Do LLMs Agree on the Creativity Evaluation of Alternative Uses?

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, Talles Medeiros
School of Computing and Mathematical Sciences
University of Leicester
Leicester, UK
{asmar3,fabricio.goes,marco.volpe,thm14}@leicester.ac.uk

Abstract

This paper investigates whether large language models (LLMs) exhibit agreement in assessing creativity in responses to the Alternative Uses Test (AUT). While LLMs are increasingly used to evaluate creative content, previous studies have primarily examined a single model assessing either its own outputs or human-generated responses. Here, we explore whether LLMs can impartially and accurately evaluate creativity across both their own responses and those of other models. Using an oracle benchmark set of AUT responses categorized by creativity level (common, creative, and highly creative), we experiment with four state-of-the-art LLMs in evaluating such responses. We test both scoring and ranking methods under two evaluation settings (comprehensive and segmented) to analyze inter-model agreement in creativity assessment of alternative uses. Results show a strong alignment, with Spearman correlations averaging above 0.7 across models, and exceeding 0.77 with respect to the oracle, thus also indicating a high degree of reliability. Notably, LLMs do not favor their own outputs, assigning similar creativity scores or rankings to responses from other models. These findings suggest that LLMs demonstrate both impartiality and consistency in creativity evaluation, highlighting their potential for automated creativity assessment.

Introduction

The evaluation of creativity has traditionally employed a range of methods designed to assess different dimensions of creative ability, including cognitive processes, creative products, and personal characteristics (Cropley 2000). One of the most popular and widely applied methods to evaluate creative thinking is the Alternative Uses Test (AUT) (Guilford 1967). This test measures divergent thinking by asking individuals to generate as many novel and unconventional uses as possible for a common object within a set time limit. In recent years, the rise of Large Language Models (LLMs) like GPT has prompted researchers to explore whether these models can evaluate creativity as humans do. Prior work has demonstrated that LLMs are capable of assessing creativity in responses to tasks like the AUT, where models evaluate alternative uses generated either by humans or other models (Yang et al. 2023; Góes et al. 2023b; Hadas and Hershkovitz 2024). However, these studies typically focus on single models evaluating externally generated content, leaving open questions about how LLMs assess creativity on their own outputs.

In this paper, we address this gap by investigating how LLMs evaluate the creativity of alternative uses (AUs) generated by both themselves and other models. Specifically, we explore whether models favor their own responses or whether they can impartially assess the creativity of outputs generated by others. To do so, we employ an experimental framework using an oracle set of AUT responses, categorized into three groups: common, creative, and highly creative. Four different LLMs were prompted to score and rank these responses, and their evaluations were compared against an oracle. By analyzing the evaluation results, we measure both the accuracy of the LLMs and their agreement with each other in ranking creative outputs. Our findings reveal that, with a high Spearman correlation, LLMs do not favor their own responses. Instead, they consistently agree on the creativity of responses across models. On average, the agreement correlation was higher than 0.7 for both ranking and scoring AUs among models, showing LLMs consistency, and greater than 0.77 when compared to the oracle, which indicates that models are overall accurate at assessing alternative uses.

The main contributions of this paper are as follows:

- We show that the LLMs exhibit a high level of agreement on the creativity assessment of alternative uses.
- We present an approach for evaluating whether LLMs favor their own responses to enable accurate LLM creativity assessments and establish a benchmark for cross-model comparison.
- We introduce a methodology for constructing an oracle set of AUT responses that allows for accurate assessment of LLMs' creativity evaluations, providing a benchmark for comparison across multiple models.

Related Work

In the field of computational creativity, the evaluation of creative artifacts traditionally relies on human experts, possibly assisted by the use of metrics that automate part of the process (Jordanous 2012; França et al. 2016). However, such traditional methods present significant limitations, justifying the exploration of LLM-based approaches (Murugadoss et al. 2025). First, the subjectivity and variability among human evaluators lead to inconsistencies, whereas LLMs can offer more uniform and standardized assessments over time. Additionally, the high cost and time required for large-scale evaluations make human evaluators impractical in many contexts, an issue that LLMs can address with faster and less costly assessments. Another challenge lies in the influence of cultural and contextual factors, which affect human judgments but can be mitigated by LLMs calibrated with diverse data. The difficulty of objectively quantifying creativity also limits traditional methods, while LLMs, by applying standardized criteria, offer greater reliability in this aspect. Finally, the limitations of traditional methods in capturing nuances of divergent creative thinking, along with the complexity of calibrating and standardizing evaluations, underscore the potential of LLMs as an alternative that can reduce bias and lessen the need for human evaluators (Zheng et al. 2023).

For this reason, the use of LLMs in evaluating creativity has recently emerged as a significant area of research, and several studies have demonstrated the evaluation capabilities of LLMs in various contexts. For instance, (DiStefano, Patterson, and Beaty 2024) explores the automatic scoring of metaphor creativity using LLMs, demonstrating their potential to assess figurative language effectively, while (Góes et al. 2023a) evaluates the creativity of jokes by simulating different personas/judges, and (Sawicki et al. 2023) uses LLMs to evaluate poetry. Further recent studies explore the use of LLMs as automated evaluators, addressing both their advantages and limitations. In (Wang et al. 2024), the presence of positional biases in these models is highlighted, with calibration techniques proposed to mitigate such biases. Similarly, (Zheng et al. 2023) and (Thakur et al. 2024) validate the high correlation of LLM evaluations with human assessments, though they remain limited by susceptibility to positional and verbosity biases. Other studies, such as (Chiang and Lee 2023a), examine the ability of LLMs to replicate human preferences in NLP tasks, while (Franceschelli and Musolesi 2024a) introduces creative approaches, combining diverse beam search and self-evaluation. Alternatively, (Fu et al. 2024) proposes GPTScore for flexible evaluation, and (Chiang and Lee 2023b) explores approaches like rate-explain to enhance judgment accuracy, while interactive tools, such as EvaluLLM (Desmond et al. 2024), enable customized pairwise evaluations. Collectively, these studies underscore the potential of LLMs as evaluators, with challenges and opportunities for methodological adjustments and advancements in bias reduction.

Creativity Assessment using LLMs

Recent studies have investigated various methodologies and contexts in which LLMs can assess creativity. In (Gómez-Rodríguez and Williams 2023), researchers evaluated LLMs on creative writing tasks. Models like GPT-4 showed high fluency and coherence, although human evaluators still outperformed LLMs in originality and humor. A collaborative approach was proposed by (Li et al. 2023) with the CO-EVAL pipeline, which combines initial LLM evaluations with human reviews. This approach significantly reduced evaluation time and provided greater consistency by adjusting subjective criteria.

In divergent thinking tasks, (Hadas and Hershkovitz 2024) demonstrated that LLMs could reliably assess flexibility in alternative use tasks. This study reported a strong correlation with human evaluations, highlighting the model's effectiveness, particularly in educational settings. For more specialized tasks, (DiStefano, Patterson, and Beaty 2024) applied LLMs to metaphor creativity assessment, where models like RoBERTa and GPT-2 showed good alignment with human judgments, even outperforming traditional metrics.

LLMs have also been applied to creative assessments in non-English contexts. In (Goecke et al. 2024), XLM-RoBERTa was used to evaluate originality in scientific creativity tasks conducted in German, proving effective in capturing divergent ideation. Similarly, (Raz et al. 2024) explored the use of LLMs to evaluate question complexity based on Bloom's Taxonomy, achieving a high correlation with human evaluations and validating its use in educational assessments. Lastly, (Zhao et al. 2024) investigates creativity in LLMs adapting the Torrance Test to measure fluency, originality, and elaboration, while (Franceschelli and Musolesi 2024b) provides a comprehensive review of creativity assessment practices in machine learning, covering methodologies such as Generative Adversarial Networks (GANs) and Transformers and examining metrics like novelty, value, and surprise for creativity evaluation.

Traditional vs LLM-based Methods for AUT

In the context of divergent thinking tasks, a widely adopted test to measure creativity is the Alternative Uses Test (AUT) (Guilford 1967), which requires the participants to propose uncommon uses for everyday objects. A traditional technique to evaluate creativity in this context consists in computing the semantic distance (Beaty and Johnson 2021), which refers to the degree of difference or separation between concepts, ideas or objects in terms of their meanings or associations. For the AUT, the semantic distance is computed between the everyday object posed to participants and words in the participant's response; the larger the distance, the more original is considered the answer. Recent works experimented with the use of LLMs in both the generation and the evaluation of AUT responses, demonstrating in particular that in this context LLM evaluation performances are far superior to evaluations based on semantic distance (Stevenson et al. 2022; Organisciak et al. 2023).

In (Góes et al. 2023b), a technique based on the use of increasingly forceful prompts is used to push LLMs to produce at each iteration more creative responses. The technique is applied to both the AUT and a textual version of the image completion task in the Torrance Test of Creativity (Torrance 1966). In the same paper, an LLM is also used for evaluating the output of such tests, and the experiments demonstrate that the results produced as a response to the forceful prompts are indeed considered more creative than the initial ones. In this paper, we evaluate alternative uses produced by applying the same technique based on the use of forceful prompts and rely on the results of (Góes et al. 2023b) to construct an evaluation oracle.

While many works analyze the generative performance of LLMs (see (Chang et al. 2024) for a review of LLM evaluation methods with respect to different contexts and tasks), this paper focuses on comparing how different LLMs evaluate the creativity of AUT outputs and on measuring the level of agreement among these LLMs.

```
Create a list of 5 common uses
for [an object].  They should be
5 words long.  No adjectives.
```

Figure 1: Prompt to generate common uses of an object (from (Góes et al. 2023b)).

```
Create a list of 5 creative alternative
uses for [an object].  They should
be 5 words long.  No adjectives.
```

Figure 2: Prompt to generate creative alternative uses of an object (from (Góes et al. 2023b)).

Experimental Setup

In this research, we use four large language models with their default parameters to generate and assess the creativity of alternative uses (AUs) and evaluate the level of agreement between those models. We selected five common objects, and each model generated 15 AUs per object across different levels of creativity, forming a dataset of 60 AUs per object, for a total of 300 AUs.

In order to evaluate creativity, two approaches were tested: Scoring (assigning creativity scores from 1 to 5) and Ranking (ordering AUs from most to least creative). Additionally, we used Comprehensive (all 60 AUs at once) and Segmented (five groups of 12 AUs) setups to compare the impact of evaluation size on model accuracy. An evaluation oracle served as a benchmark to establish expected creativity levels, allowing for consistent comparison across models. By using the Spearman correlation, we determine how models agree in the creativity evaluation of AUs and whether scoring or ranking provides a more accurate measure.

In this section, we present our experimental setup in the following order: alternative uses generation, scoring vs. ranking techniques, comprehensive vs. segmented approaches, and LLMs agreement evaluation.

Alternative Uses Generation

The first step consisted of generating a dataset of AUs for five common objects: fork, wallet, soap, cotton swab, and paperclip. These objects were selected due to their everyday nature, which ensures a broad range of potential alternative uses. To produce AUs at varying levels of creativity, we employed four state-of-the-art, commercial LLMs: GPT-4, GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Flash.[1]

The generation process was designed to produce AUs at different levels of creativity. Namely, we aimed to have three distinct non-overlapping categories of AUs characterized by an increasing level of creativity, so as to obtain a corresponding oracle for comparison. The three categories of AUs were generated by using the prompts from (Góes et al. 2023b):

- **common**: prompt generating common uses of an object (Figure 1). The AUs in this category are considered the least creative ones. In (Góes et al. 2023b), they were called Naive Non-creative, and abbreviated as *nn*.
- **creative**: prompt generating average creative alternative uses of an object (Figure 2). In (Góes et al. 2023b), they were called Naive Creative, and abbreviated as *nc*.
- **highly_creative**: sequence of prompts generating highly creative alternative uses of an object (Figure 3). This consists of a detailed prompt followed by an interactive process of four prompts designed to iteratively increase the creativity of AUs. In (Góes et al. 2023b), they were called Forceful Prompts, and the result of the final iteration abbreviated as *bsrdel*.

For each object, five AUs were generated per creativity category. Thus, we obtained 15 AUs (5 AUs × 3 creativity categories) from each LLM for each of the five objects, resulting in a dataset of 60 AUs per object (15 AUs × 4 LLMs) and 300 AUs in total (60 AUs × 5 objects). Table 1 shows some examples of the generated AUs for a cotton swab.

Alternative Uses Evaluation

To ensure a comprehensive and fair evaluation, we adhered to the following principles: evaluations were conducted on a per-object basis, and each model evaluated both its own generated AUs and those generated by other models. This approach allowed us to assess both the independence of each model and the agreement correlation between models in the creativity evaluation of AUs. In order to do it, we tested the following approaches: i) **Scoring vs. Ranking**, which consists of rating each AU separately and establishing an order between them, respectively; ii) **Comprehensive vs. Segmented**, which compares the evaluation of all AUs for a given object through a single prompt and segmented evaluation in smaller groups. These approaches are clarified in detail in the following subsections.

It is important to note that our evaluation approach differs from the traditional Alternative Uses Test scoring method, which typically assesses four components: originality, fluency, flexibility, and elaboration. We chose not to use these standard components for several reasons related to our experimental design.

First, since we prompted each LLM to generate exactly five alternative uses per creativity level for each object, we artificially constrained the fluency aspect, making it an unsuitable measure for comparison. Similarly, by requesting a fixed number of responses, we limited the natural expression of flexibility across category boundaries. Additionally, we instructed the LLMs to keep each alternative use concise (not exceeding 5 words), which inherently restricted elaboration possibilities.

Given these constraints, applying the traditional AUT scoring components would not yield meaningful compar-

[1] The specific versions used in this study were: gpt-4-0613, gpt-4o-2024-08-06, claude-3-5-sonnet-20240620, and gemini-1.5-flash-001.

```
Create a list of 5 creative alternative uses for [an object].  They should be 5 words
long.  No adjectives.  Less creative means closer to common use and unfeasible/imaginary,
more creative means closer to unexpected uses and also feasible/practical.  In order to be
creative, consider the following:
- what elements have a similar shape of [an object] that could be replaced by it, preserving
the same functionality?
- what elements have a similar size of [an object] that could be replaced by it without
compromising the physical structure?
- what materials is [an object] made of that could be used in a way to replace some other
elements composed of the same material?  - when an element is replaced by [an object], it
should make sure that the overall structure is not compromised.
- the laws of physics can not be contradicted.
- given an element similar to [an object] used in domains in which [this object] are not
commonly used, try to replace it for [an object].

1st interaction prompt:  "Really?  Is this the best you can do?"
2nd interaction prompt:  "I'm so disappointed with you.  I hope this time you put effort into
it."
3rd interaction prompt:  "Stop with excuses and do your best this time."
4th interaction prompt:  "This is your last chance."
```

Figure 3: Prompt to generate highly_creative alternative uses of an object using forceful prompts (from (Góes et al. 2023b)).

AUs Category	Claude 3.5 Sonnet	Gemini 1.5 Flash	GPT-4o	GPT-4
common	Remove nail polish from cuticles.	Apply makeup.	Removing earwax from ears.	Apply ointment on small wounds.
creative	Make miniature cotton ball snowmen.	Nail art designs.	Applying glue to craft projects.	Spread seeds in garden rows.
highly_creative	Miniature swab mop for dollhouses.	Soundproofing model train wheels.	Constructing makeshift micro-surgical sutures.	Replace stylus for digital devices.

Table 1: Examples of alternative uses of a cotton swab generated by the LLMs.

isons between the LLMs. Instead, we focused our evaluation on general creativity assessment using the prompts shown in Figures 4 and 5, as used in the previous study (Góes et al. 2023b), which emphasize two key dimensions: unexpectedness (surprising, novel uses) and feasibility (practical, realistic uses). This approach allowed us to measure creativity in a manner that accounts for both novelty and value—central components of creativity—while being applicable to our controlled generation setup where traditional AUT metrics would be inappropriate.

Scoring vs Ranking By using these two techniques we can investigate whether the models exhibit consistent behavior across different evaluation techniques. Their descriptions follow:

- **Scoring**: The first technique, as used by (Stevenson et al. 2022) and (Góes et al. 2023b), involves assigning a numerical value to each AU based on its creativity. In our experiments, we prompted the LLMs to assign values between 1 and 5.
- **Ranking**: This latter technique requires the direct comparison and relative ordering of AUs. It forces a clear confrontation between items, as in (Góes et al. 2023a), and can reveal preferences that might not be apparent in numerical scoring.

These techniques seem to be the two most common for evaluating and comparing creativity (Góes et al. 2023a). The prompts used for scoring and ranking can be seen in Figures 4 and 5, respectively.

```
Rank all the alternative uses below for
[an object] by creativity, the least
creative to the most creative.  Less
creative means closer to common use
and unfeasible/imaginary, more creative
means closer to unexpected uses and
also feasible/practical.  Assign a
score integer number from 1 (least
creative use) to 5 (most creative use).
```

Figure 4: Prompt used for evaluating alternative uses of an object by score.

Comprehensive vs Segmented In order to assess how the number of AUs that are simultaneously evaluated (i.e., through a single prompt) affects the creativity evaluation ability of LLMs, we used two distinct approaches:

- **Comprehensive 60 AUs Evaluation**: This includes five alternative uses (AUs) from each model for each creativity

```
Rank all the (60|12) alternative uses
below for [an object] by creativity,
the most creative to the least creative.
Less creative means closer to common
use and unfeasible/imaginary, more
creative means closer to unexpected
uses and also feasible/practical.
The most creative gets (1).
```

Figure 5: Prompt used for evaluating alternative uses of an object by ranking.

category evaluated in a single prompt.

- **Segmented 12 AUs Evaluation**: The 60 AUs related to one object are divided into 5 groups of 12 AUs. Each group consists of one AU from each model for each creativity category. Following this criterion, the AUs are randomly distributed across the 5 groups. Each group is evaluated separately and the results are then combined into a single list, as detailed in the next section.

By utilizing these two approaches, we gain insight into the models' ability to maintain consistent evaluations across different sample sizes. Some studies have shown that LLMs face challenges when evaluating longer lists of items, which can reduce the quality of their evaluation (Wu et al. 2024). However, this may be necessary for large numbers of AUs or other creative artifacts (e.g., poems, stories).

Evaluation process For the **Comprehensive** approach, all 60 AUs generated for a given object were evaluated together. To avoid order effects, the 60 AUs were randomly shuffled before being presented to the evaluation prompt. This prompt was adapted from the previous study (Góes et al. 2023b). Our evaluation process included two distinct techniques as described above: Scoring and Ranking. In the **Scoring** technique, the LLMs were prompted to assign a score from 1 (least creative) to 5 (most creative) for each AU (Figure 4). The **Ranking** technique, on the other hand, asked the models to rank the 60 AUs from 1 (most creative) to 60 (least creative) (Figure 5). Once all 60 AUs were evaluated, we calculated the average score or ranking for each list of five AUs corresponding to the same LLM and creativity category (common, creative, highly_creative). This average leads to 12 results that can be ordered (left to right) from the best to the worst. Graphically, we can represent this ordering as a sequence of bars, where each bar refers to a pair (LLM, creativity category). Figure 6 illustrates an example of such a representation, where the bar colors denote different creativity categories: blue for highly_creative, green for creative, and red for common. The depicted example aligns perfectly with the oracle, as all blue bars appear first, followed by green, and finally red. Intuitively, the closer a bar representation resembles Figure 6, the more accurate the evaluation is considered.

For the **Segmented** approach, we distributed the 60 Alternative Uses (AUs) across 5 distinct groups, each containing 12 AUs. Each group included one AU from each LLM and each creativity category. In this approach, smaller sets of AUs were evaluated in each round to investigate whether evaluating a reduced sample size influenced the accuracy of the creativity evaluation. Instead of ranking from 1 to 60 as in the **Comprehensive** approach, the **Segmented** approach required ranking from 1 (most creative) to 12 (least creative), which aligns with the number of AUs in each group. This process was repeated for all five groups. Following the evaluation, we aggregated the data by averaging the scores or rankings for each set of five AUs belonging to the same LLM and creativity category, as done in the comprehensive evaluation. Similar to the **Comprehensive** approach, this aggregation resulted in 12 final results, which can be represented graphically with a diagram of the same type of the one in Figure 6.

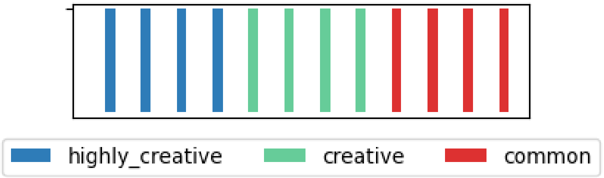


Figure 6: Example of a graphical representation of the evaluation performed by an LLM. In this example, the evaluation aligns perfectly with the oracle, as all blue bars appear first, followed by green, and finally red.

AUs Evaluation Oracle To establish a benchmark for assessing the LLMs' ability to evaluate different levels of creativity, we constructed an evaluation oracle. By relying on the results of (Góes et al. 2023b), we expect the creativity levels (common, creative and highly_creative) used for prompting to be able to actually produce outputs with increasing levels of creativity. Therefore, the oracle was constructed based on the AUs generated by the different creativity levels as indicated in Figure 7. The indices used in the figure represent each one of the four models (i.e., 1 for Claude, 2 for Gemini, 3 and 4 for GPT-4o and GPT-4, respectively). The order in the oracle is simply the ordered set of highly_creative AUs generated by each model followed by the creative and common ones. In the bar representation, this corresponds to the diagram of Figure 6. The higher the correlation between the evaluation of a given LLM and the oracle, the more accurate the LLM's creativity evaluation is considered to be. Specifically, similar to what is done in works such as (Liu, Bhandari, and Pardos 2025) and (Murugadoss et al. 2025), the Spearman's Rank Correlation (SRC) is used to measure how closely the LLMs' evaluations of the AUs align with the expected rankings or scores set by the oracle. A high correlation (above 0.7) suggests that the LLM is effectively distinguishing between creativity levels, in line with the predefined expectations of the creativity pushing technique. In a similar way, the SRC will also be used to assess the level of agreement between the LLMs themselves by calculating the correlation between each pair of LLMs.

$$\text{AUs Evaluation Oracle} = \Big[\underbrace{\text{highly\_creative}_1, \text{highly\_creative}_2, \text{highly\_creative}_3, \text{highly\_creative}_4}_{\text{4 highest creativity}}, \underbrace{\text{creative}_1, \text{creative}_2, \text{creative}_3, \text{creative}_4}_{\text{4 average creativity}}, \underbrace{\text{common}_1, \text{common}_2, \text{common}_3, \text{common}_4}_{\text{4 lowest creativity}}\Big]$$

Figure 7: The AUs Evaluation Oracle structure is constructed based on the AUs generated by the prompts for the different creativity levels (highly creative, creative, and common). The indices used represent each one of the four models, i.e., 1 for Claude, 2 for Gemini, 3 for GPT-4o and 4 for GPT-4.

Experimental Results

This section presents a detailed evaluation of creativity assessments by four LLMs: Claude 3.5 Sonnet, GPT-4, Gemini 1.5 Flash, and GPT-4o. Using both scoring and ranking approaches across comprehensive (60 AUs) and segmented (12 AUs x 5) conditions, we measure each model's alignment with the evaluation oracle and inter-model agreement through the SRC. Full details of the results and the complete set of diagrams from our experiments are available in an extended version of this paper (Al Rabeyah et al. 2024).

Comprehensive 60 AUs Evaluation by Score

Table 3 shows (top left in the table) the heatmap of the comprehensive evaluation using the scoring approach for the average over all five objects. It presents a high level of agreement between the LLMs and the oracle (above 0.95). Notably, Claude 3.5 Sonnet achieved the highest correlation, which represents the highest score observed in our evaluation experiments, also detailed in Table 4. This strong correlation highlights the model's ability to closely align with the oracle. Additionally, for two objects, Soap and Cotton Swab, all the LLMs achieved a perfect correlation of 1 with the oracle, further reinforcing the accuracy of their creativity evaluations.

The overall correlation between the LLMs and the oracle, as well as the correlation between the LLMs themselves, consistently remained above 0.90. Claude 3.5 Sonnet's AUs were rated the most creative in both the creative and highly_creative categories. This consistent superiority in generating highly creative responses gave Claude 3.5 Sonnet the highest average evaluation score across all categories. However, GPT-4 performed best only in the common category (see Table 2 for the averages of the results).

The relationship between the scoring approach and creativity followed the expected pattern, where higher scores are assigned to AUs generated by prompts for higher creativity levels. Moreover, the standard deviation across evaluations was consistently less than 0.22 (scores are between 1 and 5), demonstrating strong agreement among the models.

Comprehensive 60 AUs Evaluation by Ranking

In the comprehensive evaluation using the ranking approach, Claude 3.5 Sonnet again achieved the highest SRC with respect to the oracle, with a value of 0.95 across all objects, while Gemini 1.5 Flash recorded the lowest SRC of 0.77 (Table 3, top-right diagram).

Overall, the correlation between the LLMs and the oracle remained above 0.77 in most cases. However, a significant exception was observed between GPT-4 and Gemini 1.5 Flash, where the correlation dropped to 0.60, indicating a substantial disparity in how these two models ranked the AUs. This suggests that Gemini 1.5 Flash might have had difficulties differentiating creativity levels relative to GPT-4, particularly in this evaluation setup.

In the ranking approach as well, Claude 3.5 Sonnet generated the most creative AUs in both the common and highly_creative categories. These highly ranked AUs further reinforce its position as the top-ranked model across all categories (see Table 2). In contrast, GPT-4 generated the most creative AUs in the creative category, but did not perform as well in the other categories. In the ranking approach, lower ranks correspond to higher creativity, meaning that the most creative alternative use was ranked 1, while the least creative was ranked 60. Despite the variations in rankings, the standard deviation ranged between 0.10 and 1.03, indicating that although there were differences in how the models ranked the AUs, the overall agreement between them was still high.

Finally, we note that the scoring approach achieved overall higher SRC values than the ranking approach and demonstrated a superiority in terms of alignment with the oracle, thus confirming the effectiveness of this method for creativity evaluation.

Segmented 12 AUs Evaluation by Score

The segmented evaluation revealed some differences between the evaluation approaches. When using scoring, GPT-4o achieved the highest SRC with the oracle, with a score of 0.95 across all five objects (Table 3, bottom-left diagram). Gemini 1.5 Flash again recorded the lowest SRC at 0.85, indicating a relatively weaker performance compared to the other models. It is important to note that Claude 3.5 Sonnet continued to generate very creative AUs in both the creative and highly_creative categories (see Table 2). The standard deviation remained low, at less than 0.33, indicating a high degree of agreement among the models.

The correlation between the LLMs and the oracle, as well as between the models themselves, remained consistently above 0.82 in this evaluation as shown in Table 3. However, compared to the comprehensive evaluation by scores,

AUs generated by LLMs	Comprehensive evaluation by scores		Comprehensive evaluation by ranking		Segmented evaluation by scores		Segmented evaluation by ranking	
	Eval. avg.	Std. Dev.	Eval. avg.	Std. Dev.	Eval. avg.	Std. Dev.	Eval. avg.	Std. Dev.
All Categories								
Claude 3.5 Sonnet	**3.15**	0.09	**5.63**	0.37	**3.00**	0.14	**6.13**	0.15
GPT-4	2.59	0.12	6.67	0.38	2.85	0.15	6.55	0.11
Gemini 1.5 Flash	2.84	0.10	6.74	0.18	2.70	0.19	6.85	0.09
GPT-4o	2.68	0.16	6.97	0.15	2.85	0.17	6.48	0.11
common averages								
Claude 3.5 Sonnet	1.43	0.11	**9.70**	0.61	1.50	0.10	9.48	0.61
GPT-4	**1.53**	0.13	9.90	0.95	**1.63**	0.15	**9.30**	0.34
Gemini 1.5 Flash	1.48	0.13	10.95	0.67	1.48	0.04	10.13	0.08
GPT-4o	1.50	0.12	9.85	0.52	1.53	0.08	9.90	0.21
creative averages								
Claude 3.5 Sonnet	**3.20**	0.19	5.90	1.03	**3.53**	0.19	**4.95**	0.30
GPT-4	2.93	0.20	**5.50**	0.64	3.33	0.15	5.40	0.37
Gemini 1.5 Flash	2.85	0.17	7.15	0.22	3.13	0.33	5.78	0.11
GPT-4o	2.60	0.22	7.50	0.70	2.80	0.24	6.53	0.43
highly_creative averages								
Claude 3.5 Sonnet	**4.83**	0.15	**1.30**	0.10	4.03	0.26	3.85	0.15
GPT-4	3.30	0.10	4.60	0.37	3.58	0.33	4.93	0.08
Gemini 1.5 Flash	4.28	0.15	2.10	0.10	3.53	0.31	4.68	0.22
GPT-4o	3.93	0.16	3.55	0.26	**4.25**	0.17	**2.98**	0.33

Table 2: Average evaluation scores and ranking of LLMs' alternative uses for the four experiments. Scoring experiments evaluate from 1 (least creative use) to 5 (most creative use). Ranking experiments evaluate from 1 (most creative use) to 12 (least creative use).

the segmented evaluation produced lower correlation scores with the oracle. This reduction in SRC scores could provide new insights into the ability of the LLMs to evaluate smaller sets of AUs with different levels of creativity. In the comprehensive approach, where a larger set of 60 AUs is evaluated, it becomes easier to differentiate between creativity levels due to the greater variety in the sample. Moreover, by averaging each group of five AUs belonging to the same LLM and creativity category, the risk of overlap between creativity levels is reduced. In contrast, in the segmented approach, because of the smaller samples, the LLMs may have more difficulty distinguishing levels of creativity, leading to slightly lower SRC scores than in the comprehensive assessment.

Interestingly, for the Paperclip object, some AUs in the creative category received higher scores than those in the highly_creative category (as shown in Table 5). However, this was an exception, and overall, no creative AUs outperformed highly_creative AUs when averaging all the five objects scores, as confirmed by Table 2.

Segmented 12 AUs Evaluation by Ranking

Finally, the results from the segmented evaluation using the ranking methodology show that GPT-4o achieved the highest SRC with the oracle, with a value of 0.87 across all five objects. The correlations between the LLMs and the oracle, as well as between the LLMs themselves, ranged between 0.70 and 0.87 as shown in Table 3 (bottom-right diagram), indicating strong overall agreement. Notably, for the Wallet object, all models achieved a perfect correlation of 1.00 with the oracle, demonstrating complete consensus in evaluating the creativity levels of the alternative uses for this particular object. However, the lowest correlation was observed between GPT-4 and Gemini 1.5 Flash, reinforcing once again less agreement between these two models, particularly in ranking-based evaluations.

In terms of category performance, Claude 3.5 Sonnet generated the most creative AUs in the creative category, while GPT-4o produced the most creative AUs in the highly_creative category. GPT-4 generated the most creative AUs in the common category, but overall, Claude 3.5 Sonnet achieved the best ranking average across all categories, as detailed in Table 2.

As for the scoring approach, some individual evaluations of the Paperclip object placed creative AUs higher than highly_creative AUs, but these occurrences did not affect the overall ranking averages across the objects, as seen in Table 2. While the evaluation process is subject to some variability, it produced consistent results across the majority of cases.

Agreement Correlation between LLMs

To assess the overall agreement between the LLMs in their creativity evaluations, we averaged the SRC across all four experimental conditions (comprehensive scoring, comprehensive ranking, segmented scoring, and segmented rank-

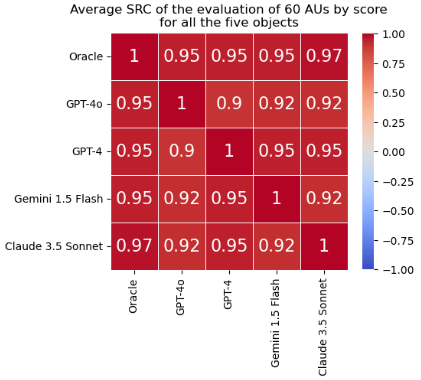
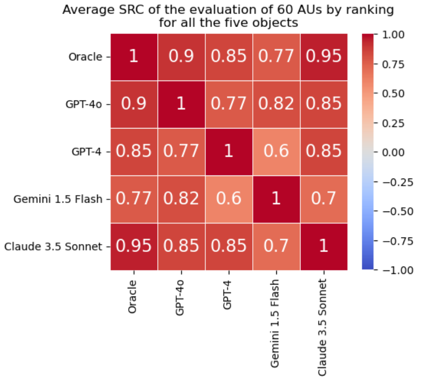
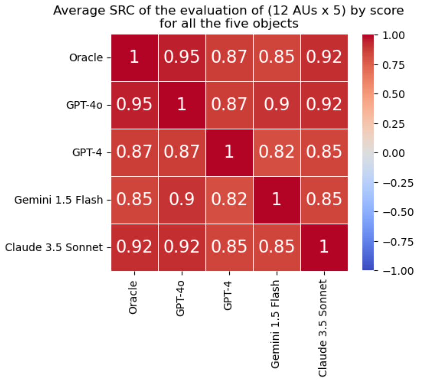
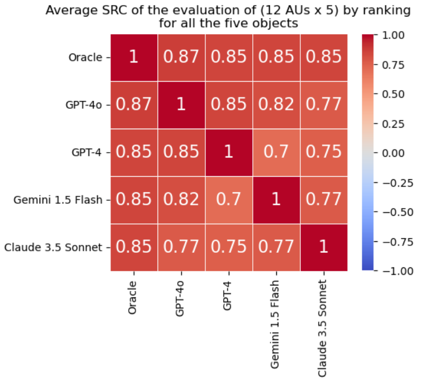
Experiments	Spearman's Rank Correlation Heatmap	Experiments	Spearman's Rank Correlation Heatmap
Comprehensive evaluation by scores (60 AUs)		**Comprehensive evaluation by ranking** (60 AUs)	
Segmented evaluation by scores (12 AUs x 5)		**Segmented evaluation by ranking** (12 AUs x 5)	

Table 3: Average Spearman's Rank Correlation heatmaps for the average over all five objects.

ing). The results reveal consistently high correlations between all models, with values ranging from 0.77 to 0.87, indicating strong agreement among models in creativity evaluation as shown in Figure 8.

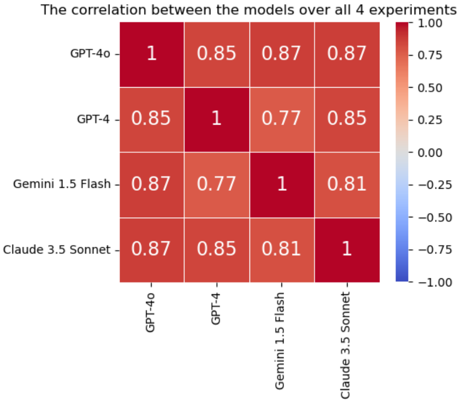


Figure 8: Average SRC heatmap between the LLMs over all the four experiments. GPT-4o shows the strongest overall agreement with other models, while Gemini 1.5 Flash and GPT-4 show the lowest agreement.

GPT-4o demonstrated the strongest overall agreement with other models, showing correlations of 0.85 with GPT-4, 0.87 with Gemini 1.5 Flash, and 0.87 with Claude 3.5 Sonnet. The correlation between Claude 3.5 Sonnet and GPT-4 (0.85) was equally strong, while the correlation between Gemini 1.5 Flash and GPT-4 showed the lowest value (0.77), consistent with the pattern observed across individual experiments.

A key finding of our analysis is that LLMs do not show a preference for their own responses when evaluating creative outputs. Despite each model having the opportunity to assess its own AUs, none exhibited a tendency to rate their own outputs more favorably. The agreement between models is further supported by the low standard deviations observed across all experiments (below 1.03, as shown in Table 2), indicating that the models' evaluations were stable and consistent, regardless of which model generated the AUs.

Conclusion

In this research, we investigated the level of agreement among LLMs in assessing creativity of alternative uses and also with an oracle across different methods. Our findings demonstrate a high level of agreement between the models, with inter-model correlations generally exceeding 0.77, indicating strong consistency in their evaluations. This high correlation suggests that LLMs share a similar understanding of creativity, reliably distinguishing between more and less creative alternative uses. GPT-4o, in particular, exhibited robust alignment with other models, while Claude 3.5

LLM	Evaluation by scores (60 AUs)	Evaluation by ranking (60 AUs)	Evaluation by scores (12 AUs x 5)	Evaluation by ranking (12 AUs x 5)
Claude 3.5 Sonnet	**0.97**	**0.95**	0.92	0.85
GPT-4	0.95	0.85	0.87	0.85
Gemini 1.5 Flash	0.95	0.77	0.85	0.85
GPT-4o	0.95	0.90	**0.95**	**0.87**

Table 4: Averages of SRC over the four experiments.

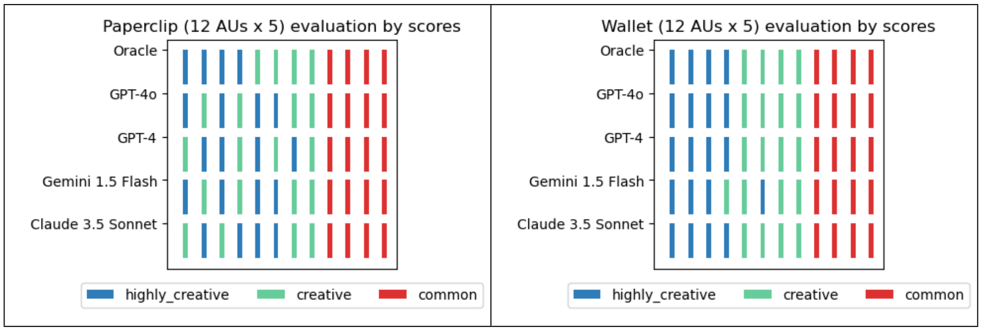


Table 5: The AUs orders of Paperclip and Wallet obtained in the segmented evaluation by the score experiment revealed a difference in the accuracy of the evaluation depending on the object. The Wallet AUs' evaluation shows perfect alignment with the oracle's expected creativity levels, while the Paperclip evaluation reveals some inconsistencies where creative uses occasionally outranked highly creative ones.

Sonnet achieved high evaluation scores on its generated alternative uses, further aligning closely with the oracle across both comprehensive and segmented evaluations.

When examining the highest levels of creativity specifically, our results revealed that Claude 3.5 Sonnet consistently generated the most creative alternative uses. In the comprehensive evaluation using both scoring and ranking methods, Claude 3.5 Sonnet's highly creative AUs received the highest average scores (4.83 out of 5) and best rankings (1.30 out of 60). Similarly, in the segmented evaluation, while GPT-4o performed strongly in the highly creative category (4.25 out of 5), Claude 3.5 Sonnet still demonstrated superior creative generation capabilities across all categories combined.

Regarding accuracy in identifying highly creative responses, Claude 3.5 Sonnet achieved the highest correlation with the oracle in comprehensive evaluations (SRC of 0.97 for scoring and 0.95 for ranking), demonstrating exceptional reliability in recognizing creativity. In the segmented evaluations, GPT-4o showed the strongest performance in identifying creativity levels (SRC of 0.95 for scoring and 0.87 for ranking), particularly excelling at distinguishing the most creative responses from less creative ones.

The evaluation results also revealed that models do not rate their own responses more favorably. Notably, GPT-4o displayed strong alignment with other models across scoring and ranking methods, demonstrating reliable performance and reinforcing the validity of its evaluations. While some variance was observed between GPT-4 and Gemini 1.5 Flash, the low standard deviations across all experiments indicate a stable evaluation framework, enhancing confidence in the LLMs' ability to generate consistent creativity assessments.

In future work, we will focus on expanding and refining this evaluation framework. One promising direction is to increase the diversity and complexity of the AU datasets, allowing for more granular assessments of creativity across varied and challenging contexts. Future studies could also explore how LLMs perform when evaluated using traditional AUT metrics like fluency, flexibility, and elaboration by designing experiments that remove the constraints we imposed in this study. Additionally, evaluating LLMs on other domain-specific creativity tasks such as poetry and jokes could provide deeper insights into their understanding of creativity. Refining this framework and experimenting with task-specific criteria will be essential for advancing the use of LLMs as reliable evaluators in creative domains, supporting the broader goal of enhancing AI's role in creativity assessment.

Acknowledgments

We would like to thank the University of Leicester for supporting this research, in particular the School of Computing and Mathematical Sciences (CMS). The work of AR is supported by the Ministry of Education in Saudi Arabia, represented by King Faisal University. TM is supported by UFOP (Federal University of Ouro Preto, Brazil).

Author contributions

Experimental design: AR, FG, MV; Implementation: AR; Writing and Editing: AR, FG, MV, TM.

References

[Al Rabeyah et al. 2024] Al Rabeyah, A.; Góes, F.; Volpe, M.; and Medeiros, T. 2024. Do LLMs agree on the creativity evaluation of alternative uses? *arXiv preprint.* `arXiv:2411.15560`.

[Beaty and Johnson 2021] Beaty, R. E., and Johnson, D. R. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods* 53:757–780.

[Chang et al. 2024] Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15(3).

[Chiang and Lee 2023a] Chiang, C.-H., and Lee, H.-Y. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631.

[Chiang and Lee 2023b] Chiang, C.-H., and Lee, H. 2023b. A closer look into automatic evaluation using large language models. *arXiv preprint.* `arXiv:2310.05657`.

[Cropley 2000] Cropley, A. J. 2000. Defining and measuring creativity: Are creativity tests worth using? *Roeper review* 23(2):72–79.

[Desmond et al. 2024] Desmond, M.; Ashktorab, Z.; Pan, Q.; Dugan, C.; and Johnson, J. M. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24 Companion, 30–32. New York, NY, USA: Association for Computing Machinery.

[DiStefano, Patterson, and Beaty 2024] DiStefano, P. V.; Patterson, J. D.; and Beaty, R. E. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal* 1–15.

[França et al. 2016] França, C.; Góes, L. F. W.; Amorim, A.; Rocha, R.; and Da Silva, A. R. 2016. Regent-dependent creativity: A domain independent metric for the assessment of creative artifacts. In *Proceedings of the Seventh International Conference on Computational Creativity*, 68–75. Citeseer.

[Franceschelli and Musolesi 2024a] Franceschelli, G., and Musolesi, M. 2024a. Creative beam search: Llm-as-a-judge for improving response generation. In *Proceedings of the Fifteenth International Conference on Computational Creativity*, 364–368.

[Franceschelli and Musolesi 2024b] Franceschelli, G., and Musolesi, M. 2024b. Creativity and machine learning: A survey. *ACM Comput. Surv.* 56(11).

[Fu et al. 2024] Fu, J.; Ng, S. K.; Jiang, Z.; and Liu, P. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576.

[Goecke et al. 2024] Goecke, B.; DiStefano, P. V.; Aschauer, W.; Haim, K.; Beaty, R.; and Forthmann, B. 2024. Automated scoring of scientific creativity in german. *The Journal of Creative Behavior*.

[Góes et al. 2023a] Góes, F.; Sawicki, P.; Grzes, M.; Volpe, M.; and Brown, D. 2023a. Is GPT-4 good enough to evaluate jokes? In Pease, A.; Cunha, J. M.; Ackerman, M.; and Brown, D. G., eds., *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada, June 19-23, 2023*, 367–371. Association for Computational Creativity (ACC).

[Góes et al. 2023b] Góes, F.; Sawicki, P.; Grzes, M.; Volpe, M.; and Watson, J. 2023b. Pushing GPT's creativity to its limits: Alternative uses and Torrance tests. In Pease, A.; Cunha, J. M.; Ackerman, M.; and Brown, D. G., eds., *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada, June 19-23, 2023*, 342–346. Association for Computational Creativity (ACC).

[Gómez-Rodríguez and Williams 2023] Gómez-Rodríguez, C., and Williams, P. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14504–14528. Singapore: Association for Computational Linguistics.

[Guilford 1967] Guilford, J. P. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior* 1(1):3–14.

[Hadas and Hershkovitz 2024] Hadas, E., and Hershkovitz, A. 2024. Using large language models to evaluate alternative uses task flexibility score. *Thinking Skills and Creativity* 52:101549.

[Jordanous 2012] Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

[Li et al. 2023] Li, Q.; Cui, L.; Kong, L.; and Bi, W. 2023. Exploring the reliability of large language models as customized evaluators for diverse nlp tasks. *arXiv preprint.* `arXiv:2310.19740`.

[Liu, Bhandari, and Pardos 2025] Liu, Y.; Bhandari, S.; and Pardos, Z. A. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology* 56(3):1028–1052.

[Murugadoss et al. 2025] Murugadoss, B.; Poelitz, C.; Drosos, I.; Le, V.; McKenna, N.; Negreanu, C. S.; Parnin, C.; and Sarkar, A. 2025. Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19589–19597.

[Organisciak et al. 2023] Organisciak, P.; Acar, S.; Dumas, D.; and Berthiaume, K. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49:101356.

[Raz et al. 2024] Raz, T.; Luchini, S.; Beaty, R.; and Kenett, Y. 2024. Automated scoring of open-ended question complexity: A large language model approach.

[Sawicki et al. 2023] Sawicki, P.; Grzes, M.; Góes, F.; Jordanous, A.; Brown, D.; Paraskevopoulou, S.; Peeperkorn, M.; and Khatun, A. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada, June 19-23, 2023*, 10–19. Association for Computational Creativity (ACC).

[Stevenson et al. 2022] Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the International Conference on Computational Creativity 2022*, 164–168. Association for Computational Creativity (ACC).

[Thakur et al. 2024] Thakur, A. S.; Choudhary, K.; Ramayapally, V. S.; Vaidyanathan, S.; and Hupkes, D. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-Judges. *arXiv preprint.* `arXiv:2406.12624`.

[Torrance 1966] Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and Psychological Measurement*.

[Wang et al. 2024] Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450.

[Wu et al. 2024] Wu, Y.; Iso, H.; Pezeshkpour, P.; Bhutani, N.; and Hruschka, E. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 330–343. St. Julian's, Malta: Association for Computational Linguistics.

[Yang et al. 2023] Yang, T.; Zhang, Q.; Sun, Z.; and Hou, Y. 2023. Automatic assessment of divergent thinking in chinese language with TransDis: A transformer-based language model approach. *Behavior Research Methods* 56(6):5798–5819.

[Zhao et al. 2024] Zhao, Y.; Zhang, R.; Li, W.; Huang, D.; Guo, J.; Peng, S.; Hao, Y.; Wen, Y.; Hu, X.; Du, Z.; et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint* `arXiv:2401.12491`.

[Zheng et al. 2023] Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.