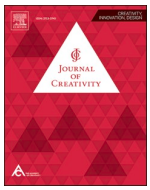
Journal of Creativity 33 (2023) 100065

Contents lists available at ScienceDirect

Journal of Creativity

journal homepage: www.elsevier.com/locate/yjoc



# The originality of machines: AI takes the Torrance Test

Erik E. Guzik [a,*], Christian Byrge [b], Christian Gilde [c]

[a] University of Montana, College of Business, 32 Campus Drive, Missoula, MT 59812, United States
[b] Vilnius University Business School, Lithuania
[c] University of Montana Western, Business Department, United States

ARTICLE INFO

*Keywords:*
Artificial intelligence
Originality
Creativity
Assessment
Torrance tests of creative thinking
Entrepreneurship
Innovation

ABSTRACT

This exploratory research investigated the creative abilities of OpenAI's large language model, ChatGPT, based on the GPT-4 architecture, as assessed by the Torrance Tests of Creative Thinking. In comparison to human samples and a national percentile from Scholastic Testing Services, ChatGPT's performance was analyzed for fluency, flexibility, and originality. Results indicated that ChatGPT scored within the top 1% for originality and fluency, and showed high scores for flexibility, thus highlighting the current creative abilities of AI and the potential of AI systems to support and augment human creativity in new and meaningful ways. The study encourages additional research to further define, measure, and develop creativity in the era of advanced AI.

## Introduction

The emergence of artificial intelligence (AI) and the ongoing development of its capabilities have opened new doors in the assessment of skills that were previously believed to be the sole province of human cognition. GPT-4, developed by OpenAI, is one such AI model, known for its remarkable performance in generating human-like responses to natural language queries (OpenAI, 2023). Demonstrating proficiency across several academic fields, GPT-4 has recently achieved exceptional scores on assessments ranging from the Law School Admission Test (LSAT) to the Graduate Record Examination, Verbal Exam (OpenAI, 2023).

These recent achievements in AI suggest interesting questions for those researching forms of cognition often described as creative thinking: How do advanced large language models (LLM) perform on creative tasks? More pointedly, how do LLM models like ChatGPT perform on accepted creativity assessments designed to measure and identify such human creative abilities as novelty and original thinking?

That AI might be capable of exhibiting creative abilities is perhaps not as surprising—or as far-fetched—as one might think. The use of AI for solving creative tasks is not new (Boden, 2004; Cope, 1989). Indeed, the very origins of AI were based in no small measure on a desire to develop new and novel ways to solve problems (Cordeschi, 2007). In addition, with the recent introduction of public-facing AI tools there has been an increase in the use of AI to generate a variety of creative work, including written content, images, video, and sound formats (Miller, 2019; Anantrasirichai & Bull, 2022).

Further, though not widely recognized or discussed, one of the explicit goals of the founders of AI in their initial proposal for the famous 1956 Dartmouth Summer Research Project on Artificial Intelligence was the development of machines to simulate all aspects of human intelligence (McCarthy et al., 1955), with specific focus on "Randomness and Creativity" (McCarthy et al., 1955, p. 2). In this same proposal, Nathaniel Rochester shared his desire to develop means to promote "Originality in Machine Performance," as well as methods to support "The Process of Invention or Discovery," stating as his seminal goal, "how can I make a machine which will exhibit originality in its solution of problems?" (McCarthy et al., 1955, pp. 7–9). Apparently, AI's founders believed creativity—including originality—were among the specific forms of intelligence that machines could emulate (Boden, 2009).

Have, then, the stated goals of AI's founders been realized? That is, is AI creative? Further, does AI exhibit novelty and originality in its solutions? This research sought to answer these questions by assessing and exploring the creative abilities of the GPT-4 AI model.

## Material and methods

To answer questions of AI's ability to exhibit creativity and originality, AI must be evaluated, just as any human would need to be evaluated to ascertain creative ability. *How* to test AI, however, is a much thornier question, especially since assessment depends in no small manner on how one defines creativity.

* Corresponding author.
*E-mail address:* erik.guzik@umontana.edu (E.E. Guzik).

https://doi.org/10.1016/j.yjoc.2023.100065
Received 14 July 2023; Received in revised form 15 August 2023; Accepted 20 August 2023
Available online 22 August 2023

In this respect, there are a variety of definitions and means of assessing creativity, including notions and measures based on personality, achievement, actions, mindset, creative behaviors, and so on (Sternberg, 1999). Definitions and assessment methods can likewise be understood in terms of frameworks targeting such aspects of creativity as Person, Product, Process, and Press (4Ps, Rhodes, 1961), or such categories as Creators, Creating, Collaborations, Contexts, Creations, Consumption, and Curricula (7Cs, Lubart, 2017).

For its evaluation method, this research focused specifically on the assessment of product (output, artifacts, or creations). While the assessment of other factors of creativity—including creative achievement, confidence, self-belief, behavior, traits, methods, etc.—are largely based on self-reporting questionnaires, the assessment of creative products is more often based on the social (external) evaluation of creative ideas and creative responses that have been produced by the test-taker. Assessment of creative products seems especially appropriate for evaluating the creative abilities of AI. Specifically: Is AI creative according to the external evaluation of its actual output?

For the purposes of this research, creative products may be defined as novel and useful (Runco & Jaeger, 2012). For some researchers, novelty has often been considered more important than usefulness for defining and identifying creativity (Caroff & Besançon, 2008; Diedrich, Benedek, Jauk, & Neubauer, 2015; Han, Forbes, & Schaefer, 2021; Runco & Charles, 1993), which seemingly supports Rochester's vision of developing machine originality as a requisite component of AI-based problem-solving abilities (McCarthy, 1955).

Some of the more established creative product assessment methods include the Torrance Tests of Creative Thinking (TTCT; Lissitz and Willhoft 1985), Runco Creativity Assessment Battery (rCAB; Runco, 2011), Consensual Assessment Technique (CAT; Amabile, 1982), Remote Associates Test (RAT; Mednick & Mednick, 1967), and Alternative Uses Test (AUT; J.P. Guilford, 1967). Dietrich and Kanso (2010) suggests that different types of assessment methods for creative products may have fundamentally different uses: the TTCT and the AUT may be primarily useful for assessing divergent thinking, while the CAT may be primarily useful for assessing artistic and real-life creativity tasks, and the RAT may be primarily useful for assessing insight and forms of convergent thinking.

In terms of evaluating AI output, these methods each have their own methodological strengths. The strength of the TTCT is its large database of historical human responses that can be used as a control and comparison group. This has become possible because of its consistency in using the same demographic makeup of test-tasker for decades and because responses have been systematically collected for scoring. The strength of the AUT and the RAT is the simplicity in administrating the tests, the scoring and analysis. The strength of the CAT is its flexibility to adopt to specific domains. This becomes possible because it uses a panel of domain related experts to judge the creative products (Kaufman, Plucker, & Baer, 2008).

More importantly, these methods may assess different levels of creativity, which may be particularly relevant for the assessment of AI artifacts. Many researchers working in the fields of AI and computer science distinguish between psychological creativity (P-creativity) and historical creativity (H-creativity), suggesting that P-creativity might be a more relevant form of creativity to examine when analyzing machine output (Boden, 2004; Miller, 2019). Interestingly, in the field of psychology, Sternberg (2018) suggests that those who are strong at everyday (or professional creativity) may not necessarily be strong at historical creativity. Further, Kaufman et al. (2010) find that for everyday "little-c" creativity, domain specificity is rather low, while for historical "Big-C" creativity, domain specificity is high. The CAT may therefore be more applicable for assessing creative products related to historical or "Big-C" creativity, while the AUT and TTCT may be more applicable for assessing psychological, everyday and professional creativity.

This study sought to assess the creative potential of ChatGPT related to psychological, everyday and professional creativity. This made the TTCT interesting as a viable and appropriate assessment tool for this study. The TTCT offers a suite of authentic activities that prompt the test-taker to engage in various types of thinking that mirror the kinds of creativity required for real-life and daily human operations, including asking questions, guessing causes and consequences, improving a product, and utilizing imagination (STS, 2017). In total, the TTCT includes six distinct creative activities designed to evaluate the operation of creativity as it is often used in business and everyday life. In addition to a standard Alternative Uses Task (Unusual Uses), the TTCT also includes tasks to evaluate Asking Questions, Guessing Causes, Guessing Consequences, and Product Improvement. Each separate activity holds promise in providing new insight into the possible creativity and originality of AI.

In addition, the TTCT does not relate to any specific domain or profession. Rather it gives insight into domain-general creativity that is recognizable to humans across all kinds of domains and professions. As such, the TTCT functions like a test battery offering a broader system of measures that can assess more perspectives of the multi-dimensional nature of creativity.

Also, the TTCT is a commercially protected assessment instrument, therefore prompts are not accessible to ChatGPT or any other AI system, or publicly available in general. This offers a unique opportunity to test ChatGPT using specific prompts that it likely has never been asked before. ChatGPT will have to generate the responses, not simply retrieve them from its database.

Furthermore, due to the historical collection and assessment by Scholastic Testing Services (STS), the TTCT offers the possibility to compare the ChatGPT responses to thousands of human responses. This makes it feasible to measure originality and novelty as uncommonness by using statistical rarity of responses. In addition, STS scorers only accept responses that are relevant (useful) for the tasks provided, thus further satisfying the definition of creativity (novelty and usefulness) offered by Runco & Jaeger (2012).

And finally, since the TTCT has been one of the most widely used and referenced tests of creativity (Davis, 1997; Lissitz & Willhoft, 1985), and because it has a clear definition of creativity and of its creativity variables (e.g., fluency, flexibility, and originality), it is easier to relate the TTCT results to previous studies where the TTCT or related methods have been used.

For these reasons, the TTCT was selected as the most viable instrument for this study.

### *Sample*

The basis of this research was a controlled study with additional comparison to a normed database of human results provided by STS. The experimental group consisted of 8 separate GPT-4 submissions generated through the OpenAI ChatGPT application. The control group was composed of 24 human submissions collected from 11 male and 13 female undergraduate students, with an age range from 20 to 31. A comparison group was taken from the national data provided by STS compiled in 2016, comprising 2718 students in grades 13-plus (STS, 2017). While the TTCT is often used for identification of gifted students in grades 1–12, the comparison group and percentile rankings included only data and results from STS for students in grades 13-plus.

### *Materials*

This research utilized the TTCT Verbal Test as an assessment instrument. The TTCT Verbal Test includes six distinct creative thinking activities: (1) Asking Questions; (2) Guessing Causes; (3) Guessing Consequences; (4) Product Improvement; (5) Unusual Uses; and (6) Just Suppose. Each activity within the TTCT includes a standard set of directions ranging from 4–12 sentences, recited verbally to test-takers prior to each timed activity. While the TTCT contains protected

intellectual property that cannot be shared publicly, the six activities were generally structured as follows:

- **Activity 1: Asking Questions.** The test-taker is tasked with asking questions about a given picture. The picture displays an action involving one or more living characters. An example would be a drawing of a dog looking at a front door. Possible questions asked by the test-taker might then be: "Does this dog have an owner?", "Is it close to 5pm, when the owner is expected to return?", and so on.
- **Activity 2: Guessing Causes**. The test-taker is tasked with guessing what might be causing the action described in the provided image. Examples might be: "The owner is late from work", "The dog is a puppy waiting for its mother to come back from work for the day, which is chasing squirrels."
- **Activity 3: Guessing Consequences**. The test-taker is tasked with guessing the consequences of the action described in the provided image. Examples might be: "The puppy will be excited when his mother returns, and tail-wagging will be prominent among all parties."
- **Activity 4: Product Improvement**. The test-taker is tasked with improving a product. The product is described in 2–3 sentences. The test taker is then asked to think of the most interesting and unusual ways to improve the product for the end user. An example is a toy train. Improvements might then include: "Add a refrigerated car to the train to transport lunch to a family member across the room", "Add a drone as the engine to fly the train in the air", and so on.
- **Activity 5: Unusual Uses**. The test-taker is tasked with considering interesting and unusual uses of an item. The name of the item is given but not described to the test-taker. A standard example is a paper clip. Responses might include: "Hang ornaments from a Christmas tree", "Combine to create a bracelet", "Fashion an impromptu fish-hook."
- **Activity 6. Just Suppose**. The test-taker is given an improbable situation and tasked with imagining what *would* happen if the improbable situation were to occur. An example would be: "JUST SUPPOSE—all children became giants for one day out of the week. What would happen?" Responses might include: "Parents would need to order larger clothes", "Sports manufacturers would need to make much bigger soccer balls", "The bigger soccer balls could cause new damage to homes", and so on.

While Activity 5 may be considered a standard Alternative Uses Test (J.P. Guilford, 1967), the other activities were designed to target a range of creative thinking functions, indicative of the types of creative thinking of usefulness and value, very often encountered, for example, in real-life business and real life contexts. For more information about these general tasks and the TTCT, see also Torrance (1979).

### *Procedure*

To generate the GPT-4 submissions, the GPT-4 model was assessed through the cloud-based ChatGPT Plus application. As ChatGPT is a conversation-based model, the task instructions were entered as prompts exactly as provided within the TTCT. ChatGPT accepted the task prompts without any additional comments and did not ask any follow-up questions. No additional information, instructions, or guidance were provided. In addition, no preparation exercises or training were provided to ChatGPT prior to, or during, testing. To ensure the integrity of the TTCT and in accordance with intellectual property rights, prompts and responses were not shared with the public OpenAI database (this research opted out of data sharing with OpenAI). GPT-4 was kept at its default settings, including its default temperature setting of 0.7.

Task instructions were entered one at a time until all six tasks were completed as part of each individual submission. As OpenAI places a limit on the number of conversations that can be completed during one session with GPT-4, once a conversation limit was reached, the testing was continued after the required waiting period was completed using the following prompt, "Please continue the task." In addition, to prevent any form of learning effect, a new ChatGPT session was initiated for each new TTCT submission, ensuring no prior history would be available for each subsequent GPT-4 test attempt. For scoring, all ChatGPT responses were transcribed by research assistants into handwritten responses, ensuring a blinded and comparable evaluation with the human submissions by STS.

Students in the control group were enrolled in collegiate level entrepreneurship and finance courses and followed a standard research ethics (IRB) protocol as part of the study. The students were given the exact instructions and amount of time recommended in the TTCT. The tests were administered by an instructor with experience and knowledge of the TTCT and its administration. Students completed the TTCT booklets by handwriting their responses within the allotted time as per TTCT instructions.

All responses—human and AI—were included in the same group of submissions and blind-scored for fluency, flexibility, and originality by STS, as per the standard TTCT scoring method. Scoring of the handwritten responses was completed by human scorers at STS.

To account for the unique capabilities and limitations of AI models, several control variables were applied to the study. For example, ChatGPT can generate responses much faster than humans. Therefore, instead of time, the limitations in the number of responses set by the TTCT were utilized. Furthermore, at the time of this research, ChatGPT did not have image processing abilities, so the images in the TTCT (part of activities 1–3) were translated into text. The text translation used generic and basic language to convey the general idea of the images contained within the TTCT.

## Results

### *Fluency scores*

For overall fluency, all eight GPT-4 submissions scored within the top national percentile scored by STS, while the 24 human control results fell between the 3rd and 99th percentile (see Fig. 1). The mean for the GPT-4 group was 99 with a standard deviation (SD) of 0. The mean for the control group was 61.2 with SD of 29.0. A power analysis indicated sufficient sample size to determine significant difference between the results of the two groups. An independent samples *t*-test and the Mann-Whitney U test performed in SPSS both indicated a significant difference between the GPT-4 group and the control group for fluency scores with a p-value $< 0.001$ (see Fig. 4 for means comparisons across groups).

### *Flexibility scores*

For flexibility, the GPT-4 group's national percentile rankings ranged from 93 to 99, with a mean of 97.1 and SD of 2.2. The control group's flexibility rankings ranged from 1 to 97, with a mean of 56.8 and SD of 30.3 (see Fig. 2). An independent samples *t*-test and the Mann-Whitney U test indicated a significant difference between the GPT-4 group and the control group for overall flexibility scores with a p-value $< 0.001$ (see Fig. 4 for means comparison). However, as mentioned in the following discussion, flexibility scores on individual tasks provided a more nuanced—and perhaps more interesting—comparison of the experimental and control groups.

### *Originality scores*

For originality, all GPT-4 scores ranked within the top percentile of TTCT test-takers, yielding a mean of 99 and SD of 0 (see Fig. 3). The control group's originality scores ranged from 5 to 99, with a mean of 59.3 and SD of 29.2. An independent samples *t*-test and the Mann-

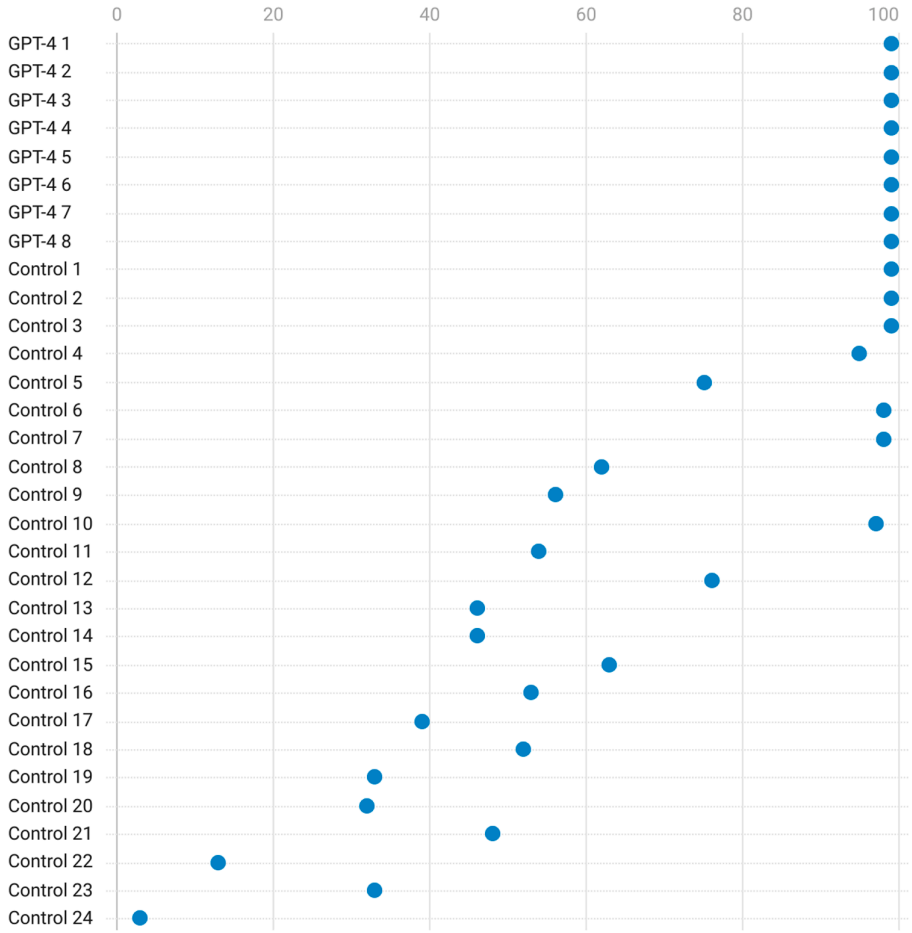



**Fig. 1.** Fluency National Percentile Ranks (GPT-4 and Control Group).

Whitney U test both indicated a significant difference between the GPT-4 group and the control group for originality scores with a p-value $< 0.001$.

*Individual activity results*

Given the TTCT's inclusion of activities designed to target a range of creative outputs, it is also instructive to review how GPT-4 performed on each individual TTCT activity in terms of fluency, flexibility, and originality. Table 1 lists each TTCT activity and the maximum score achievable for each individual measure.

Figs. 5–7 illustrate GPT-4′s performance on each task relative to the control group and the maximum points available in each activity. Interestingly, an independent samples test indicated significant difference for all results at the p=.01 level with the exception of individual flexibility scores for activities 1–5, which did not indicate significant difference (see Fig. 6). Likewise, a Mann-Whitney test indicated a significant difference for all outcome measures at the p=.01 level with the exception of flexibility scores for activities 2–4, which did not indicate a significant difference between the two groups.

## Discussion

While the results obtained from this study highlight several interesting areas of discussion, three points seem of particular relevance: (1) the unexpected originality of the most recent iterations of AI; (2) the lower relative flexibility scores of AI on certain individual tasks and activities; (3) the possible limits of current creativity assessments and/or conceptions of creativity.

*The originality of AI*

While GPT-4′s ability to generate large numbers of ideas (fluency) is perhaps expected, the ability of GPT-4 to generate novel and unexpected ideas (originality) is surprising. The results help illustrate the creative advances of AI, including OpenAI's GPT model. In 2022, a study by Stevenson et al. (2022) found that GPT-3′s performance on creativity tests was "impressive, and in many cases appears human-like" (Stevenson et al. 2022, p. 3). But the researchers found that GPT-3′s ability to generate unexpected and novel ideas, that is, engage in what is often defined as original thinking, did not match that of humans (Stevenson et al., 2022).

This research seems to be the first to show that AI matches or exceeds human abilities for original thinking. Based on the research, not only are the latest forms of AI generating large numbers of ideas (fluency) and different types, variations, and categories of ideas (flexibility), they are, for the first time, generating new, unique, and unexpected ideas (originality), performing in the top percentile for original thinking. In short, AI models like GPT-4 are becoming capable of producing ideas that humans consider to be original, novel, and unique.

It is perhaps worth repeating that the Torrance Tests include not only a standard Alternative Uses Task (AUT), but additional activities designed to evaluate a variety of real-life, everyday, and practical activities in business and modern life that require creative thought,

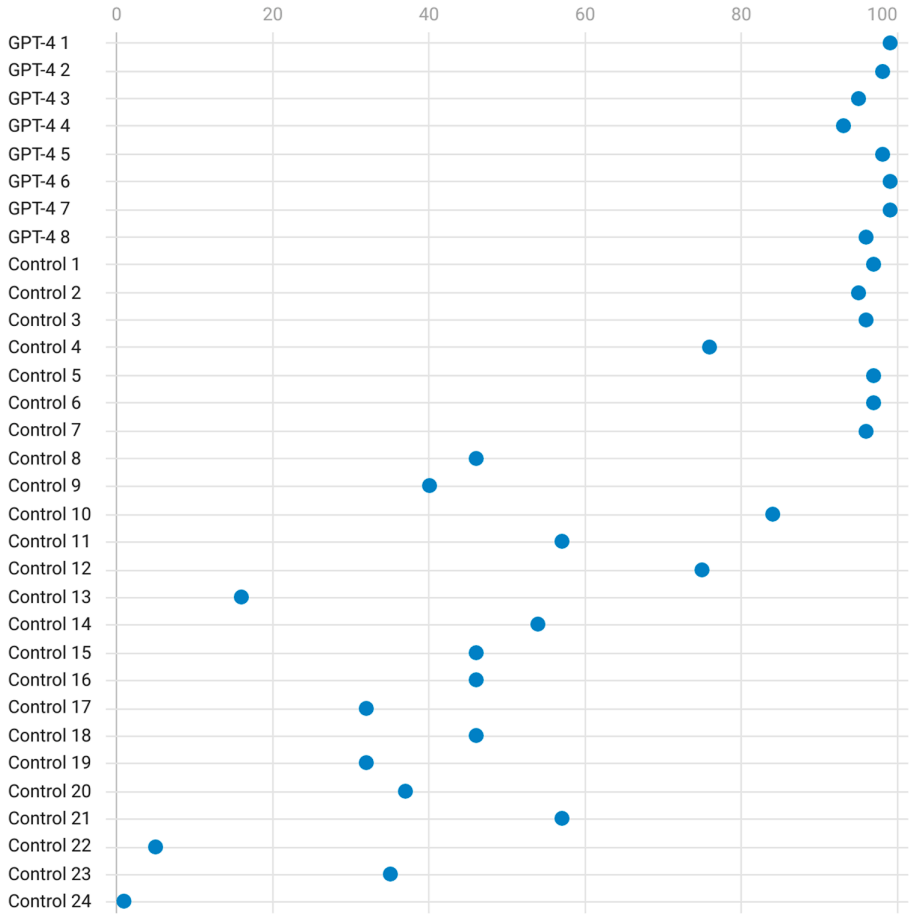



**Fig. 2.** Flexibility National Percentile Ranks (GPT-4 and Control Group).

including Asking Questions, Guessing Consequences, and Product Improvement. GPT-4 outperformed the human control group in all such activities. The originality of GPT-4 on Activity 6, Just Suppose, is especially intriguing. This activity has been designed to elicit creative responses from the test-taker based on the reading of an improbable, fictitious scenario, usually requiring strong imagination skills and truly novel thinking on the part of the human test-taker. That AI scored so much higher than humans on this type of imagination task (95% of its responses were considered original by the reviewers as compared to only 24% of responses from the control group) is striking. In this respect, GPT-4 marks a dramatic and unexpected shift in the creative abilities—and originality—of AI.

*The lower relative scores for AI flexibility*

GPT-4 scored lower on flexibility on certain activities of the TTCT. While the overall flexibility scores of GPT-4 still proved to be significantly different than the control group, this was not the case for the flexibility scores for the following activities: guessing causes, guessing consequences, and product improvement. A number of reasons might explain this disparity, such as, the lack of precise wording within prompts provided to GPT-4 to target different categories of ideas, or the possibility that the required training and current algorithms used by GPT-4 have not yet been fully developed to exploit flexibility. These results might suggest areas for future improvement of LLMs to better promote flexible thinking within AI models, or perhaps how human creativity might complement AI creativity to provide a more diverse and robust set of ideas when flexibility is required.

*Assessment and conception of creativity in the era of AI*

While the TTCT has long been considered a valid and reliable measure of creativity, the results of GPT-4 testing may simply highlight the limitations of existing creativity assessments. Although GPT-4 demonstrated high fluency, flexibility, and originality, assessments such as the TTCT may not fully capture the nuances and complexities of human creativity especially as related to person, process, and press—in this respect, current human scores on tests such as the TTCT may understate or incompletely measure human creativity. Further, the performance of GPT-4 could suggest that traditional creativity assessments need to be revised to differentiate between human and AI-generated creative outputs and evaluate other aspects of creativity beyond those currently assessed, including processes of convergent thinking.

Current product-based measures like the TTCT also raise interesting questions about creativity assessment in general. For example, *which* human raters are capable of best measuring human creativity? And which of these raters are capable of measuring AI creativity? In addition, is a truly accurate and unbiased assessment of creativity and originality possible within current assessments? A study by Licuanan et al. (2007) found that participants preferred ideas of low originality when evaluating highly original ideas. They also found that ideas of high originality were discounted because raters lacked the knowledge for accurately recognizing the originality of these ideas. In another study by Blair and Mumford (2007) it was found that human raters were more positive about readily understandable ideas that had short-term benefits, and that were consistent with current social norms, rather than truly novel ideas.

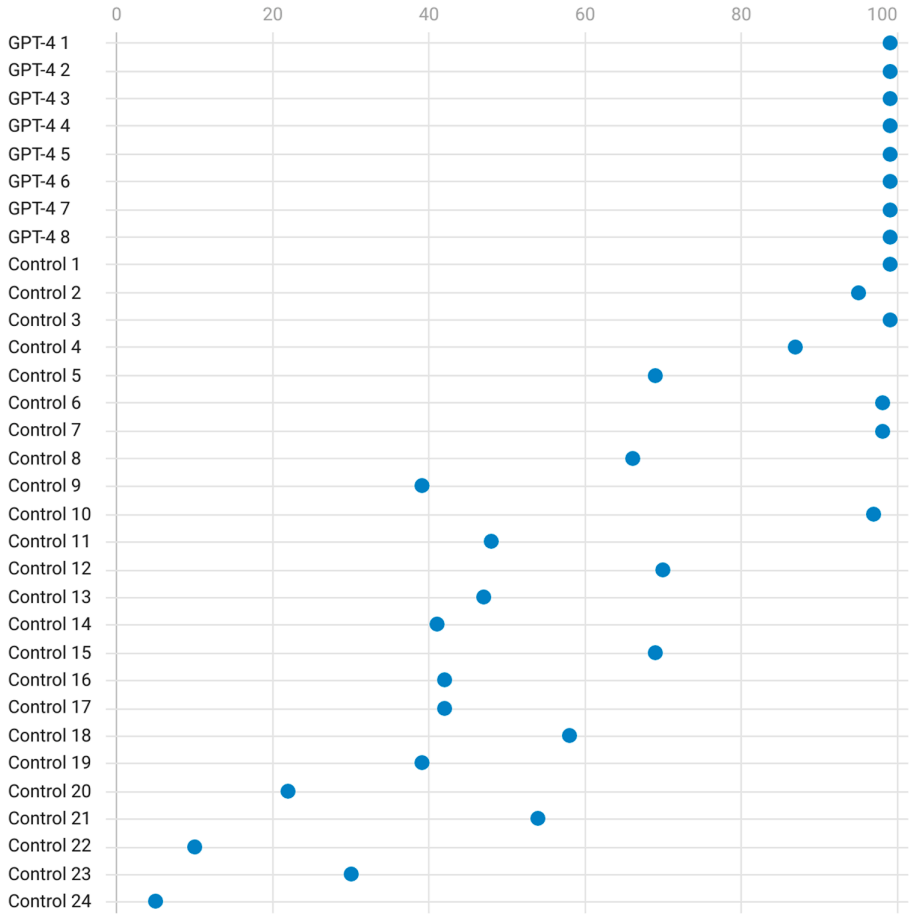



**Fig. 3.** Originality National Percentile Ranks (GPT-4 and Control Group).


**Fig. 4.** Mean Percentile Ranking (GPT-4 and Control Group).

**Table 1**
TTCT Tasks and Maximum Scores for Fluency, Flexibility, and Originality.

| Activity | Maximum Possible Score |
|---|---|
| 1. Asking Questions | 25 |
| 2. Guessing Causes | 25 |
| 3. Guessing Consequences | 25 |
| 4. Product Improvement | 32 |
| 5. Unusual Uses | 50 |
| 6. Just Suppose | 27 |

Interestingly, given the "all-knowing nature" of AI, tools such as GPT-4 may be uniquely suited for evaluating (or assisting in evaluating) creative ideas (see, for instance, Organisciak et al. (2022) for recent work in this area). Given how GPT-4 responded to the TTCT prompts, designed as they are to elicit creative output, AI seems to be able to distinguish between common responses and a request for more unique and original ideas. As such, AI may help evaluators better identify and become aware of the truly novel ideas when assessing creativity—and may provide a needed and valuable tool to better assess and develop human creativity.

Finally, the results from this study may demonstrate a need to reevaluate and/or expand current conceptions of creativity, especially as related to the notions of effectiveness, usefulness and value. That is, while GPT-4 generated surprisingly high marks for fluency, flexibility, and originality, AI may not yet be able to discern the effectiveness of its proposed original ideas, perhaps suggesting a limitation of current AI-based creativity as based on usefulness or effective solution-finding, an important area for future research and study.

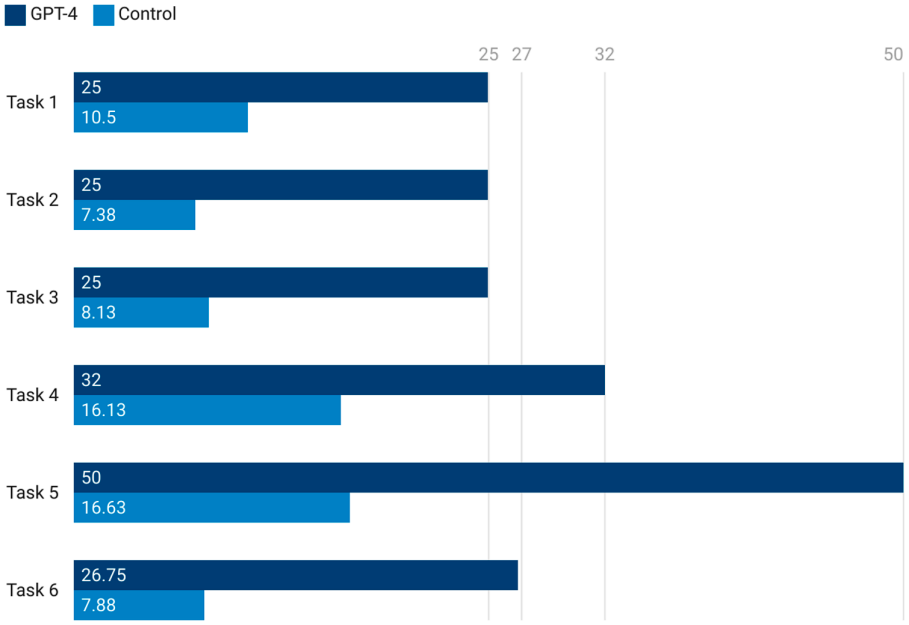



**Fig. 5.** Fluency Scores for Each Task (GPT-4 and Control).

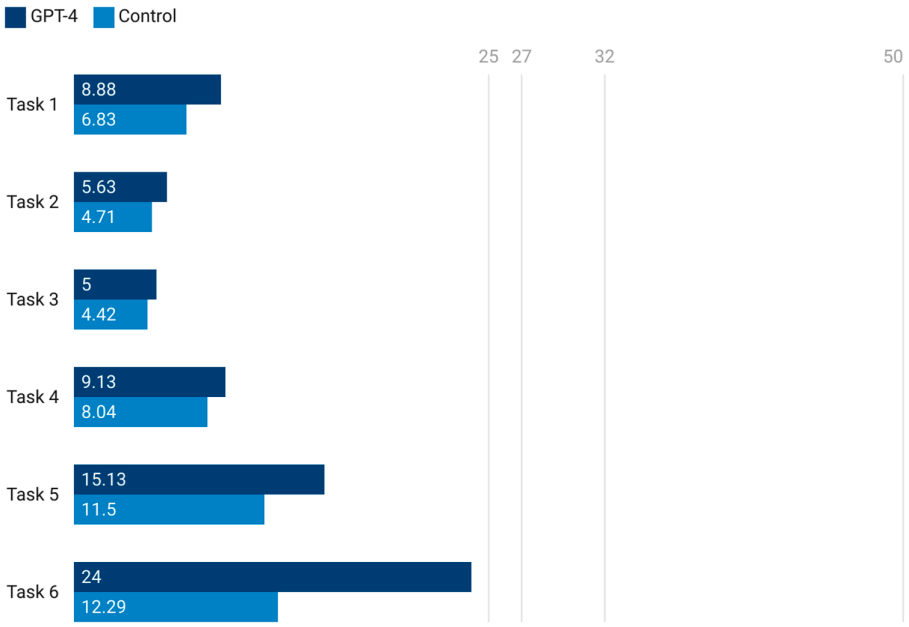



**Fig. 6.** Flexibility Scores for Each Task (GPT-4 and Control).

*Final comments*

The original founders of artificial intelligence—John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon—stated in 1955 that their goals for developing AI included, ‘‘the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it’’ (McCarthy et al., 1955, p. 2). The simulation by AI of a particular feature of human intelligence—creativity—seems to now be upon humankind.

This study provides new insight into the creative abilities of GPT-4 as assessed by the TTCT. The GPT-4 model generated impressive results for the TTCT dimensions of fluency, flexibility, and originality, suggesting that AI systems have the potential to produce viable creative output. Indeed, for the first time, an AI model demonstrated the ability to generate new, unique, and unexpected ideas that match or exceed the abilities of human originality.

Is, then, AI creative? From the perspective of generating novel and unexpected output—and based on currently accepted conceptions and assessment methods of creativity—this study must conclude that, yes, it is. The impact of this fact will likely shape not only the practical applications of AI’s simulated creativity in business and overall human life, but how we understand the unique operation of human creativity as well. The study therefore encourages additional research to further

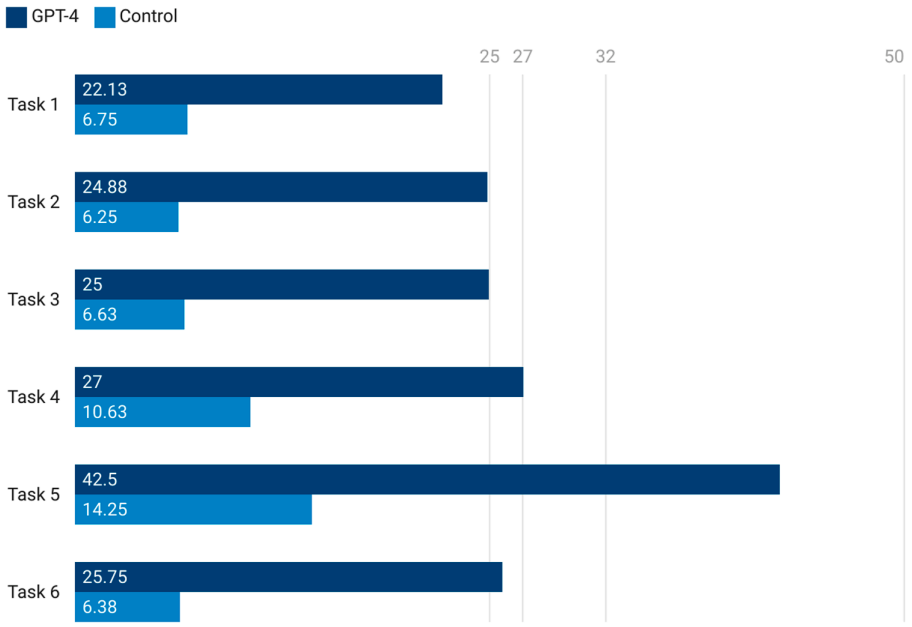



**Fig. 7.** Originality Scores for Each Task (GPT-4 and Control).

define, measure, and develop creativity—human and simulated—in the era of advanced AI.

## Funding

This research did not receive any type of grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013.
Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review, 55*, 589–656.
Blair, C. S., & Mumford, M. D. (2007). Errors in idea evaluation: Preference for the unoriginal? *Journal of Creative Behavior, 41*(3), 197–222.
Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. London: Routledge.
Boden, M. A. (2009). Computer models of creativity. *AI Magazine, 30*(3), 23. https://doi.org/10.1609/aimag.v30i3.2254
Caroff, X., & Besançon, M. (2008). Variability of creativity judgments. *Learning and Individual Differences, 18*(4), 367–371.
Cope, D. (1989). Experiments in musical intelligence (EMI): Non-linear linguistic-based composition. *Journal of New Music Research, 18*, 117–139.
Cordeschi, R (2007). AI turns fifty: Revisiting its origins. *Applied Artificial Intelligence, 21* (4-5), 259–279. https://doi.org/10.1080/08839510701252304
Davis, G. A. (1997). *Creativity is forever* (5th ed.). Kendall Hunt Publishing Company.
Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin, 136*(5), 822–848.
Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 35–40.
Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.
Han, J., Forbes, H., & Schaefer, D. (2021). An exploration of how creativity, functionality, and aesthetics are related in design. *Research in Engineering Design, 32* (3), 289–307.
Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20*, 171–178.
Kaufman, J. C., Beghetto, R. A., Baer, J., & Ivcevic, Z. (2010). Creativity polymathy: What Benjamin Franklin can teach your kindergartener. *Learning & Individual Differences, 20*, 380–387.
Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. *Journal of Creative Behavior, 41*, 1–27.
Lissitz, R. W., & Willhoft, J. L. (1985). A methodological study of the Torrance Tests of Creativity. *Journal of Educational Measurement, 22*, 1–11.
Lubart, T. (2017). The 7 C's of creativity. *Journal of Creative Behavior, 51*, 293–296. https://doi.org/10.1002/jocb.190
McCarthy, J., Minsky, M.L., Rochester, N., & Shannon, C.E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.
Mednick, S. A., & Mednick, M. T. (1967). *Examiner's manual, remote associates test: college and adult forms 1 and 2*. Boston: Houghton Mifflin.
Miller, A. I. (2019). *The artist in the machine: the world of AI-Powered creativity*. The MIT Press.
OpenAI. (2023). *GPT-4 technical report*.
Rhodes, M. (1961). An analysis of creativity. *Phi Beta Kappen, 42*, 305–310.
Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences, 15*(5), 537–546.
Runco, M.A. (2011) *Runco creativity assessment battery (rCAB)*. Creativity Testing Services.
Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92–96.
Sternberg, R. J. (1999). *Handbook of creativity*. New York, NY: Cambridge University Press.
Sternberg, R. J. (2018). What's wrong with creativity testing? *The Journal of Creative Behavior, 54*(1), 20–36.
Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H (2022). Putting GPT-3's creativity to the (Alternative uses) test. In *International conference on computational creativity*.
Organisciak, P., Acar, S., Dumas, D., Berthiaume, K. (2022). Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. Submitted for publication.
STS. (2017). TTCT administration manual. Scholastic testing services.
Torrance, E. P. (1979). *The search for satori and creativity*. New York: Creative Education Foundation Press.