: Are we biased against AI creativity?', *Journal of Business and Psychology*, vol. 39(3), pp. 643-656.

McNeish, D., Wolf, M. G. (2020). 'Thinking twice about sum scores', *Behavior Research Methods*, vol. 52, pp. 2287-2305.

Mumford, M.D. (2003). 'Where have we been, where are we going? Taking stock in creativity research', *Creativity Research Journal*, vol. 15, pp. 107–120.

Nielsen, B. D., Pickett, C. L., Simonton, D. K. (2008). 'Conceptual versus experimental creativity: Which works best on convergent and divergent thinking tasks?', *Psychology of Aesthetics, Creativity, and the Arts*, vol. 2(3), pp. 131.

OECD (2023). 'AI language models: Technological, socio-economic and policy considerations', *OECD Digital Economy Papers,* No. 352, Paris: OECD Publishing.

Pievani, T. (2026). *MULTIPLICITY. An Adventure in the Great Library of Evolution,* Cambridge (MA): THE MIT PRESS.

Romano, J. P., Wolf, M. (2005a). 'Exact and approximate stepdown methods for multiple hypothesis testing', *Journal of the American Statistical Association,* vol. 100(469), pp. 94-108.

Romano, J. P., Wolf, M. (2005b). 'Stepwise multiple testing as formalized data snooping', *Econometrica,* vol. 73(4), pp. 1237-1282.

Romano, J. P., Wolf, M. (2016). 'Efficient computation of adjusted p-values for resampling-based stepdown multiple testing', *Statistics & Probability Letters*, vol. 113, pp. 38-40.

Schwarz, C. (2018). 'ldagibbs: A command for topic modeling in Stata using latent Dirichlet allocation', *The Stata Journal*, vol. 18(1), pp. 101-117.

Tattersall, I. (2001). 'How we came to be human', Scientific American, vol. 285(6), pp. 56-63.

Unsworth, K. (2001). 'Unpacking creativity', *Academy of Management Review*, vol. 26(2), pp. 289-297.

Yarger, L., Cobb Payton, F., Neupane, B. (2020). 'Algorithmic equity in the hiring of underrepresented IT job candidates', *Online Information Review*, vol. 44, pp. 383–395.

Zhou, E., & Lee, D. (2024). 'Generative artificial intelligence, human creativity, and art', *PNAS Nexus*, vol. 3(3), pp. 1-8.