Machine Intelligence Research
www.mi-research.net

22(3), June 2025, 417-436
DOI: 10.1007/s11633-025-1546-4

Assessing and Understanding Creativity in Large Language Models

Yunpu Zhao[1,2] Rui Zhang[2] Wenyi Li[3] Ling Li[3]

[1] Department of Computer Science, University of Science and Technology of China, Hefei 230026, China

[2] State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

[3] Institute of Software, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract: In the field of natural language processing, the rapid development of large language model (LLM) has attracted increasing attention. LLMs have shown a high level of creativity in various tasks, but the methods for assessing such creativity are inadequate. Assessment of LLM creativity needs to consider differences from humans, requiring multiple dimensional measurement while balancing accuracy and efficiency. This paper aims to establish an efficient framework for assessing the level of creativity in LLMs. By adapting the modified Torrance tests of creative thinking, the research evaluates the creative performance of various LLMs across 7 tasks, emphasizing 4 criteria including fluency, flexibility, originality, and elaboration. In this context, we develop a comprehensive dataset of 700 questions for testing and an LLM-based evaluation method. In addition, this study presents a novel analysis of LLMs′ responses to diverse prompts and role-play situations. We found that the creativity of LLMs primarily falls short in originality, while excelling in elaboration. In addition, the use of prompts and role-play settings of the model significantly influence creativity. Additionally, the experimental results also indicate that collaboration among multiple LLMs can enhance originality. Notably, our findings reveal a consensus between human evaluations and LLMs regarding the personality traits that influence creativity. The findings underscore the significant impact of LLM design on creativity and bridge artificial intelligence and human creativity, offering insights into LLMs′ creativity and potential applications.

Keywords: Large language models (LLMs), creativity assessment, prompt engineering, cognitive psychology, divergent thinking.

Citation: Y. Zhao, R. Zhang, W. Li, L. Li. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, vol.22, no.3, pp.417–436, 2025. http://doi.org/10.1007/s11633-025-1546-4

1 Introduction

In recent years, the realm of artificial intelligence (AI) has witnessed a meteoric rise in the development and sophistication of large language models (LLMs)[1, 2]. LLMs have significantly advanced in their capabilities in addressing a variety of conventional natural language processing tasks, such as reasoning and natural language understanding[3–6]. Moreover, LLMs also have demonstrated significant value in widespread applications. From transforming rudimentary text into compelling narratives[7, 8], unlocking a new realm of storytelling, to solving complex algorithmic problems[9], these models have shown a semblance of what could be interpreted as creativity. The practical manifestations of this creativity have penetrated various sectors, including science research, where they assist in idea generation and suggestion[6]; education, by providing personalized learning experiences[10]; and in the entertainment industry, creating music and art[11, 12]. In many of their applications, LLMs seem to exhibit the ability to generate original text, aiding tasks related to imagination and creativity, suggesting that they may indeed possess elements of creativity.

From the broad capabilities demonstrated by LLMs, the creativity they exhibit is a key reason they are considered powerful. However, behind the impressive abilities of LLMs lies a significant question that warrants careful examination: Do these models actually possess real creativity, or is their apparent intelligence merely an illusion – a complex imitation of human thinking created by their training paradigm? This question touches on the very nature of LLM intelligence, which may not be easily explained. Since LLMs have shown considerable creativity, understanding the extent and characteristics of this creativity is essential. Gaining deeper insight into the creativity of LLMs can not only guide us in further improving their performance but also in enhancing our understanding of the nature of their creativity. This, in turn, informs our daily use and application of these models, underscoring the need for an effective method to measure and assess their creativity. Specifically, creative abilities

Research Article
Manuscript received on August 24, 2024; accepted on January 15, 2025; published online on April 28, 2025
Recommended by Associate Editor Zhiyuan Liu
Colored figures are available in the online version at https://link.springer.com/journal/11633
© The Author(s) 2025

are critical for the following application scenarios. First, LLM can inspire humans on creative tasks and provide novel ideas, especially in research idea generation[13, 14]. It has also been suggested that the use of LLM can also lead to homogenization of creativity[15]. Second, humor generation with LLMs offer significant value in both creative and practical applications. By simulating human-like humor, LLMs can assist in content creation for entertainment, marketing, and social media. Finally, LLMs can serve as powerful cocreators in creative writings by generating narrative ideas, suggesting plot developments, or even drafting sections of text that inspire further refinement by human writers.

Creativity, as a term, traditionally refers to the natural ability to think innovatively, to make unconventional connections, and to devise solutions that are both novel and effective[16]. Assessing the creativity of LLMs is fraught with challenges. First, the question of creativity does not have clear answers to refer to. When we ask an LLM a question such as "what is the speed of light in vacuum in meters per second?", the answer can be formally vetted, given the objective nature of the topic. However, when posed with a prompt such as "what would be the implications if animals could talk?", the situation becomes different in this case because there is no definitive answer and the answer is open and divergent, making it challenging to judge the correctness of the output[17]. Additionally, since creativity encompasses various aspects, including originality and flexibility, it is necessary to design diverse tasks and criteria to measure these qualities effectively in LLMs. In addition, there are differences between LLMs and humans, which might lead to irrelevant responses or serious logical issues, requiring us to additionally assess these aspects. Finally, evaluating creativity necessitates a delicate balance between accuracy and efficiency, rendering traditional human-based evaluation methods less practical. Therefore, it is imperative to address the challenges outlined above to make a robust and sound assessment of creativity in LLMs.

Recognizing the need for a comprehensive assessment of LLM′s creativity, we design an efficient framework to automatically assess the creativity of LLMs by adapting and modifying the Torrance tests of creative thinking (TTCT)[18], a widely recognized tool in psychometrics′ research for human creativity assessment. To enhance the credibility of the results and reduce the randomness, seven verbal tasks, which use verbal stimuli, were selected. We employed GPT-4, the most advanced LLM, to expand the question set for each task, thereby constructing the testing dataset. To ensure a thorough and objective evaluation of creativity and capture creativity′s various manifestations, we combine diverse tasks and criteria. We design a comprehensive test protocol incorporating four criteria for measuring creativity: Fluency, flexibility, originality, and elaboration. We let the LLMs answer questions from the constructed dataset, obtaining many question-answer pairs. We utilized GPT-4 as an evaluator to assess each answer, as the GPT-4 is capable of effectively assessing the openness of responses and identifying their shortcomings and errors. Under proper prompt engineering, GPT-4 can efficiently and effectively complete the evaluation of the entire dataset results. Thus, we can achieve a balance between efficiency and accuracy in our assessment method.

We selected six popular LLMs as test subjects, each possessing different architectures and parameter scales. In addition to the overall testing, we conducted some additional exploratory experiments that investigate the changes of creativity levels exhibited by LLMs when given different types of prompts and different roles that LLMs play. Then, we designed a collaboration mechanism for LLMs to explore the impact of multiple LLMs collaborating on creativity. Last, we also performed some psychological experiments related to personality traits on the LLMs, including emotional intelligence (EI), empathy, the big five inventory (BFI) and self-efficacy. Because we found in relevant psychological research showing that human creativity is correlated with these personality traits and we verified the consistency between LLMs and humans in this regard.

Our experiments and analysis yielded several conclusions. First, there are significant differences in creative performance among different models, even among those of the same scale with an equal number of parameters. This variation primarily exists between different types of models. Their differences are reflected mainly in the model architecture, parameter settings during training, alignment strategies, and the datasets used for training. Additionally, we observed that models generally excel in the elaboration metric, but tend to be less adept in demonstrating originality. In addition, the type of prompt and the specific role-play request given to the model also plays a significant role in influencing its creative output. When the models are given instructive prompts or chain-of-thought prompts, there is a significant increase in the level of creativity. Additionally, having LLM play different roles leads to notable differences; the role of a scientist demonstrates the highest level of creativity. Many roles even show a decrease compared to the default scenario, but there is generally an improvement in originality. Then, collaboration among multiple LLMs can enhance the level of creativity, with the most notable improvement in originality. Finally, the results of the psychological scale revealed consistency between LLMs and humans in terms of associated creativity factors, such as emotional intelligence (EI), empathy, self-efficacy, and others.

2 Related works

2.1 Creativity assessment in psychological research

The question of creativity assessment has been a prominent focus on the creativity research, especially since the 1950s, marking the inception of a systematic study into individual differences in creativity[19]. For example, Guilford pioneered the research on creativity and his famous structure of intellect model was mainly about defining and analyzing the factors constituting intelligence, where creativity plays a major driving force in his theory[20]. In recent years, many new developments regarding the measurement of divergent thinking, consensual assessment technique and subjective ratings, and self-report methodology[21–23] have emerged. Although advances in methodology and technology have led to important developments regarding creativity assessment, some assessment methods have long been described as "gold standard" for creativity assessment[24, 25]. Among them, TTCT[18] has been the most widely used and researched test of creativity, having extensive data to support its reliability and validity. Research on TTCT reports good reliability scores for scoring and test-retest reliability[26].

TTCT is designed to identify and assess an individual′s creative potential by exploring various dimensions. Contrasting conventional assessments that emphasize convergent thinking, the test fosters divergent thinking, encouraging participants to generate multiple solutions to open-ended, ambiguous problems. TTCT has been widely applied in educational settings, organizational assessments, demonstrating its versatility and comprehensive approach to measuring creativity. Its ability to tap into various facets of creative thinking has made TTCT a reliable and respected tool[27]. Owing to the authority and comprehensiveness of the TTCT, we select tasks from the TTCT to construct our dataset.

2.2 Creativity and personality: Findings in psychological research

Research has revealed that creativity is not solely a fixed human personality trait. It evolves from a combination of individual processes such as cognitive, affective, behavioral, and contextual factors. Some psychologists have conducted a detailed meta-analysis of papers exploring the relationship between creativity and various personality traits[28, 29].

These studies′ results highlight a correlation between creativity and a plethora of personal factors. Notably, elements such as emotional intelligence, divergent thinking, openness to experience, and intrinsic motivation stand out as strong influencers. However, factors such as age, intelligence, and gender exhibit a relatively milder association with creativity, signifying a varied spectrum of influence across different personal traits. Since large language models have exhibited some personality traits, we conducted experiments to test whether these findings also hold true in LLMs.

2.3 Assessing the creativity of large language models

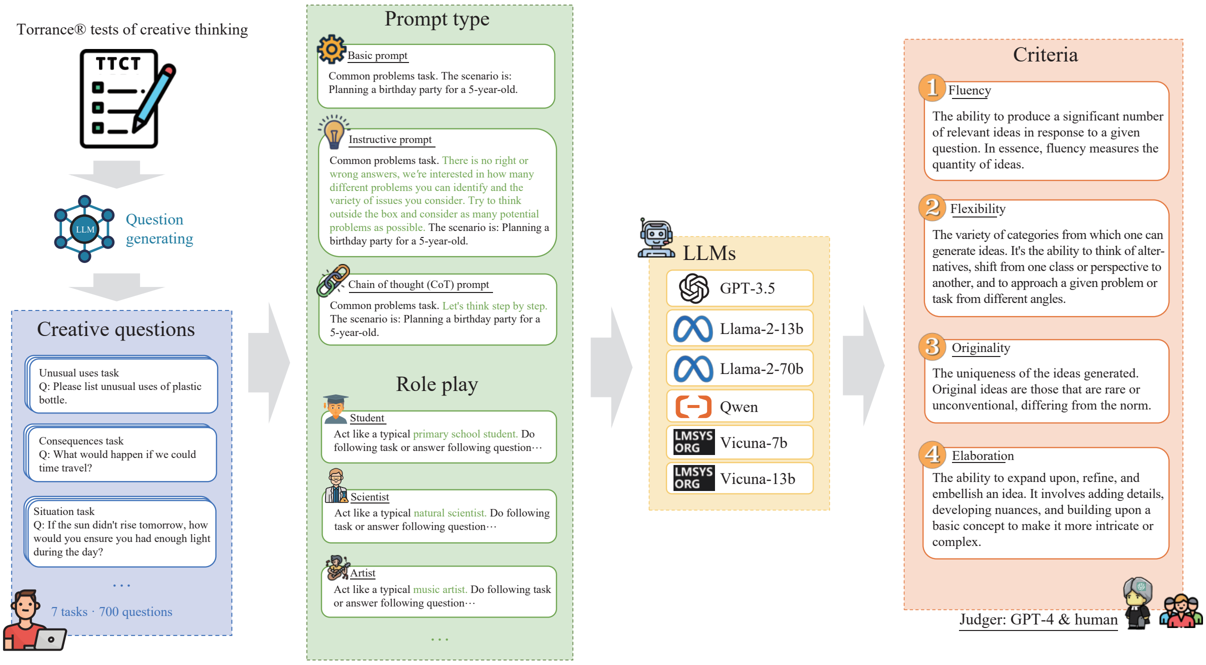
The emergence of abilities from LLMs continually surpasses people′s expectations, and the evaluation of various abilities of LLMs has received widespread attention[30]. Currently, most evaluations focus on the ability of LLMs to solve tasks, with fewer evaluations combining aspects of psychology.

Although some studies have focused on the intersection of LLM with psychology and cognitive science[31], work discussing the creativity of LLM is still in a relatively early stage. Current studies somewhat focused on exploring the creativity of LLMs, primarily from the standpoint of creativity theory, which aims to elucidate the definitions and challenges of applying creativity theory within the context of LLMs[32]. Some initial evaluations of creativity in LLMs have also been undertaken[33–35]. However, these works only employed simple tasks such as the alternative uses task (AUT) to assess creativity, and the lack of comparison between various LLMs limits the validity of their conclusions. It is worth mentioning that in [36], the authors used the standard TTCT to assess GPT-4′s creativity. The results show that GPT-4 achieved human top 1% levels in fluency and originality, along with a high score in flexibility. This study leans more towards comparing advanced large language models (LLMs) with human benchmarks. The original TTCT test protocol does not seamlessly adapt to assessing creativity in LLMs, as the limited sample of questions could induce randomness and accidental outcomes, making hypothesis testing challenging when comparing different models. Furthermore, expanding the number of question sets leads to high time costs in human-based evaluations.

Due to the differences between humans and LLM, it is problematic to directly use the TTCT′s test protocol to benchmark LLMs′ creativity. To address this dilemma, we propose a new framework for systematic analysis LLM′s creativity. This framework comprises carefully crafted metrics used in TTCT and a dataset that accounts for seven tasks. We will dive into detail of the framework in Section 3.

3 Overview of the framework

In this work, we design an overall framework to evaluate LLM′s creativity, as shown in Fig. 1. First, we constructed a dataset containing 700 questions of 7 tasks



Unusual uses task
Q: Please list unusual uses of plastic bottle.

Consequences task
Q: What would happen if we could time travel?

Situation task
Q: If the sun didn't rise tomorrow, how would you ensure you had enough light during the day?

…

7 tasks · 700 questions

Common problems task. The scenario is: Planning a birthday party for a 5-year-old.

Common problems task. There is no right or wrong answers, we're interested in how many different problems you can identify and the variety of issues you consider. Try to think outside the box and consider as many potential problems as possible. The scenario is: Planning a birthday party for a 5-year-old.

Common problems task. Let's think step by step. The scenario is: Planning a birthday party for a 5-year-old.

Act like a typical primary school student. Do following task or answer following question···

Act like a typical natural scientist. Do following task or answer following question···

Act like a typical music artist. Do following task or answer following question···

…

The ability to produce a significant number of relevant ideas in response to a given question. In essence, fluency measures the quantity of ideas.

The variety of categories from which one can generate ideas. It's the ability to think of alternatives, shift from one class or perspective to another, and to approach a given problem or task from different angles.

The uniqueness of the ideas generated. Original ideas are those that are rare or unconventional, differing from the norm.

The ability to expand upon, refine, and embellish an idea. It involves adding details, developing nuances, and building upon a basic concept to make it more intricate or complex.

Fig. 1 Overview of the creativity assessment framework. A TTCT-inspired dataset was constructed to evaluate LLMs under varied prompts and role-play settings. GPT-4 served as the evaluator to score model outputs.

that were derived and modified from the psychology scale of the TTCT and expanded the number of questions via GPT-4. We tested six models on four different criteria using the dataset we constructed. Following this, we conducted a series of experiments on the creativity of LLMs when giving different types of prompts and assigning different roles to LLMs. Finally, we used the GPT-4 as the evaluator to obtain the performance results of the LLMs and verify the consistency of the LLM-based evaluation with humans.

3.1 Dataset construction

This research utilized a modified version of the TTCT verbal test, which includes tasks based on verbal stimuli. The seven selected tasks: 1) Unusual uses, 2) consequences, 3) just suppose, 4) situations, 5) common problems, 6) improvements, and 7) imaginative stories, were chosen to capture a broad spectrum of creative thinking abilities. These tasks are adapted from the widely used TTCT, which has also served as the basis for recent work in the field of LLM evaluation[37]. The tasks we choose align with widely accepted models of creativity such as Guilford's structure of the intellect model and involve both divergent and convergent thinking[20]. Meanwhile, TTCT tasks, especially in their divergent thinking focus, align with the Geneplore model[38] by emphasizing idea generation (fluency and originality) and flexibility (the ability to shift between categories or approaches). Thus, the tasks capture both novelty and usefulness, which are central to most modern definitions of creativity. This makes them sufficient for assessing a holistic view of creative potential.

Specifically, each task includes one hundred questions generated by GPT-4 using few-shot prompts. The seven tasks were generally structured as follows:

1) Task 1: Unusual uses. This task challenges individuals in their ability to think of as many unusual and diverse uses as possible for a common object within a limited time frame. The object in question is typically everyday and familiar, such as a brick, paper clip, or newspaper.

2) Task 2: Consequences. This task focuses on the ability to foresee consequences or outcomes of an unusual or hypothetical situation. For example, what would be the implications if animals could talk?

3) Task 3: Just suppose. This task encourages imaginative and speculative thinking by asking participants to consider hypothetical, often fantastical, scenarios and their implications. For example, just suppose you woke up one morning and found you could fly. What would you do? List as many things as you can think of.

4) Task 4: Situation task. This task is designed to assess creative thinking by evaluating how individuals respond to and interpret a given situation. This task emphasizes understanding social dynamics, empathy, and the ability to consider multiple perspectives or solutions. For example, if all books were to disappear, how would you gain knowledge?

5) Task 5: Common problem. This task focuses on everyday problems that are familiar to most people, requiring participants to generate innovative and effective solutions. For example, organizing a cross-country road trip or building a tree house.

6) Task 6: Improvement. This task focuses on assessing an individual's ability to enhance or modify existing objects or ideas. The given object is similar to the unusual uses task.

7) Task 7: Imaginative stories. This task is designed to assess creativity through narrative and storytelling with a given prompt. This task emphasizes the ability to construct original, coherent, and imaginative stories, showcasing an individual's creative potential in terms of narrative ability. Examples of given prompts are "The Invisible Elephant" or "The Book that Wrote Itself".

Each task includes 100 questions generated by GPT-4 via few-shot prompts. GPT-4 can generate a diverse and comprehensive set of similar problems based on the given examples, and all problems have been validated by humans to ensure usability. In addition, we conducted experimental validation of domain generality across different tasks. Cronbach's Alpha and inter-task correlations indicate that our task selection is effective and sufficient.

3.2 Evaluation criteria

To provide a comprehensive evaluation of an individual's creative abilities, we should consider not only the quantity of ideas they produce, but also the quality, diversity, and depth of those ideas. We have four criteria for creativity evaluation:

1) Fluency. This refers to the ability to produce a significant number of relevant ideas in response to a given question. In essence, fluency measures the quantity of ideas.

2) Flexibility. This assesses the variety of categories from which one can generate ideas. It is the ability to think of alternatives, shift from one class or perspective to another, and to approach a given problem or task from different angles.

3) Originality. This measures the uniqueness of the ideas generated. Original ideas are rare or unconventional, differing from the norm.

4) Elaboration. This refers to the ability to expand upon, refine, and embellish an idea. It involves adding details, developing nuances, and building upon a basic concept to make it more intricate or complex.

These criteria aim to provide a comprehensive assessment of an individual's creative potential. The motivation behind using these specific dimensions is grounded in the theoretical and empirical research on creativity[39, 40], which suggests that creative thinking involves not just the generation of new ideas but also the ability to manipulate, refine, and apply these ideas effectively. The four criteria are based on long-standing psychological frameworks for creativity assessment, particularly the TTCT. These dimensions collectively capture distinct and complementary facets of creative thinking and have been extensively validated in psychological and educational research and are considered gold standards in creativity assessment.

3.3 LLM-based evaluation

Standard TTCT evaluation methods require trained psychologists to follow professional manuals to assess the results, and an individual's single test only contains answers to a very limited number of questions. When evaluating creativity in LLM, both the insufficient sample of responses and the high human resource costs limit the application of creativity tests on LLMs. Recent psychological research has focused on the automated assessment of creativity[41, 42]. However, these methods often have limitations, such as being tailored to specific tasks or requiring prepared reference answers, which prevent their generalization to a variety of tasks and a larger number of questions.

With the rapid development of LLM capabilities, the evaluation methods for many natural language processing tasks have evolved from traditional human annotation to reference-based automated methods, and now, to methods on the basis of LLMs. LLMs are increasingly playing the role of judges in tasks such as question-answering, translation, and text quality assessment[43–46], giving rise to various evaluation framework[47–49]. According to experimental results from relevant literature, LLM exhibits higher correlation with human evaluations compared with traditional automated technologies[50, 51]. In this study, on the basis of the evaluation criteria from Section 3.2, we utilize GPT-4 to score the answer. For each criterion, the LLM needs to complete the Likert scale based on the responses. Additionally, we verified the consistency between the evaluations made by LLM and human evaluations.

4 Evaluation and results

We conducted a statistical analysis of the creativity scores of 6 popular LLMs across seven tasks, totaling 700 questions. We unveiled hidden conclusions within the data results from various dimensions. We compared the differences in creativity levels between the models, and we compared the performance variations under different criteria within the same model. Subsequently, we experimented with many types of prompts to see whether changes in prompts would affect the models' levels of creativity. Since LLMs possess the ability to play user-specified roles, we select six typical human identities to explore the impact on creativity under different role-playing conditions. Finally, we utilize some psychological scales to test the LLMs, investigating the correlation between the personality traits of the LLMs and creativity.

4.1 Experimental settings

4.1.1 Tested models

We tested six of the most advanced LLMs, which are listed below. All the models were implemented with the open-source repository HuggingFace[52].

1) GPT-3.5. GPT-3.5 is a language model developed by OpenAI, which is an advanced version of the GPT-3 model. It is capable of generating natural language text and code. GPT-3.5 was trained on an Azure AI supercomputing infrastructure. The versions we used in the experiments are GPT-3.5-turbo-0613.

2) LLaMA-2. LLaMA-2 is a family of state-of-the-art open-access large language models released by Meta and Microsoft[2]. It is built upon success of its predecessor, LLaMA-1. LLaMA-2 is specifically designed to facilitate the development of generative AI-powered tools and experiences. It is available for free research and commercial use. LLaMA-2 release introduces a family of pretrained and fine-tuned LLMs, ranging in scale from 7 B to 70 B parameters. The versions we used in the experiments are LLaMA-2-13b-chat-hf and LLaMA-2-70b-chat-hf.

3) Vicuna. Vicuna is a lightweight, accurate, and efficient language model developed by a team of researchers from several universities, including UC Berkeley, Carnegie Mellon University, Stanford University, and UC San Diego[44]. It was built from Meta's adaptable LLaMA model, which was fine-tuned on a dataset of around 70 000 human-generated conversations from the ShareGPT website. The versions we used in the experiments are Vicuna-7b-v1.5 and Vicuna-13b-v1.5.

4) Qwen. Qwen (abbr. Tongyi Qianwen), proposed by Alibaba Cloud[53]. It is a transformer-based large language model, which is pretrained on a large volume of data, including web texts, books, codes, etc. The versions we used in the experiments are Qwen-7b-chat.

4.1.2 Details of hyperparameters

The models used in our experiment primarily originate from the open-source HuggingFace platform. The specific versions of these models have already been reported above. In this section, we present the experimental parameters and other settings related to the experiment.

For an LLM based on the transformer architecture, there are certain parameters that directly affect the output of the model.

1) Max tokens. This parameter controls the maximum number of tokens to generate in the chat completion. In our experiment, this value is uniformly set to 512, ensuring that the output length is sufficient to maintain the quality of the answers.

2) Temperature. The parameter is a crucial factor in determining the nature of the model's responses. This is a hyperparameter that influences the randomness or unpredictability in the model's responses. Essentially, its mechanism is to change the probability distribution of the model's output logits. However, according to our experiments, changes in temperature do not significantly affect creative performance, which appears quite random. Therefore, in our experiments, the temperature is uniformly set to 1.

3) Top_p. Top_p is also a parameter used to control the diversity of the generated text, also known as "nucleus sampling". This parameter's full name is "top probability", which is typically represented by a value between 0 and 1, indicating the cumulative threshold of the highest probabilities chosen in the probability distribution when generating the next token. In our experiments, top_p is uniformly set to 1.

4) Top_k. This parameter is used when generating the next token to limit the model to consider only the top_k tokens with the highest probability. This strategy can reduce the likelihood of the model generating meaningless or repetitive outputs, while also improving the speed and efficiency of the model generation. In our experiments, the top_k is uniformly set to 50.

GPT-4 serves as the judge for our LLM-based evaluation, with its relevant parameters set to default. The version used is GPT-4-0613. In addition, all prompt templates used in the experiment are provided in the appendix.

4.2 Results of different models and criteria

We assessed the responses of six language models to 700 questions, with GPT-4 serving as the evaluator across all creativity dimensions. We first evaluate the average score of each model across all tasks, as shown in Fig. 2(a) and Table 1. It can be observed that GPT-3.5 has the highest level of creativity, followed by the LLaMA-2 architecture models, then the LLaMA-based fine-tuned model vicuna, and finally Qwen. The experimental results from the perspective of the model suggest that the type of model has a significant effect on creativity, whereas the scale of parameters does not have a decisive influence. Different types of models vary in their architectures, alignment strategies, and the datasets used during training. These factors are likely to be key determinants of the level of creativity. Similar findings can also be observed in other LLM evaluation papers[54–56]. For example, in Toolbench[56], the 30 B version of LLaMA outperforms the 65 B version of LLaMA in many tasks, and text-daVinci-003 also performs better overall than GPT-3.5.

To further validate the ranks of the models, we conducted pairwise comparisons between the models, as shown in Fig. 2(b). Each cell in this heatmap represents the win rate of the model on the y-axis in terms of creativity score compared to the model on the x-axis. The win rate scores are consistent with the strengths and weaknesses of the models shown in Fig. 2(a), and we conducted statistical tests for significance, which are marked

Table 1 Comparative creativity scores across LLMs

	Fluency	Flexibility	Originality	Elaboration
Common problem task				
GPT-3.5	**4.975**	**4.650**	**3.870**	4.735
LLaMA-2-13b	4.940	4.480	3.770	4.890
LLaMA-2-70b	4.920	4.545	3.720	**4.905**
Qwen	3.090	2.890	2.360	3.360
Vicuna-13b	4.910	4.320	3.510	4.415
Vicuna-7b	4.880	4.270	3.380	4.200
Consequences task				
GPT-3.5	4.855	4.810	**4.105**	**5.000**
LLaMA-2-13b	4.910	**4.830**	4.080	**5.000**
LLaMA-2-70b	**4.930**	**4.830**	3.995	4.995
Qwen	4.410	4.430	3.610	4.875
Vicuna-13b	4.260	4.295	3.580	4.850
Vicuna-7b	4.535	4.435	3.660	4.920
Improvement task				
GPT-3.5	**5.000**	**4.970**	**4.620**	**4.980**
LLaMA-2-13b	4.980	4.850	4.150	4.890
LLaMA-2-70b	4.965	4.800	4.085	4.900
Qwen	4.870	4.550	3.760	4.700
Vicuna-13b	4.970	4.410	3.600	4.380
Vicuna-7b	4.950	4.560	3.860	4.640
Imaginative stories task				
GPT-3.5	**4.160**	**4.200**	**4.475**	**4.925**
LLaMA-2-13b	3.720	3.620	4.030	4.730
LLaMA-2-70b	3.830	3.660	4.050	4.700
Qwen	3.240	3.510	3.740	4.430
Vicuna-13b	3.310	3.610	3.750	4.490
Vicuna-7b	3.280	3.470	3.760	4.580
Just suppose task				
GPT-3.5	**3.960**	**4.310**	4.030	**4.930**
LLaMA-2-13b	3.830	4.160	**4.040**	**4.930**
LLaMA-2-70b	3.795	4.090	3.750	4.870
Qwen	3.580	3.840	3.250	4.640
Vicuna-13b	3.410	3.580	3.030	4.550
Vicuna-7b	3.480	3.860	3.240	4.600
Situation task				
GPT-3.5	4.790	4.670	3.940	4.970
LLaMA-2-13b	4.195	4.390	3.920	4.850
LLaMA-2-70b	**4.850**	**4.800**	**4.050**	**4.990**
Qwen	3.940	4.010	3.170	4.590
Vicuna-13b	3.970	3.970	3.140	4.600
Vicuna-7b	4.020	3.980	3.210	4.620

Table 1 (continued) Comparative creativity scores across LLMs

	Fluency	Flexibility	Originality	Elaboration
		Unusual uses task		
GPT-3.5	**5.000**	**4.920**	**4.670**	4.895
LLaMA-2-13b	4.990	4.860	4.280	**4.910**
LLaMA-2-70b	4.980	4.850	4.255	4.880
Qwen	4.905	4.210	3.690	4.130
Vicuna-13b	4.860	4.060	3.670	3.760
Vicuna-7b	4.910	4.640	3.940	4.300

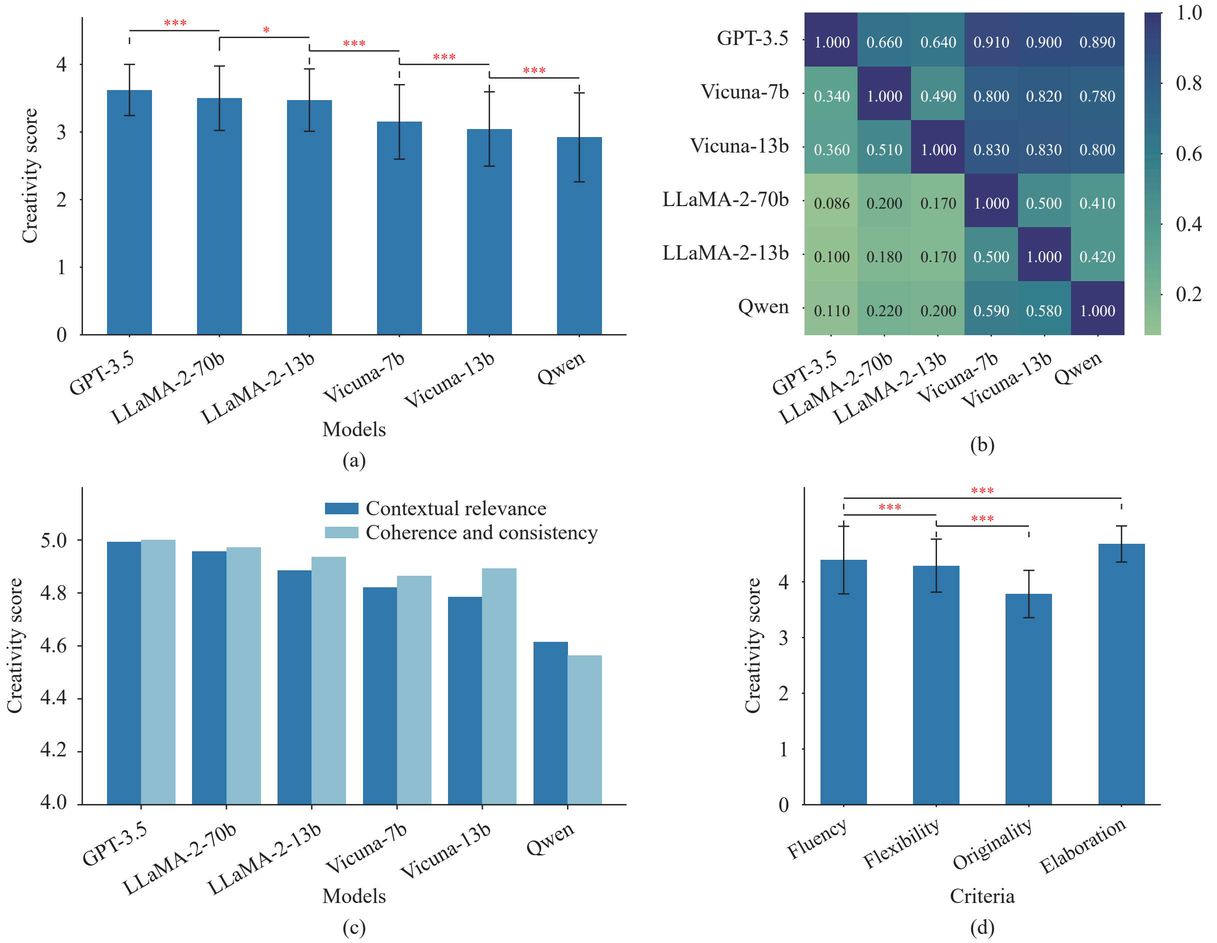


Fig. 2 Creativity performance of different LLMs across models and criteria. (a) Overall creativity scores with error bars showing standard deviations. Significance is marked using the Wilcoxon signed-rank test. (b) Pairwise win rate heatmap. (c) Scores for relevance and consistency. (d) Average scores across four creativity dimensions. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

in Fig. 2.

Next, we evaluate the average scores of each criterion across all tasks, as shown in Fig. 2(d). The score for elaboration is consistently high across all tasks, while originality is relatively lower, with fluency and flexibility scoring in the middle. The capabilities of LLMs inherently stem from training on human language corpora, so it is intuitive that they score relatively lower in originality. The creativity of LLM is likely to be a manifestation of the combination of existing human knowledge, and how to improve the originality of LLM is an important future endeavor. The elaboration metric reflects the degree of refinement of a creative idea, and LLM′s ability to articulate this has always been outstanding.

However, if we focus on the performance of different tasks, we will find that there are significant variances in creativity performance under different tasks, as shown in Fig. 3, which shows the radar charts of the performance of six models across seven tasks. It can be observed that most models exhibit a higher level of overall creativity in the common problem, consequences, and unusual uses tasks, while the overall creativity level is lower in the just suppose and imaginative stories tasks, reflecting the varying degrees of creative difficulty presented by different tasks.

At last, there are some differences between humans and LLMs when answering questions. In the case of LLMs responding to human prompts, issues such as irrelevance to the topic or logical errors may arise. On the other hand, humans generally maintain consistency in their answers. So we have evaluated the responses of all models in this regard, and it is observable that there are significant differences in relevance and coherence among the various models, as shown in Fig. 2(c). The results show that the GPT-3.5 and LLaMA models performed well, while the Vicuna and Qwen models had poorer performance. Sometimes, Vicuna and Qwen fail to understand the question properly, leading to irrelevant answers. Sometimes, due to a misunderstanding of the question, they refuse to answer. We all consider these as manifestations of a lack of creativity. This issue may be related to the alignment strategies employed and the training datasets of the models.

4.3 Results of different prompt types

The prompt is a crucial component of the LLM model, as it provides the necessary context and information for LLMs to generate a relevant and coherent response. The quality and type of prompt can significantly impact the quality of the generated response. Therefore, we believe that the type of prompt can greatly influence the creativity of LLMs.

In our experiment, we designed and compared four different types of prompts: basic prompt, instructive prompt, post-instructive prompt and chain of thought

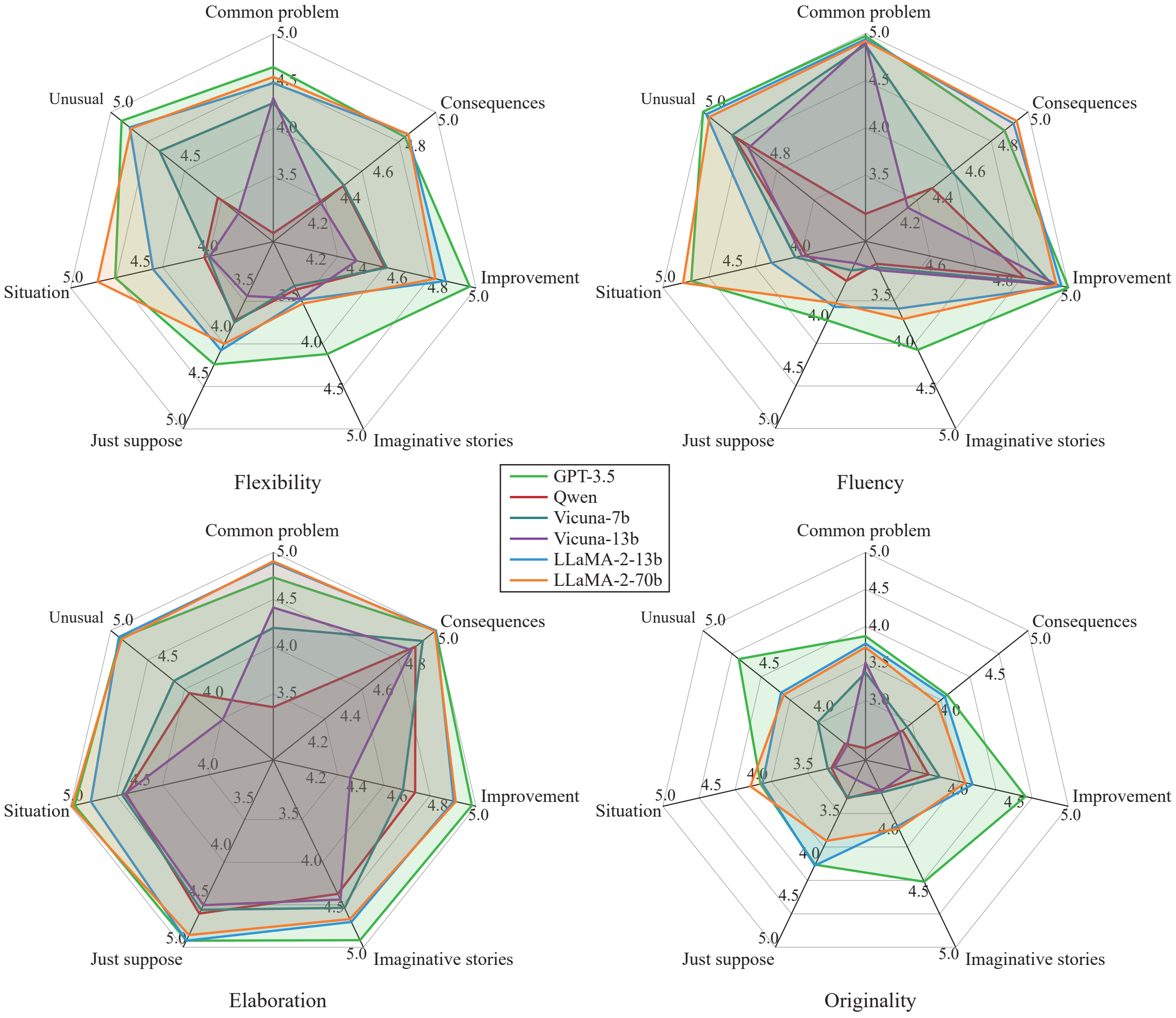


Fig. 3 Radar comparison of LLM creativity across tasks (Colored figures are available in the online version at https://link.springer.com/journal/11633)

(CoT) prompt. Herein, the basic prompt contains only the essential information needed to describe the task, simple and clear. The instructive prompt provides a detailed description of the expected answer, outlining what constitutes a creative response. The post instruction prompt uses two rounds of prompts, starting with a basic prompt for the LLM to give a basic answer, then giving some instruction about creativity (the same as instructive prompt). LLM revises the answer given in the first round on the basis of the given instruction and gives the revised answer. Chain of thought is a technique that enables complex reasoning capabilities through intermediate reasoning steps or just a single explicit prompt like "let's think step by step". We utilize the technique used in [57] to design our CoT prompt. The example of prompts is shown in Fig. 1.

As shown in Figs. 4(a)–4(c), we obtained data on the performance of LLMs in terms of creativity across all tasks and all criteria under different prompt types. From the perspective of the task, the inclusion of instructive language in prompts has improved creativity in all tasks except for "unusual uses". The reason for the lack of improvement in "unusual uses" may be that the task description is already clear enough and the required divergent thinking ability is relatively simple. When the CoT prompt is used, there has been an increase in the level of creativity in three tasks, indicating that some tasks require a higher level of convergent thinking ability to demonstrate creativity. In the case of post instructions, the greatest differences were shown between tasks. While a few tasks, such as imaginative stories and just suppose, showed some rise, most of the rest did not have a significant boosting effect, and even produced a drastic drop on the unusual uses task. We speculate that the main reason may be that under multiple rounds of dialog, the post-instruction actually implicitly negates the initial response, resulting in the inability to be in a position to come up with a more creative response on a relatively simple task, such as unusual uses. From the perspective of creativity criteria, instructive prompts clearly significantly enhance both flexibility and originality but do not increase elaboration. On the other hand, CoT prompts slightly improve elaboration. Both types of prompts are beneficial to the fluency of the responses. In summary, the creative performance of LLMs, like their other abilities, is significantly influenced by the prompts. Effective prompt engineering is greatly beneficial for better harnessing the potential creativity of LLMs. For the post instruction prompt, only originality criteria have an obvious increase, and even a significant decrease in the fluency and flexibility criteria. For the same reason as previously stated, the second round of the responses will naturally negate the initial responses, resulting in a lack of flexibility and fluency in the final answer.

4.4 Results of playing different roles

LLMs possess the remarkable capability to adopt the roles specified by users, which can subsequently influence their outputs. This adaptability enables the models to deliver tailored responses, aligning with the context and characteristics of the assumed identities. In our experiment, we attempted to specify the exact identity and role of the LLM within the system prompt. The primary objective of this approach was to ascertain whether the LLM could enhance its creative expression by adopting specific roles and to determine if this influence is consistent with the cognitive patterns observed in reality.

As shown in Fig. 4(d), we assigned six distinct roles to the LLM: engineer, farmer, merchant, scientist, artist, and primary school student, requiring the model to perform tasks in alignment with the characteristics of the respective roles. The results demonstrated that across all creativity assessment criteria, the creativity level of the scientist surpassed that of the other six roles, reflecting a correlation between the accumulation of knowledge, educational attainment, and the level of creativity. Furthermore, when the LLM was playing different roles than the scientist was, the values of fluency and flexibility have decreased, yet originality has increased significantly. This suggests that giving LLM specific roles induces more original responses. This experiment reveals the weakness of LLM's lack of originality in its default situation.

4.5 Results of creativity under collaboration

In reality, creative activities can be accomplished through collaboration and discussion among multiple individuals. The literature indicates that the process of creative collaboration can increase the innovativeness of the outcomes[58, 59]. Inspired by this, we believe that the results produced through the collaboration with LLMs have stronger creativity than those generated by a single LLM.

Based on the above analysis, this section explores the use of multiple agents engaging in multi-round discussions on questions from the dataset, ultimately producing a joint final answer. After the previous LLM provides an answer, the subsequent LLM will use that answer as inspiration to give its own response. Once a predetermined number of rounds is reached, the final result is presented. In our experiment, using GPT-3.5 as the base model, we explored the changes in scores under different creativity criteria when the number of LLMs is 2 and 3 (we call it agent) and the number of rounds is 2 and 3. We compared these scores with the creativity scores obtained under default conditions.

As shown in Fig. 5, we presented scatter plots of the creativity scores under different criteria, varying by the number of rounds and agents. The area of each scatter

(a)

Task name	Instructive prompt	CoT prompt
Unusual uses task	–	–
Consequences task	√ *	√ ***
Just suppose task	√ ***	√ ***
Situation task	√ ***	√ ***
Common problem task	√ *	–
Improvement task	√ *	–
Imaginative stories task	√ ***	–

Criteria	Instructive prompt	CoT prompt
Fluency	√ ***	√ *
Flexibility	√ ***	–
Originality	√ ***	–
Elaboration	–	√ *

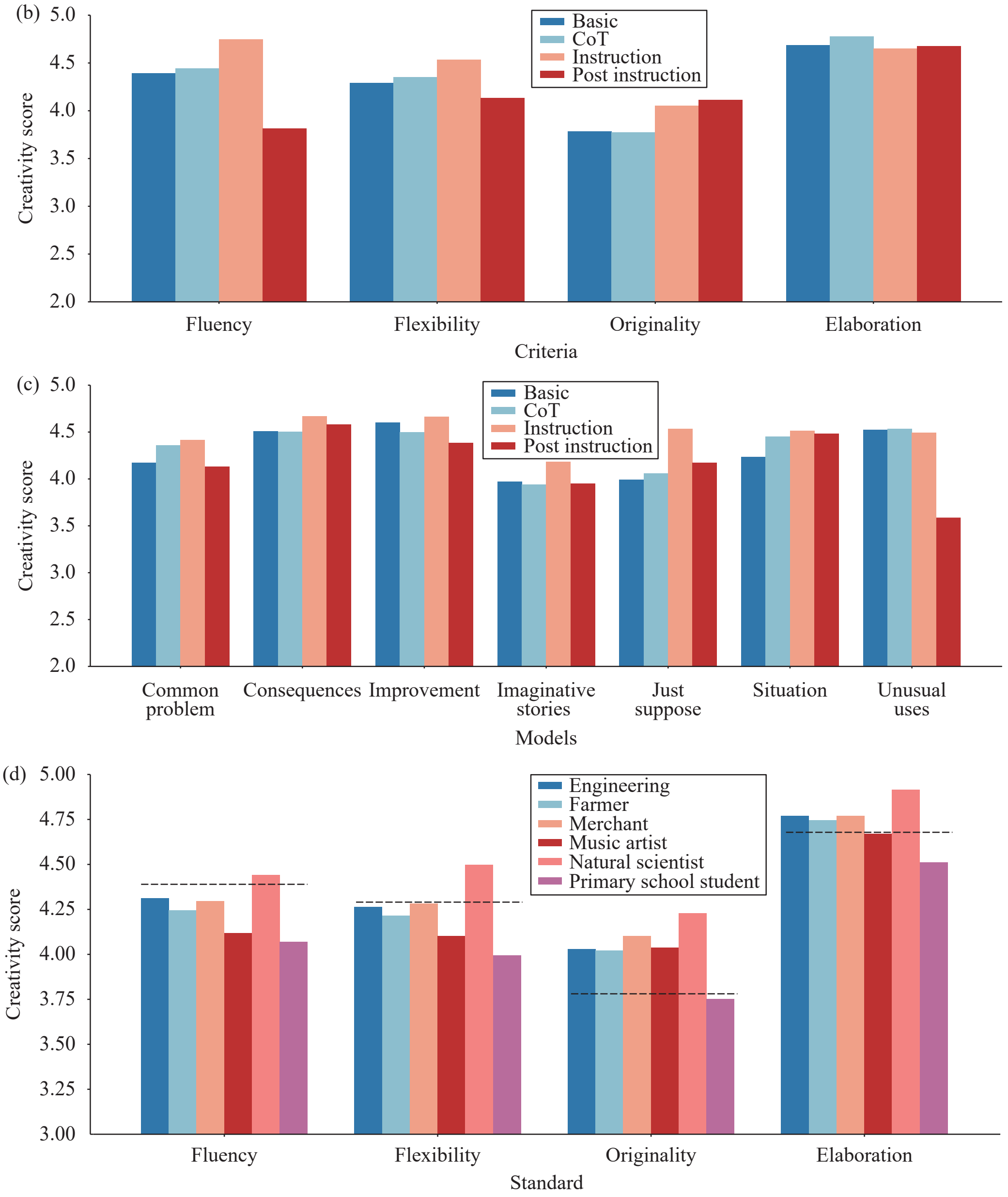


Fig. 4 Effects of prompt types and role-play settings on LLM creativity. (a) Prompt type impact across tasks and criteria; "√" denotes significant improvement, "–" denotes no effect. Significance is marked ($***$ for $p < 0.000\,1$, $*$ for $p < 0.05$; Wilcoxon signed-rank test). (b) Creativity scores across criteria by prompt type. (c) Creativity across tasks by prompt type. (d) Role-based performance across all tasks; the horizontal line indicates baseline without role-play. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

point represents the level of creativity. From the results, we can see some interesting findings: First, when the number of rounds is one, an increase in the number of agents leads to a decrease in the level of creativity across

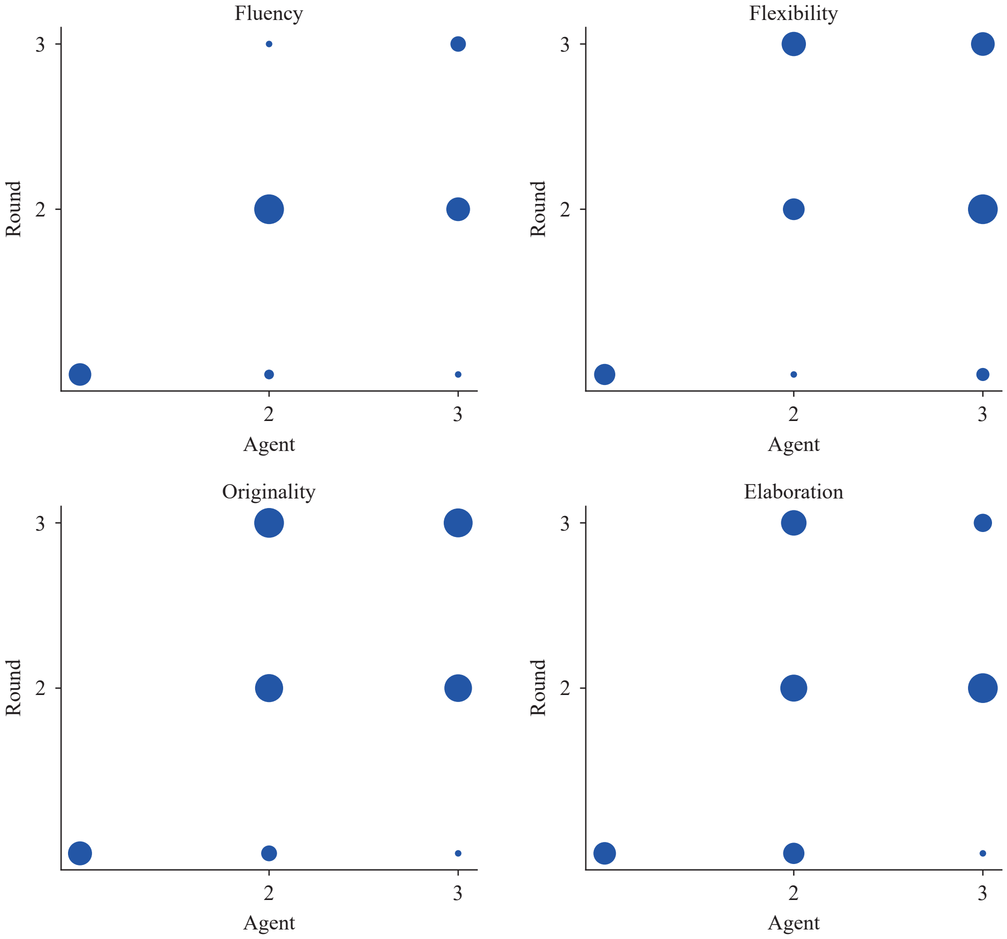


Fig. 5 Effect of agent collaboration on creativity scores

the four criteria. This might be due to the lack of multiple reviews of the answers by the same agent in a single round of interaction, leading to the answers of later-ranked agents constantly negating previous answers. Additionally, in the cases of originality, flexibility, and elaboration, an increase in both rounds and agents enhances the level of creativity, with the most significant improvement observed in originality. This supports the conclusion that collaboration can enhance creativity and is consistent with human behavior. Lastly, there are some exceptions to the above conclusion, such as a decrease in fluency when there are two agents and three rounds. This could be due to excessive discussion that makes the answers overly concise.

4.6 Investigation of the relationship between LLM′s creativity and its personality traits

4.6.1 Psychology scales

In this experiment, we explored the relationships between personality traits and creative performance of some large models, using some public psychological scales and related literature.

We use the situational test of emotion management (STEM) for the assessment of emotional intelligence[60]. STEM evaluates an individual′s ability to manage emotions in various situations. It is based on the concept of emotional intelligence, which involves recognizing, understanding, and managing one′s own emotions and those of others. The test typically presents a series of hypothetical scenarios to the participants. Each scenario is designed to assess different aspects of emotional intelligence, such as emotional awareness and regulation. The participants are asked how they would respond to each situation. Their responses are then analyzed to determine their EI levels. STEM is used in various settings, including organizational training, psychological research, and personal development. It helps in identifying areas where emotional intelligence can be improved, which is valuable in both personal and professional contexts.

We use the Toronto empathy questionnaire (TEQ) for assessing LLM′s empathy level[61]. The TEQ was developed by researchers at the University of Toronto. It is grounded in the idea that empathy is a multi-dimensional construct, involving both cognitive and affective elements. The questionnaire consists of 16 items, each rated on a 5-point Likert scale. These items are designed to measure the respondent's emotional and cognitive responses to the experiences and feelings of others. The TEQ is used in various fields, including psychological research, clinical settings, and social science studies. It helps in understanding how individuals emotionally connect with others, which can be important in contexts such as therapy, counselling, and social work.

We use generalized self-efficacy scale[62] to assess LLM's self-efficacy. The scale was developed by Ralf Schwarzer and Matthias Jerusalem in 1995. This is part of a larger body of research on self-efficacy and psychological well-being. The generalized self-efficacy scale is a short survey consisting of 10 items. The respondents rate each item on a scale, typically from 1 to 4, where higher scores indicate greater self-efficacy. Unlike scales that measure task-specific or situation-specific self-efficacy, this scale assesses a general sense of personal competence to deal effectively with a variety of stressful situations. It is widely used in psychological research, clinical psychology, and health psychology. It's also utilized in organizational and educational settings to understand and enhance individuals' beliefs in their own capabilities. The generalized self-efficacy scale has been validated in numerous studies across different cultures and is known for its reliability and construct validity.

Finally, we applied the classic big five inventory (BFI) test to LLMs[63]. The big five personality traits include openness, conscientiousness, extraversion, agreeableness, and neuroticism (often abbreviated as OCEAN). These traits represent a broad range of human personality characteristics and are believed to be universal. The BFI typically comprises short statements that respondents rate based on how accurately they reflect their own behavior or personality traits. The BFI is valued for its balance between brevity and comprehensive coverage of the five-factor model. This demonstrates good reliability and validity, making it a trusted tool in personality assessment.

4.6.2 Investigation results

We have mentioned that, creativity evolves from a combination of individual processes such as cognitive, affective, behavioral, and contextual factors. In this section, we subject LLMs to a series of psychometric tests traditionally used to assess human personality traits. Our aim is to explore whether, akin to humans, there is a correlation between various personality factors and the creative capabilities of these advanced computational systems.

In our experiment, we selected eight personality traits: emotional intelligence, empathy, self-efficacy, openness, conscientiousness, extraversion, agreeableness, and neuroticism. The latter five are the classic big five personality traits. The meta-analytic literature[28, 29] in the field of psychology suggests that each of these eight traits correlates with levels of creativity.

As shown in Table 2, we conducted experiments on LLMs and reported the correlations between the mentioned personality traits and creativity. We chose the Kendall τ and Spearman ρ as the correlation coefficients and performed hypothesis testing. The experimental results indicate that the levels of emotional intelligence, empathy, conscientiousness, extraversion, and neuroticism in large language models have a significant positive correlation with creativity levels, whereas agreeableness shows a significant negative correlation. Apart from agreeableness and openness, the influence of the remaining personality traits on creativity in large models is consistent with human performance.

4.7 Task domain generality evaluation

From the perspective of domain generality[64], creativity is viewed as a transferable skill that can be applied across different fields or domains (e.g., arts, sciences, business). The selected tasks – unusual uses, consequences, just suppose, situation, common problem, improvement, and imaginative stories – are domain-general in nature, meaning they assess creativity without being tied to any specific subject matter or expertise. These tasks are content-neutral and focus on core cognitive processes such as idea generation, problem-solving, and ima-

Table 2 Reports of Kendall's τ, Spearman's ρ, and p-values for correlations between selected personality traits and LLM creativity. Significant results are highlighted in bold.

		Emotional intelligence	Empathy	Self-efficacy	Openness
Kendall τ	Correlation coefficient	**0.382 5**	**0.382 5**	**0.440 1**	0.032 9
	p-value	< 0.000 1	< 0.000 1	< 0.000 1	0.605 4
Spearman ρ	Correlation coefficient	**0.502 6**	**0.502 6**	**0.561 3**	0.052 6
	p-value	< 0.000 1	< 0.000 1	< 0.000 1	0.536 9
		Conscientiousness	Extraversion	Agreeableness	Neuroticism
Kendall τ	Correlation coefficient	**0.369 1**	**0.263 6**	**–0.370 0**	**0.253 3**
	p-value	< 0.000 1	< 0.000 1	< 0.000 1	< 0.000 1
Spearman ρ	Correlation coefficient	**0.489 7**	**0.339 4**	**–0.496 7**	**0.345 5**
	p-value	< 0.000 1	< 0.000 1	< 0.000 1	< 0.000 1

gination, which are fundamental to creativity across all domains. Because they are not specialized in a particular field (e.g., only artistic creativity or scientific innovation), they can assess the general creative potential of an individual that applies across different contexts. This domain-general approach ensures that the tasks are broadly applicable and can capture creative abilities that are relevant in multiple disciplines, making the assessment more holistic and inclusive. Thus, their domain generality contributes to the completeness of the task set, as it ensures that the creativity being measured is not limited to a single context but represents a more universal capability.

To validate the domain generality under our experimental settings, we computed Cronbach′s Alpha for each creativity criterion across the tasks, with results showing high internal consistency ($\alpha > 0.8$ for all criteria), indicating that the tasks are sufficient for measuring domain-general creativity, as shown in Fig. 6. The strong correlations between tasks across different models further support the idea that the task set captures a shared underlying construct, ensuring that the evaluation provides a holistic view of the models′ creative capabilities.

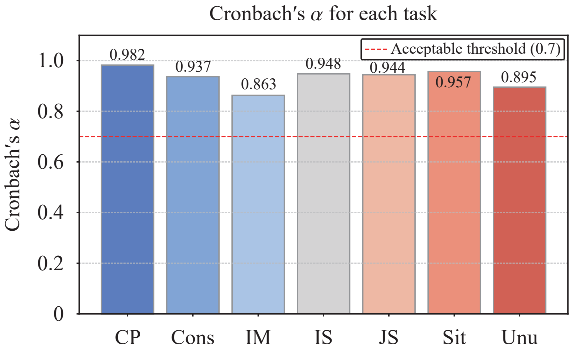


Fig. 6 Bar chart showing Cronbach′s α values for each task, with the acceptable threshold (0.8) indicated by a red dashed line. All tasks exceed this threshold. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

What′s more, we do inter-task correlation analysis between the scores of the 7 tasks across the 6 models. As shown in Fig. 7, results suggest that all tasks measure a shared, domain-general aspect of creativity. In summary, the TTCT tasks can be justified as sufficiently broad and representative of core creative processes such as divergent thinking, originality, and flexibility. However, they are not exhaustive, and their sufficiency largely depends on the specific definition of creativity being employed.

4.8 Evaluating reliability of GPT-4 as a judge

As in the experiments above, GPT-3.5 generally achieves high creativity performance across multiple tasks, though specific tasks, such as the consequences and situation tasks under certain criteria, do not rank it as the top performer. Despite GPT-4 and GPT-3.5 being developed by OpenAI and likely sharing similar model structures, we posit that GPT-4 does not exhibit significant preference when serving as a judge. To validate that GPT-4 does not favor GPT-3.5 responses, we conducted additional comparison tests. We select LLaMA-3-8b from the LLaMA model family, which is close to GPT-3.5 in creativity performance. We used it as an examiner to score the LLM responses in our experiment with the same system prompt. This allowed us to obtain comparative scores for the same data from two different judges.

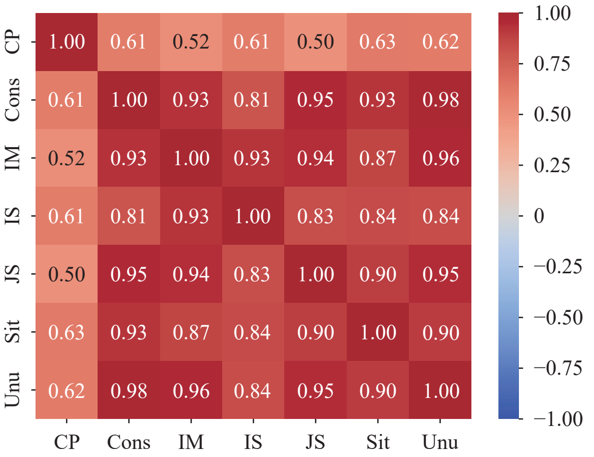


Fig. 7 The heatmap displays Pearson correlation coefficients between creativity scores across task pairs, as evaluated by the same judge. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

To assess the consistency between judges, we applied the intra-class correlation coefficient (ICC), a statistical measure evaluating agreement across different raters, specifically using the ICC(3,1) model to gauge single-rater, absolute agreement. This method helps determine if the two judges provide similar ratings under identical conditions. Additionally, we calculated conventional correlation coefficients to further validate consistency. As shown in Fig. 8, the results show a high level of agreement, with an ICC of 0.99 between GPT-4 and LLaMA-3-8b, indicating excellent consistency. Pearson, Spearman, and Kendall correlations of 0.64, 0.61, and 0.46, respectively, indicate moderate correlation in overall rankings, although there are variations in how the models rank specific tasks. These findings suggest that while GPT-4 and LLaMA-3-8b are consistent in their ratings, they may differ in interpreting and prioritizing certain task aspects.

4.9 Model-human agreement evaluation

To confirm that the assessment methods based on LLMs are overall reasonable and consistent with human judgement, we sample the responses generated by these models and hire humans to evaluate them. We presented the answer pairs generated by the LLMs to 20 native English-speaking participants globally (10 male) recrui-

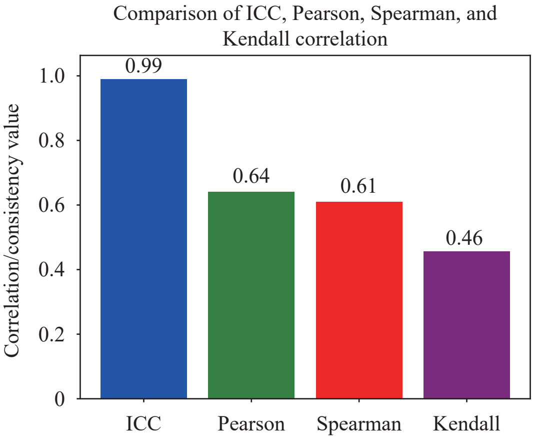


Fig. 8 Comparison of ICC, Pearson, Spearman, and Kendall correlation values, demonstrating the consistency and agreement between GPT-4 and LLaMA-3-8b as judges across tasks. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

ted from Prolific (https://www.prolific.co/), and paid each participant £15. The average reward per hour for the participants was £14.59. The average participant age was 32.9 ± 20.1. In the experiment, we sampled seven tasks, resulting in 84 pairs of questions and answers, which means there are 84 trials. These pairs consist of answers from different models to the same question within the same task, and are presented to the participants.

On each trial of the task, participants were asked to make a binary decision about which of the two answers is more creative according to the given criteria. The participants also have the option to choose that there is no significant difference in creativity between the two responses. A progress bar at the top of the screen indicated to participants how many trials they had completed and had remained to complete. After the final human evaluation data are obtained, we calculate the consistency between the human assessment results and those of the LLMs for the overall score and each criterion.

We use Kendall′s coefficient and Spearman′s coefficient for this calculation. Since the participants′ data are based on relative win-loss relationships, we need to preprocess the human evaluation results. For the tie results, we convert the human assessment results to the average score of two answers evaluated by LLM; for non-tie results, we assign the higher score evaluated by the LLM to the winning response in the human results. The results are shown in Table 3.

These moderate correlations indicate statistically significant alignment between GPT-4′s evaluations and human judgments, despite not achieving perfect consensus. For subjective and multifaceted constructs such as creativity, moderate correlations are notable, as complete agreement among human raters themselves is often difficult to achieve. This alignment supports the view that GPT-4 captures key aspects of human evaluation while acknowledging that further refinement is needed to enhance the alignment.

5 Conclusions and discussions

In this article, we have presented a framework to assess and understand the creativity of LLMs. The core of this framework consists of 7 tasks and LLM-based evaluation protocol that can be used to assess LLM′s creativity along four criteria. The proposed framework can be used to assess the creative performance of LLMs from multiple dimensions, while also exploring the factors that influence the creativity of these models and the relationships with other model characteristics. To illustrate the use and usefulness of our framework, we constructed a dataset containing 700 questions that encompass various types of tasks measuring divergent thinking.

Through our further analysis and experiments, we demonstrated that the creativity of LLMs is significantly influenced by the type of model architecture, the type of prompts it receives, and the model′s system prompts. At the same time, we also revealed a correlation between the levels of creativity of LLMs and their personality traits. This work is beneficial for our deeper understanding of the representations of LLMs and trying to establish a bridge between artificial intelligence models and human cognitive models.

Although we propose an effective framework for measuring the creativity of LLMs, it still has some limitations that need to be addressed by future work. First, LLMs use text as both input and output, which allows them to borrow from psychological methods of creativity assessment such as TTCT, which uses a verbal task with verbal stimuli. However, with the rapid development of AI models, those accepting multimodal inputs are emerging[65], which we call large multi-modal model (LMM). Designing a variety of tasks beyond verbal question-and-

Table 3 Correlations and corresponding p-values between GPT-4-based and human-based evaluations across creativity criteria (fluency, flexibility, originality, elaboration, and overall)

		Fluency	Flexibility	Originality	Elaboration	Overall
Kendall τ	Correlation coefficient	0.588 9	0.578 2	0.556 7	0.451 2	0.499 6
	p-value	$< 0.000\,1$	$< 0.000\,1$	$< 0.000\,1$	$< 0.000\,1$	$< 0.000\,1$
Spearman ρ	Correlation coefficient	0.621 4	0.614 9	0.592 3	0.468 5	0.556 4
	p-value	$< 0.000\,1$	$< 0.000\,1$	$< 0.000\,1$	$< 0.000\,1$	$< 0.000\,1$

answer formats for assessing the creativity of these LMM is an important direction for future research.

Second, LLMs are not the only generative models being capable of expressing creativity; there are also image generation models that are based on diffusion models and models for generating music[66, 67], in other words, can generate multi-modal outputs. How to assess the content produced by these other types of models to measure their level of creativity is also a question worth considering. In addition, the power of LLM allows developers to use it to develop a wide variety of plug-ins, integrate it with external programs or software, and even construct an agent system[68], and the creativity in these cases is bound to be different and needs to be investigated.

Moreover, while this study demonstrates the overall alignment between GPT-4 and human evaluations of creativity, it does not delve into criterion-specific correlations (e.g., fluency, flexibility, originality, elaboration). By expanding the dataset, leveraging separate evaluations for each creativity criterion, and employing advanced statistical techniques such as criterion-specific intraclass correlation coefficients, finer-grained alignment studies can illuminate the strengths and limitations of LLM-based evaluation frameworks.

Last, we believe that the creativity exhibited by LLMs is only an outcome-oriented interpretation. Whether AI models possess true creativity from a human cognitive perspective remains an open question in the field of artificial intelligence. LLM′s expression of creativity is likely to be an imitation of human creativity through a large amount of learning. Understanding the creativity of LLMs is also beneficial for uncovering the inner secrets of the model “black box”, and for a deeper understanding of the nature of intelligence and cognition. Although analysing the nature of creativity is difficult, our analysis and evaluation of LLM creativity performance is fundamental to study of the kernel of creativity.

Appendix

A.1 Example prompts

Likert scale scoring.

You are an expert of psychology. Your objective is to assess the subject′s creativity through their answers to some question/answering task related to divergent thinking. You will be given a question-answer pair. Your task is to score the answer.

You should rate the answer on five metrics. For all five metrics, assign a score between 1 and 5, with 5 being the highest. Five metrics are:

1) Fluency. Fluency refers to the ability to generate a large quantity of ideas or solutions to a given problem. This measure isn't concerned with the quality or uniqueness of the ideas, but rather the sheer volume. The more ideas one can produce, the higher the fluency is.

2) Flexibility. Flexibility is the capacity to shift one′s thinking and to produce a wide range of ideas from different categories or perspectives. It involves being able to think outside of the box and to switch from one type of idea to another.

3) Originality. Originality refers to the ability to come up with unique or novel ideas that differ from the norm. It′s not just about producing many ideas (fluency), but also about producing ideas that are different from what others might typically think of.

4) Elaboration. Elaboration is the ability to expand upon or add detail to ideas. It involves taking a simple idea and building upon it, adding complexity and depth. Elaboration isn't just about creating more, but about deepening what is there.

5) Finally, you will provide an overall score between 1 and 5, with 5 being the highest.

You should only give the score, format like: Fluency: 3

Question: {Question} Answer: {Answer}

Instructive prompts (unusual uses task as the example):

Unusual uses task.

The purpose of this task is to measure your ability to come up with creative and unique uses for everyday objects. We're looking for out-of-the-box thinking here.

You will be presented with a common object, and your task is to suggest as many unusual, innovative, or non-traditional uses for the object as you can think of. Please remember, the goal is not to think of the most common or typical uses, but to try and imagine unique or unusual ways the object could be used.

Here are the objects: {Objects}

Collaboration prompts.

These are answers to the question from other agents:

One agent solution: {Answers}

…

One agent solution: {Answers}

Using the answers from other agents as reference and inspiration, can you give an updated answer? Make sure to give your answer at the end of the response.

Question: {Question}

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Nos. U22A2028, 61925208, 62102399, 62302478, 62302483, 62222214, 62372436, 62302482 and 62302480), CAS Project for Young Scientists in Basic Research, China (No. YSBR-029), Youth Innovation Promotion Association CAS and Xplore Prize, China.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Z. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4, [Online], Available: https://arxiv.org/abs/2303.12712, 2023.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Y. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Y. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. H. Lu, Y. N. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. X. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. X. Xu, Z. Yan, I. Zarov, Y. C. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom. Llama 2: Open foundation and fine-tuned chat models, [Online], Available: https://arxiv.org/abs/2307.09288, 2023.

[3] Y. H. Wu, A. Q. Jiang, W. D. Li, M. N. Rabe, C. Staats, M. Jamnik, C. Szegedy. Autoformalization with large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 2344, 2022.

[4] T. R. Laskar, M. S. Bari, M. Rahman, A. H. Bhuiyan, S. Joty, J. Huang. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Proceedings of Findings of the Association for Computational Linguistics*, Toronto, Canada, pp. 431–469, 2023. DOI: 10.18653/v1/2023.findings-acl.29.

[5] J. Devlin, M. W. Chang, L. Kenton, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[6] G. Di Fede, D. Rocchesso, S. P. Dow, S. Andolina. The idea machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*, Venice, Italy, pp. 623–627, 2022. DOI: 10.1145/3527927.3535197.

[7] M. Elzohbi, R. Zhao. Creative data generation: A review focusing on text and poetry. In *Proceedings of the 14th International Conference on Computational Creativity*, Ontario, Canada, pp. 29–38, 2023.

[8] Z. Zhao, S. Song, B. Duah, J. Macbeth, S. Carter, M. P. Van, N. S. Bravo, M. Klenk, K. Sick, A. L. S. Filipowicz. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, NewYork, USA, pp. 368–370, 2023. DOI: 10.1145/3591196.3596612.

[9] Y. J. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, T. Hubert, P. Choy, C. De Masson d'Autume, I. Babuschkin, X. Y. Chen, P. S. Huang, J. Welbl, S. Gowal, A. Cherepanov, J. Molloy, D. J. Mankowitz, E. Sutherland Robson, P. Kohli, N. De Freitas, K. Kavukcuoglu, O. Vinyals. Competition-level code generation with AlphaCode. *Science*, vol. 378, no. 6624, pp. 1092–1097, 2022. DOI: 10.1126/science.abq1158.

[10] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, vol. 103, Article number 102274, 2023. DOI: 10.1016/j.lindif.2023.102274.

[11] S. Y. Li, W. J. Yu, T. P. Gu, C. Z. Lin, Q. Wang, C. Qian, C. C. Loy, Z. W. Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 11050–11059, 2022. DOI: 10.1109/CVPR52688.2022.01077.

[12] B. Banar, S. Colton. A systematic evaluation of GPT-2-based music generation. In *Proceedings of the 11th International Conference on Artificial Intelligence in Music, Sound, Art and Design*, Madrid, Spain, pp. 19–35, 2022. DOI: 10.1007/978-3-031-03789-4_2.

[13] Y. R. Liu, S. Chen, H. C. Cheng, M. X. Yu, X. Ran, A. Mo, Y. L. Tang, Y. Huang. How AI processing delays foster creativity: Exploring research question Co-creation with an LLM-based agent. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, Article number 17, 2024. DOI: 10.1145/3613904.3642698.

[14] H. Shin, S. Choi, J. Y. Cho, S. Admoni, H. Lim, T. Kim, H. Hong, M. Lee, Kim, J. Towards an evaluation of LLM-generated inspiration by developing and validating inspiration scale. In *Proceedings of the 1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2024.

[15] B. R. Anderson, J. H. Shah, M. Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, Chicago, USA, pp. 413–425, 2024. DOI: 10.1145/3635636.3656204.

[16] M. A. Runco, G. J. Jaeger. The standard definition of cre-

ativity. *Creativity Research Journal*, vol. 24, no. 1, pp. 92–96, 2012. DOI: 10.1080/10400419.2012.650092.

[17] T. Chakraborty, S. Masud. Judging the creative prowess of AI. *Nature Machine Intelligence*, vol. 5, no. 6, Article number 558, 2023. DOI: 10.1038/s42256-023-00664-y.

[18] E. P. Torrance. *Torrance Test of Creative Thinking: Directions Manual and Scoring Guide*, Lexington, USA: Personnel Press, 1966.

[19] B. Barbot, R. Reiter-Palmon. Creativity assessment: Pitfalls, solutions, and standards. *Psychology of Aesthetics, Creativity, and the Arts*, vol. 13, no. 2, pp. 131–132, 2019. DOI: 10.1037/aca0000251.

[20] R. J. Sternberg, E. L. Grigorenko. Guilford's structure of intellect model and model of creativity: Contributions and limitations. *Creativity Research Journal*, vol. 13, no. 3–4, pp. 309–316, 2001. DOI: 10.1207/S15326934CRJ1334_08.

[21] S. Acar, M. A. Runco. Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts*, vol. 13, no. 2, pp. 153–158, 2019. DOI: 10.1037/aca0000231.

[22] G. M. Cseh, K. K. Jeffries. A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, vol. 13, no. 2, pp. 159–166, 2019. DOI: 10.1037/aca0000220.

[23] J. C. Kaufman. Self-assessments of creativity: Not ideal, but better than you think. *Psychology of Aesthetics, Creativity, and the Arts*, vol. 13, no. 2, pp. 187–192, 2019. DOI: 10.1037/aca0000217.

[24] B. Barbot, R. W. Hass, R. Reiter-Palmon. Creativity assessment in psychological research: (Re) setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, vol. 13, no. 2, pp. 233–240, 2019. DOI: 10.1037/aca0000233.

[25] K. H. Kim. The APA 2009 division 10 debate: Are the Torrance tests of creative thinking still relevant in the 21st century? *Psychology of Aesthetics, Creativity, and the Arts*, vol. 5, no. 4, pp. 302–308, 2011. DOI: 10.1037/a0021917.

[26] J. A. Plucker. Is the proof in the pudding? Reanalyses of torrance's (1958 to present) longitudinal data. *Longitudinal Studies of Creativity*, M. A. Runco, Ed., New York, USA: Routledge, pp. 103–114, 1999. DOI: 10.4324/9780203063330.

[27] K. H. Kim. Can we trust creativity tests? A review of the Torrance tests of creative thinking (TTCT). *Creativity Research Journal*, vol. 18, no. 1, pp. 3–14, 2006. DOI: 10.1207/s15326934crj1801_2.

[28] S. da Costa, D. Páez, F. Sánchez, M. Garaigordobil, S. Gondim. Personal factors of creativity: A second order meta-analysis. *Journal of Work and Organizational Psychology*, vol. 31, no. 3, pp. 165–173, 2015. DOI: 10.1016/j.rpto.2015.06.002.

[29] H. H. Ma. The effect size of variables associated with creativity: A meta-analysis. *Creativity Research Journal*, vol. 21, no. 1, pp. 30–42, 2009. DOI: 10.1080/10400410802633400.

[30] Y. P. Chang, X. Wang, J. D. Wang, Y. Wu, L. Y. Yang, K. J. Zhu, H. Chen, X. Y. Yi, C. X. Wang, Y. D. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, Article number 39, 2024. DOI: 10.1145/3641289.

[31] R. Shiffrin, M. Mitchell. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, no. 10, Article number e2300963120, 2023. DOI: 10.1073/pnas.2300963120.

[32] G. Franceschelli, M. Musolesi. On the creativity of large language models, [Online], Available: https://arxiv.org/abs/2304.00008, 2023.

[33] D. Summers-Stay, S. Lukin, C. Voss. Brainstorm, then select: A generative language model improves its creativity score. In *Proceedings of AAAI Workshop on Creative AI Across Modalities*, 2023.

[34] C. Stevenson, I. Smal, M. Baas, R. P. P. P. Grasman, H. L. J. van der Maas. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*, Bozen-Bolzano, Italy, pp. 164–168, 2022.

[35] S. A. Naeini, R. Saqur, M. Saeidi, J. Giorgi, B. Taati. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 246, 2023.

[36] E. E. Guzik, C. Byrge, C. Gilde. The originality of machines: AI takes the Torrance test. *Journal of Creativity*, vol. 33, no. 3, Article number 100065, 2023. DOI: 10.1016/j.yjoc.2023.100065.

[37] T. Chakrabarty, P. Laban, D. Agarwal, S. Muresan, C. S. Wu. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, Article number 30, 2024. DOI: 10.1145/3613904.3642731.

[38] T. I. Lubart. Models of the creative process: Past, present and future. *Creativity Research Journal*, vol. 13, no. 3–4, pp. 295–308, 2001. DOI: 10.1207/S15326934CRJ1334_07.

[39] M. Batey. The measurement of creativity: From definitional consensus to the introduction of a new heuristic framework. *Creativity Research Journal*, vol. 24, no. 1, pp. 55–65, 2012. DOI: 10.1080/10400419.2012.649181.

[40] D. Piffer. Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, vol. 7, no. 3, pp. 258–264, 2012. DOI: 10.1016/j.tsc.2012.04.009.

[41] R. E. Beaty, D. R. Johnson. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, vol. 53, no. 2, pp. 757–780, 2021. DOI: 10.3758/s13428-020-01453-w.

[42] S. Acar, K. Berthiaume, K. Grajzel, D. Dumas, C. Flemister, P. Organisciak. Applying automated originality scoring to the verbal form of Torrance tests of creative thinking. *Gifted Child Quarterly*, vol. 67, no. 1, pp. 3–17, 2023. DOI: 10.1177/00169862211061874.

[43] Y. S. Bai, J. H. Ying, Y. X. Cao, X. Lv, Y. Z. He, X. Z. Wang, J. F. Yu, K. S. Zeng, Y. J. Xiao, H. Z. Lyu, J. Y. Zhang, J. Z. Li, L. Hou. Benchmarking foundation models with language-model-as-an-examiner. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, 2023.

[44] L. M. Zheng, W. L. Chiang, Y. Sheng, S. Y. Zhuang, Z. H. Wu, Y. H. Zhuang, Z. Lin, Z. H. Li, D. C. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article num-

ber 2020, 2023.

[45] Y. J. Bang, S. Cahyawijaya, N. Lee, W. L. Dai, D. Su, B. Wilie, H. Lovenia, Z. W. Ji, T. Z. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Nusa Dua, Indonesia, 2023. DOI: 10.18653/v1/2023.ijcnlp-main.45.

[46] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, et al. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[47] C. M. Chan, W. Z. Chen, Y. S. Su, J. X. Yu, W. Xue, S. H. Zhang, J. Fu, Z. Y. Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[48] W. J. Zhong, R. X. Cui, Y. D. Guo, Y. B. Liang, S. Lu, Y. L. Wang, A. Saied, W. Z. Chen, N. Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Proceedings of Findings of the Association for Computational Linguistics*, Mexico City, Mexico, pp. 2299–2314, 2024. DOI: 10.18653/v1/2024.findings-naacl.149.

[49] Y. Dubois, X. C. Li, R. Taori, T. Y. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, T. B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1308, 2024.

[50] Y. Liu, D. Iter, Y. C. Xu, S. H. Wang, R. C. Xu, C. G. Zhu. G-Eval: NLG evaluation using Gpt-4 with better human alignment. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 2511–2522, 2023. DOI: 10.18653/v1/2023.emnlp-main.153.

[51] D. Demszky, D. Y. Yang, D. S. Yeager, C. J. Bryan, M. Clapper, S. Chandhok, J. C. Eichstaedt, C. Hecht, J. Jamieson, M. Johnson, M. Jones, D. Krettek-Cobb, L. Lai, N. Jonesmitchell, D. C. Ong, C. S. Dweck, J. J. Gross, J. W. Pennebaker. Using large language models in psychology. *Nature Reviews Psychology*, vol. 2, no. 11, pp. 688–701, 2023. DOI: 10.1038/S44159-023-00241-5.

[52] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. W. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. DOI: 10.18653/v1/2020.emnlp-demos.6.

[53] J. Z. Bai, S. Bai, Y. F. Chu, Z. Y. Cui, K. Dang, X. D. Deng, Y. Fan, W. B. Ge, Y. Han, F. Huang, B. Y. Hui, L. Ji, M. Li, J. Y. Lin, R. J. Lin, D. H. Liu, G. Liu, C. Q. Lu, K. M. Lu, J. X. Ma, R. Men, X. Z. Ren, X. C. Ren, C. Q. Tan, S. N. Tan, J. H. Tu, P. Wang, S. J. Wang, W. Wang, S. G. Wu, B. F. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. S. Yang, Y. Yao, B. W. Yu, H. Y. Yuan, Z. Yuan, J. W. Zhang, X. X. Zhang, Y. C. Zhang, Z. R. Zhang, C. Zhou, J. R. Zhou, X. H. Zhou, T. H. Zhu. Qwen technical report, [Online], Available: https://arxiv.org/abs/2309.16609, 2023.

[54] S. Lin, J. Hilton, O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, pp. 3214–3252, 2022. DOI: 10.18653/v1/2022.acl-long.229.

[55] T. Xiang, L. Z. Li, W. Y. Li, M. B. Bai, L. Wei, B. W. Wang, N. Garcia. CARE-MI: Chinese benchmark for misinformation evaluation in maternity and infant care. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

[56] Q. T. Xu, F. L. Hong, B. Li, C. R. Hu, Z. Y. Chen, J. Zhang. On the tool manipulation capability of open-source large language models. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

[57] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 22199–22213, 2022.

[58] P. B. Paulus, M. Dzindolet, N. W. Kohn. Collaborative creativity-group creativity and team innovation. *Handbook of Organizational Creativity*, M. D. Mumford, Ed., Amsterdam, The Netherlands: Elsevier, pp. 327–357, 2012. DOI: 10.1016/B978-0-12-374714-3.00014-8.

[59] M. S. Barrett, A. Creech, K. Zhukov. Creative collaboration and collaborative creativity: A systematic literature review. *Frontiers in Psychology*, vol. 12, Article number 713445, 2021. DOI: 10.3389/fpsyg.2021.713445.

[60] C. MacCann, R. D. Roberts. New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, vol. 8, no. 4, pp. 540–551, 2008. DOI: 10.1037/a0012746.

[61] R. N. Spreng, M. C. McKinnon, R. A. Mar, B. Levine. The Toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, vol. 91, no. 1, pp. 62–71, 2009. DOI: 10.1080/00223890802484381.

[62] R. Schwarzer, M. Jerusalem. Generalized self-efficacy scale. *Measures in Health Psychology: A User's Portfolio*, J. Weinman, S. Wright, M. Johnston, Eds., Windsor, UK: NFER-NELSON, pp. 35–37, 1995.

[63] O. P. John, S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2nd ed., L. A. Pervin, O. P. John, Eds., New York, USA: Guilford Press, 1999.

[64] M. H. Qian, J. A. Plucker, X. D. Yang. Is creativity domain specific or domain general? Evidence from multilevel explanatory item response theory models. *Thinking Skills and Creativity*, vol. 33, Article number 100571, 2019. DOI: 10.1016/j.tsc.2019.100571.

[65] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. H. Yu, W. L. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence. PaLM-E: An embodied multimodal language model, [Online], Available: https://arxiv.org/abs/2303.03378, 2023.

[66] F. A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023. DOI: 10.1109/TPAMI.2023.3261988.

[67] F. Carnovalini, A. Rodà. Computational creativity and music generation systems: An introduction to the state of

the art. *Frontiers in Artificial Intelligence*, vol. 3, Article number 14, 2020. DOI: 10.3389/frai.2020.00014.

[68] L. Wang, C. Ma, X. Y. Feng, Z. Y. Zhang, H. Yang, J. S. Zhang, Z. Y. Chen, J. K. Tang, X. Chen, Y. K. Lin, W. X. Zhao, Z. W. Wei, J. R. Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, vol. 18, no. 6, Article number 186345, 2024. DOI: 10.1007/s11704-024-40231-1.

Yunpu Zhao received the B. Sc. degree in computer science from Wuhan University, China in 2022. He is a Ph. D. degree candidate at the University of Science and Technology of China, China.

His research interests include deep learning and multimodal large language models.

E-mail: zyp351791@mail.ustc.edu.cn

ORCID iD: 0000-0001-7747-7040

Rui Zhang received the Ph. D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, China in 2019. She is an associate professor in State Key Lab (SKL) of Processors at the Institute of Computing Technology, Chinese Academy of Sciences, China.

Her research interests include deep learning and multimodal large language model.

E-mail: zhangrui@ict.ac.cn (Corresponding author)

ORCID iD: 0000-0001-8691-8549

Wenyi Li received the B. Eng. degree in computer science and technology from the University of Chinese Academy of Sciences, China in 2023. He is currently a Ph. D. degree candidate at the Institute of Software, Chinese Academy of Sciences (ISCAS), China.

His research interests include computer vision, code generation, and large language model reasoning.

E-mail: liwenyi2023@iscas.ac.cn

Ling Li received the B. Sc. degree in computer science and technology from Wuhan University, China in 2004, and the Ph. D. degree in computer architecture from the Institute of Computing Technology, Chinese Academy of Sciences, China in 2009. She is currently a professor at the Institute of Software, Chinese Academy of Sciences, China. She has authored four books (including *AI Computing Systems*) and over 70 papers on journals and conferences (including TIP, TC, TPDS, ICML, NeurIPS, AAAI, ISCA, MICRO).

Her research interest is intelligent computing.

E-mail: liling@iscas.ac.cn