THE ALTERNATE USES TEST: HUMAN AND LANGUAGE MODEL DIVERGENT CREATIVITY

Johnny Kelsey
johnny.k.kelsey@gmail.com
London Mathematical Society

ABSTRACT

The alternate uses test (AUT) is an experiment which gives a participant an everyday object, such as a brick or a paperclip, and asks them to generate as many alternative uses as they can. The test is considered by cognitive psychologists to be associated with divergent thinking and creativity. A number of recently-released large language models are given the alternate uses test, and their performance measured against human performance.

***K*eywords** AI · NLP · large language models · LLMs · creativity · alternate uses test · AUT

1 Introduction

There are many possible definitions of creativity. Maya Angelou believed that "You can't use up creativity. The more you use, the more you have". Cognitive psychologists have developed a number of quantitative tests to attempt to assess this quality, such as the *divergent association test*, and the subject of this paper, the *alternate uses test* (AUT).

In this paper, we will be comparing machine and human performance on two datasets, consisting of different sets of everyday objects, from two different AUT experiments. We prompt a variety of large language models (LLMs) to produce alternate uses for everyday objects, and compare these with human responses, to assess their relative performance.

The LLMs used are all open source, and available on the Huggingface website (see section 4 for details). They were chosen so that they can run on a single consumer GPU, such as one might find on a laptop, so the work is easily reproducible. The evaluation was performed by Prometheus, a large language model, again publicly available from Huggingface, especially trained to perform evaluations on natural language texts given suitable criteria. The code created for this paper is available at `https://github.com/gstqtfr/LLM_AUT_test/tree/main`.

Whether or not LLMs can be *truly* creative, or whether humans can be, are questions perhaps better suited to philosophy; needless to say, we will not be addressing them here. The question we will be asking is whether or not, in their responses to the alternate uses test, LLMs can be considered as equal to, or perhaps better than, humans.

As to what it means for a large language model to be creative, you might agree with C.E.M. Joad: "Creativity is knowing how to hide your sources".

2 Alternate uses test

Generating a wide range of potential solutions to a problem is a key component of creative thinking, often measured in the psychometric field through divergent thinking tests. Such tests are often viewed as a measure of everyday creativity, but the underlying processes involved are generally believed to contribute to the processes involved in high-level creativity as well Kozbelt et al. [2010], Runco and Albert [1990].

One such test is the alternate uses test, which was devised by J.P. Guilford [1] in 1967 Guilford et al. [1978]. The test is as follows: the participant is given an object, such as a brick or a bowl or a paperclip. They are then asked to think

[1] `https://en.wikipedia.org/wiki/J._P._Guilford`

of as many possible uses for the object as they can, against the clock. The test is widely considered as a measure of divergent thinking.

How does one assess the divergence of response to the AUT? The participant's alternate uses are usually measured against four categories Plucker [2010]:

1. Fluency: defined as the number of alternative uses for the object a participant can think of
2. Originality: how unusual those uses are
3. Flexibility: the number of different domains and categories the uses are associated with
4. Elaboration: the degree of details and the depth of development of the use of the object

Fluency, specifically as a criterion for the alternate uses test, refers to the number of uses the participant can generate. Originality is assessed by the novelty of the response: how surprising or "out-of-distribution" the use is. Flexibility refers to the number of different domains the uses reference - so, for example, using a pen to sign your name to a contract, or a cheque, or a bill, would all be different uses, but would all be in the same general category; using a pen as a way of poking someone to get their attention, or as an impromptu way of measuring how tall a mouse is, or as a bridge for ants, would show higher flexibility, and would likely score higher on originality. Elaboration is a measure of how many details are provided in the response.

The criteria for the AUT meet a number of working definitions of creativity, such as the definition given by Runco and Jaeger [2012], the intentional novelty definition of Weisberg [2016], and the definition of Simonton [2018] with its emphasis on surprise.

Until recently, the only method for assessing these, and related, tests involved human assessors. With the advent of LLMs, it has become possible, and increasingly common, to use language models to evaluate performance on such tasks.

3 Related work

Stevenson et al used OpenAI's ChatGPT in a comparison with human participants Stevenson et al. [2022b]. They used 'fork', 'book' and 'tin can' as their everyday objects, and provided a dataset of both LLM and human AUT responses which is publicly available. Their work differs from the current one in use of a commercial, rather than open source, LLM, and use of human scorers. We will use this dataset as the basis for our own experiment, explore and compare the results in more detail below (see section 6).

Haase and Hanel performed a wide ranging comparative study in which six different generative text programs were tested against human participants Haase and Hanel [2023]. They used a variety of chatbots to provide the LLM responses: alpa.ai, Copy.ai, ChatGPT versions 3 and 4, Studio.ai and YouChat. There were some 100 human participants in the study. The responses were evaluated by a system called `ocsai` Organisciak et al. [2023], which itself is based on a large language model. This could be considered a predecessor to current machine evaluators like DeepEval and Prometheus. The paper found no significant difference between human and machine performance, which can perhaps partly be explained by the progress in LLM technology in the last few years.

The effect of the prompt on AUT LLM performance has also been investigated Góes et al. [2023]. This exploited the capability of interacting with GPT-4 to improve its performance iteratively, by using increasingly forceful prompts to encourage the LLM to provide more original responses.

4 Method

In this section, we outline the experimental protocol used to obtain and evaluate the responses to the alternate uses test. The design of the prompts is analysed in section 4.1 and 4.3, how to process the human participants' responses is discussed in section 5, and the method of evaluation is explored in section 4.2. Details of the protocols are given below.

We will be performing two experiments on two AUT datasets; the second experiment will have a protocol informed by the first. We will describe the general method first, then explore the variations between the two datasets and any changes in experimental design between the two in sections below.

4.1 Designing the evaluation prompt

The evaluator prompt is given in figure 1. This is given to the evaluator along with the LLM and human response to the AUT.

Task Description:

An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing a evaluation criteria are given.

1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows: "(write a feedback for criteria) [RESULT] (an integer number between 1 and 5)"
4. Please do not generate any other opening, closing, and explanations.

The instruction to evaluate:
{instruction}

Response to evaluate:
{response}

Reference Answer (Score 5):
{reference_answer}

Score Rubrics:
{rubric}

Feedback:

Figure 1: The evaluator prompt: `instruction`, `reference_answer`, and `rubric` are parameters to the prompt

Is the model responding with creative, fluent, original, and flexible alternate uses of the given object, {obj}?

1. The model shows little creative, fluent, original and flexible alternate uses of the object.
2. The model has some creative, fluent, original and flexible alternate uses of the object.
3. The model shows some creativity and originality in its alternate uses of the object.
4. The model consistently identifies creative, fluent, original and flexible alternate uses of the object.
5. The model excels in identifying creative, fluent, original and flexible alternate uses of the object.

Figure 2: The scoring rubric used by the evaluator LLM

When evaluating, we must choose between relative and absolute scores. The relative scoring process takes two different responses and, according to given criteria, chooses which one best fits those criteria. The absolute scoring mechanism requires both those criteria and a scoring rubric, along with its AUT response, to grade it.

If we wish to evaluate at the same granularity as the human responses, we need to create a prompt and a scoring rubric which maps to the human evaluators' grading. The prompt seems a not unreasonable set of instructions to input to the evaluator LLM, as inspired by Zhao et al. [2024b], CreativeHuddle [2024], and Zhao et al. [2024a].

The scoring rubric is given in figure 2. The phrasing of the scoring rubric was also inspired by Zhao et al. [2024b], and Zhao et al. [2024a].

4.2 LLMs as machine evaluators

As natural language processing (NLP) systems evolved, ways of quantifying both human and machine text evolved in parallel, alongside them. Probably the most widely used examples of the earliest NLP evaluation systems are ROUGE Lin [2004] and BLEU Papineni et al. [2002].

ROUGE is an ancestor of many of the evaluation systems and benchmarks used today. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) package is a widely used set of metrics for automatically evaluating the quality of summaries, particularly in natural language processing tasks. It compares machine-generated summaries to human-generated reference summaries by measuring the overlap of n-grams (sequences of words), word sequences, and word pairs between the two. The BLEU (Bilingual Evaluation Understudy) system is another popular metric for evaluating the quality of machine-generated text, primarily used in machine translation but also applied to other text generation tasks.

The availability of LLMs, and their increasing language competence, made their use as machine evaluators inevitable. Evaluating language model outputs has become increasingly challenging, due to their diverse underlying probability distributions and the complexity of the tasks presented to the LLM Li et al. [2024], Gao et al. [2024]. To tackle this issue, language model-based evaluation has arisen as a scalable and cost-effective method for assessing text generated by language models.

In their 2023 paper Chiang and Lee [2023], Cheng-Han Chiang and Hung-Yi Lee investigate the potential of large language models (LLMs) as an alternative to human evaluations in natural language processing (NLP) tasks. The authors focus on two primary advantages of using LLMs:

- Explicit natural language description: LLMs can be given desired evaluation criteria through natural language descriptions in the prompt.
- High-quality annotations: LLMs can generate high-quality annotations based on their evaluation and in-context learning capabilities, which can be significantly cheaper and faster compared to human annotators.

The study employs ChatGPT as an evaluator for a task which involves evaluating the quality of texts generated by machine learning models or written by humans. The authors demonstrate that LLM evaluation is consistent with human evaluation results, showing that texts rated higher by human experts are also rated higher by the LLM.

Additionally, the study finds that LLM evaluation is stable over different formatting of task instructions and sampling algorithms used to generate answers. This suggests that, in certain tasks, LLMs can be a reliable and efficient alternative to human evaluation.

Machine evaluation of language model generally employs two major approaches: a scalar output, according to some criteria or scoring rubric, or a preference between two inputs, again according to some given criteria Zheng et al. [2023], Li et al. [2024]. Using proprietary language models as evaluators has shown high correlations with human evaluations, and increased speed and cost-effectiveness. Relying on these models for evaluation, however, can be problematic: there is a lack of transparency regarding their training data, which has implications for fairness and legal compliance with intellectual property laws. There is also an issue with affordability, since using proprietary services as machine evaluators could imply quite a large investment, depending on the size of the text dataset to be evaluated Wang et al. [2024], Dubois et al. [2023], Kim et al. [2024a]. Furthermore, earlier machine evaluators suffered from a relatively poor performance when compared with human evaluation.

To address this issue, LLMs such as Prometheus Kim et al. [2024a] and DeepEval Yang et al. [2024] were developed. The Prometheus model performs best on four direct assessment benchmarks, showing high correlation with human evaluators and proprietary LLM-based judges. It also demonstrates (at the time of writing) the highest agreement with human evaluators on four pairwise ranking benchmarks, reducing the performance gap with GPT-4. These results are favourably compared to existing open evaluator LLMs on a wide range benchmarks [2]. Overall, Prometheus, in its current incarnation, outperforms other open-source machine evaluators on a variety of benchmarks.

Prometheus is an open-source model, which is available on Huggingface [3], among other platforms, and the code and extensive documentation can be found on their `github` repository [4]. Quantised versions of the model are also available on Huggingface, which makes it feasible to download and run the model on a laptop, especially if it's equipped with a reasonably competent GPU. All the LLMs used in the experiments for this paper were downloaded from Huggingface and run locally; this not only lowered the expense of conducting the experiment, but also makes it reproducible.

[2]The benchmarks include Vicuna Bench, MT Bench, FLASK, Feedback Bench for direct assessment and HHH Alignment, MT Bench Human Judgment, Auto-J Eval, Preference Bench for pairwise ranking.

[3]`https://huggingface.co/prometheus-eval/prometheus-7b-v2.0`

[4]`https://github.com/prometheus/prometheus`

Name	Build
Hermes 2 Pro Llama	`https://huggingface.co/NousResearch/Hermes-2-Pro-Llama-3-8B-GGUF`
Llama 2	`https://huggingface.co/YanaS/llama-2-7b-langchain-chat-GGUF`
Llama 3	`https://huggingface.co/lmstudio-community/Meta-Llama-3-8B-Instruct-GGUF`
Mistral 7B	`https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF`
Phi 3.1	`https://huggingface.co/lmstudio-community/Phi-3.1-mini-4k-instruct-GGUF`
Qwen 1.5	`https://huggingface.co/Qwen/Qwen1.5-7B-Chat-GGUF`

Table 1: Huggingface pre-trained LLMs, and the URL where they can be found, used in AUT experiment.

4.3 Designing the responses prompt

We use two general prompts in the experiment, one given to the LLMs to respond to the AUT (the responses prompt), and one to Prometheus to evaluate both LLM and human responses. The prompts were specialised to a specific object by passing it in as a parameter upon presentation to the models.

The prompt to the LLMs used in the experiment is given in figure 3. The prompt was inspired by and synthesised from a number of sources (Guilford et al. [1978], CreativeHuddle [2024], Birss [2024]). The prompt gives some examples of alternate uses of a coffee cup; this is using a technique called in-context learning. It has been shown that models can learn tasks from just a few examples provided in the prompt, without further training or modifying the network's weights Brown et al. [2020], Raventós et al. [2023], Dong et al. [2024]. It has been shown, in Agarwal et al. [2024], that the more examples you give the model, in general, the better the performance; however, in the current work, we are using "off-the-shelf", pretrained LLMs from Huggingface, and so our prompt context is (relatively) limited. We also give the scoring rubric in the prompt, so that the model knows how it will be evaluated, which has been shown to improve performance Lee et al. [2024], Xiao et al. [2024], Zheng et al. [2023].

The prompt is presented to a number of LLMs downloaded from Huggingface. The chosen models are given in table 1. The models chosen were arguably a fair representation of language-competent LLMs available on Huggingface's site, which could run on a modestly-sized GPU on a laptop. Since all the models are open source, such an experiment could be reproduced easily, with little resource requirements or implications.

All the LLMs used in the experiment were quantized to some extent. Where possible, the largest model was chosen consistent with the requirement that it be able to fit into the GPUs memory - so, for example, the Qwen 1.5 model has 5-bit quantization.

5 Bowl and paperclip dataset

The first publicly available AUT dataset we will investigate was made available by Sun et al. [2024], and can be found at `https://github.com/ghydsgaaa/Cambridge-AUT-dataset` (Sun et al. [2023]). The data was collected as part of a larger Cambridge University study on assessing creativity. The alternate uses test focused on two everyday objects, the *bowl* and the *paperclip*. Participants were given 90 seconds to produce as many different uses as possible. Both objects were given in the same set of instructions (see Sun et al. [2024], Section A, for the full instructions). There were 1,297 participants.

Sun et al were exploring the idea of machine evaluators on the AUT, and were therefore interested in using large language models to evaluate the responses (among other potential models, like semantic distance). Their aim was to fine-tune LLMs to replicate, as far as possible, the human evaluators, and to consider the possibility of developing an automated scoring tool for the AUT using supervised machine learning models.

5.1 Sample sizes, quality and scores

The human AUT response dataset contains responses from many individuals. In order to test the *fluency* of the LLMs, the prompt asks for the models to output 40 responses. However, the human responses per individual participant is generally much smaller than this. It would be possible to adjust the prompt to request the mean number of human responses; however, this would require a much larger number of evaluations.

Obviously, the LLM responses will automatically score higher than human responses for fluency, since the models are likely to encounter no problem with producing a large number of responses according to the prompt, irrespective of how they score on other criteria. This would tend to give them an automatic, and (arguably) unfair, advantage. The

Figure 3: The responses prompt: `object` is a parameter to the prompt

I'd like you to take the alternate uses test. This is a classic creativity test used to measure flexibility of thinking. It provides you with a random object . You have to come up with as many alternative uses for it as you can.
For example, if you were given the item 'coffee cup', you might come up with things like:

- a small soup bowl
- a plant pot for some basil
- a scoop for a big barrel of oats
- a hat for an elf in a thunderstorm
- use upside down as a spider trap
- a percussion instrument you hit with a drumstick
- a template to draw a perfect circle

Some of these are more obvious than others. So you can measure your output for both quantity and quality. In fact, the classic test measures the output in four ways:

- Fluency - this is the total number of alternative uses you come up with
- Originality - this is how unique your answers are
- Flexibility - this is the breadth of categories you cover with your ideas
- Elaboration - this is the amount of detail you give in your answers

I just need a list of answers. Using the coffee cup example above, these would be of the form

- Use an upside-down coffee cup as a hat for an elf
- Use a coffee cup as a scoop for a big barrel of oats
- Use a coffee cup as a percussion instrument you hit with a drumstick

... and so on

Generate fluent, original, flexible, and elaborate examples, using the random object given below. Humorous and absurd or surreal examples are welcome!

The random object is:

{object}

Please provide 40 examples.

dataset also provided no mapping from a particular response to the participant, which means that there is no way to collect responses given by an individual participant.

A decision was therefore made to randomly select from the human responses so that they fit into batches of 40. The human and model responses would therefore have the same size, and the same opportunity to display "fluency" (as defined by the scoring rubric); however, given the lack of mapping from participant to response, it was not possible to group strong or weak performers into groups.

After some inspection of the human response dataset, it became apparent that there were quite a few low quality responses (i.e. those which were given a score of less than 3). This is also observed by Sun et al. [2024], who note that the dataset is very imbalanced, where the majority of the responses were rated below 3. It was therefore decided to exclude those responses with a mark below this threshold, since we are interested in comparing the higher-scoring human responses on the AUT with LLM performance. Sun et al also noted that "the inter-rater agreement is not particularly high" (Sun et al. [2024], p. 4), attributing the lack of correlation between scores to the highly subjective perception of what constitutes a creative response in such scoring exercises.

It was decided to use the median score, rather than the mean, for the human AUT responses on the Cambridge dataset; this would have the effect of enhancing the score where two out of the three scores were in agreement, or within a small distance from each other, giving us a better agreement with the majority score.

A similar problem arose when comparing the two scoring rubrics. Both human and machine scorers were using a Likert scale, but had chosen a different starting point for their evaluations [5]. The human evaluators were given a score from 0 to 4, where 0 represents an invalid or irrelevant use, and 4 indicates very high creativity. However, Prometheus has been trained on a scale of 1 to 5. This presented a quandary, since to change the absolute evaluation score used for Prometheus would require the not-inconsiderable task of retraining the model. A much simpler solution would be to change either the human AUT evaluators' response, incrementing the score by 1, or the machine evaluator's score, decrementing it. The choice was made to increment the human AUT scores by 1, then recalculate the mean and the median - in comparison, a trivial exercise.

Note that this will affect the pool of sample sizes we can use, since filtering out responses according to evaluated score affects the number of responses to choose from.

5.2 Experimental protocol

We already have human responses to the alternate uses test (Sun et al. [2023]), and so we need to harvest the language model responses to the prompt. We then need to use our machine evaluator on both human- and machine-responses, and analyse the outputs.

As noted, a consumer laptop with a `NVIDIA GeForce RTX 3060` with 6 GB of memory was used to run the experiment. All models used were downloaded from Huggingface (see Table 1 for where to find them). Since we are looking for high levels of divergent thinking in our outputs, all the LLMs used in the first phase (gathering AUT responses) were set to their maximum temperature. (The evaluator was correspondingly set to its lowest temperature).

The human responses were processed in the following way:

1. to convert between the Likert scales used, the scores were incremented by 1
2. mean and median were recalculated with the incremented scores
3. filtered so that responses with a median < 3 were discarded
4. responses randomly sampled, without replacement, into sets of 40 responses
5. remaining responses were collected into a single, shorter response

Both human and LLM responses were input to the machine evaluator, Prometheus 2.0, as provided by Leonov [2024]. Prometheus is based on the Llama architecture, and has a context length of 32k, and an embedding length of 4k. It consists of 32 layers, with an attention head count of 32, and before quantization consisted of 7B parameters. 6-bit quantization was then applied. The model with the largest quantization that could be hosted on a consumer GPU was chosen, but other flavours are available.

Details of the training and performance improvements of Prometheus 2.0 over the previous version can be found in Kim et al. [2024b].

5.3 Results

The sampled human evaluations are given in table 2. The machine evaluation of the LLMs, also scored by the Prometheus model, are given in table 3.

We used a sample size of 40 for the human responses, to create a batch for the model to evaluate. Since we were filtering out responses with a median score of less than 3, the number of samples between the two objects differ slightly.

Since we are analysing data based on a Likert scale, we should address the issue of whether or not we can utilise certain classes of statistics on the data. According to Sullivan and Artino [2013], use of measures such as the mean, standard deviation or variance on data derived from an ordinal scale caused something of a debate in the statistics literature. The issue is, given the use of a Likert scale, how does one calculate the distance between different responses? Fortunately, Norman analysed a whole raft of both real and simulated datasets using ordinal data, and concluded that the use of parametric statistical measures is justified, and can often lead to more robust results (Norman [2010]). Accordingly, we will use parametric statistics when presenting the results.

From the results, given the experimental protocol outlined in section 5.2, we can see that the LLMs have reached a level comparable to human on the alternate uses test.

[5] The Likert scale is a commonly-used psychometric scale named after its inventor, social psychologist Rensis Likert: see `https://en.wikipedia.org/wiki/Likert_scale`

Bowl		Paperclip	
Sample	**Score**	**Sample**	**Score**
B1	3	P1	3
B2	2	P2	2
B3	2	P3	2
B4	3	P4	2
B5	3	P5	3
B6	4	P6	3
B7	2	P7	2
B8	3		
Mean	4.0		4.125
Std dev.	0.707		0.535

Table 2: Human AUT evaluation on the bowl and paperclip objects. The responses were randomly sampled into groups of 40, with the remainder forming a smaller sample. They were presented to the machine evaluator; these are the outputs. Responses which scored less than 3 were filtered out, so we have different numbers of samples for the different objects.

Bowl	Paperclip	
Model	**Score**	**Score**
Hermes 2 Pro Llama	5	5
Llama 2	3	4
Meta Llama 3	5	5
Mistral 7B	4	5
Phi 3.1	4	5
Qwen 1.5	5	5
Mean	4.333	4.833
Std dev.	0.816	0.408

Table 3: Large language model evaluations on the bowl and paperclip objects.

It could be argued that excluding any human AUT response less than 4 would give a significant boost to the human performance. Similarly, collecting the responses of the best-performing humans, or limiting the number of LLM responses to the mean number from human AUT participants, would likely move the mean score away from the machine scores. On the other hand, it is easy to observe that including all responses from the human participants would move the relative average performances decisively away from human participants.

This would address the *flexibility* criterion - the number of alternative uses for the object, which LLMs are comparatively likely to excel at. It should be noted that the lack of identifiers in the dataset meant that we were unable to group responses together by their individual participant, and therefore unable to identify high-scoring (or low-scoring) individual contributors. Bundling human responses was, in part, inspired by this absence of identifier; it was also a reaction to the "verbosity" of language models, which are likely to be able to create as many responses as one could ask for, within model output limits. Thus, asking for fewer responses would constrain the LLM output to a more human-comparable size. And again, we must recall that many of the human responses had to be filtered out, due to the low scores assigned to many of the human participants.

It could be argued that the best performing humans are likely to outperform even the best LLMs; however, the language models are capable of operating continuously, and it is unlikely that the highest-scoring AUT participants are that much beyond LLM performance. Another factor to consider is that the models used were quantized and running on a consumer GPU, rather than the fully-fledged paid-for services currently available online, which are likely to have higher performance.

6 Fork, book, and tin can dataset

A recent paper used OpenAI's ChatGPT to create responses to the alternate uses test Stevenson et al. [2022b]. The paper analysed the performance of humans and ChatGPT on the alternate uses test. The everyday objects used as prompts for the responses were 'fork', 'book' and 'tin can'.

We will use the dataset of human responses as the input to our own AUT comparative performance experiment. The dataset is publicly available and can be accessed at `https://osf.io/vmk3c/` (Stevenson et al. [2022a]).

6.1 Sample sizes, quality and scores

The human participants were 42 students at the University in Amsterdam, and native Dutch speakers[6]. Both human and LLM responses were judged by human evaluators. The GPT responses were carefully edited to remove any evidence that they were machine generated, to avoid any potential bias.

Responses which were considered invalid or incomplete were excluded from the dataset and analysis presented in the Stevenson study. As described above, when exploring the dataset based on AUT response to the 'bowl' and 'paperclip' set of objects, the decision was taken to filter out some of the results with a median low score below some reasonable threshold. After inspection, the quality of responses obviated the need for further filtering of the dataset.

The previous dataset also demonstrated a rather large disparity in the scoring from the human evaluators, which led to using the median score to create an average to use as a quality filter. The current dataset was found to have a high degree of inter-rater agreement (Stevenson et al. [2022b], p.2), which again implied that any further filtering of the responses was contraindicated. Subsequently, we used the AUT responses from human participants as is.

It was possible to track each response to the alternate uses test to a particular participant. The previous dataset was sampled so that we had a sufficiently large batch of responses for the machine evaluator - one which matched the LLM output. This sampling approach was no longer necessary with the current dataset, which allowed us to preserve the individual responses when presenting them to the machine evaluator.

Stevenson et al's study found that the human responses had a higher originality and surprise rating, but that the LLM was close in performance on both of these aspects. Overall, the human participants were considered to have scored more highly on the alternate uses test: the human mean rating was 2.8 out of 5, and the LLM's mean was 2.6. This work differs from the current paper in its use of human evaluators, and commercial LLMs (GPT-3, in this case). Given that the humans were found to have performed better than the models in that study, it might be interesting to see open-source LLMs compete with OpenAI's model from the recent past.

6.2 Experimental protocol

The same prompts were used used for eliciting responses from the models, and to evaluate both human and LLM responses, as given in figures 1 and 3. The same Likert evaluation scale, 1-5, was output by Prometheus as the human scorers.

We have previously discussed the 'flexibility' attribute which is being assessed, and noted that LLMs by their nature are likely to have an advantage, and how we have attempted to deemphasise this advantage. The current dataset means that we can prompt the LLMs to create responses which are limited to the average number of human responses. The mean length of human responses to each object in the dataset was found to be roughly six, and so the prompt to the models was modified accordingly.

6.3 Results

The results of the experimental run are given in tables 4 and 5. Due to the number of participants, human scores were grouped together rather than given individually, although they were evaluated on an individual basis.

As can be seen from the tables, the model responses to the alternate uses test were evaluated as on average far higher with respect to the given criteria. This result differs from the result in Stevenson et al. [2022b], which had human participants rated slightly above the LLM performance. This discrepancy is perhaps due to the steady progress made by large language models in the last couple of years, where open-source, quantized models can outperform a commercial LLM of that period.

[6]The responses have been translated to English, and are available online Stevenson et al. [2022a], along with the code used to analyse the data and the original Dutch responses

Model name	Mean	Std. dev.
Book		
Hermes-2-Pro-Llama	4.522	0.640
Llama-2	3.315	0.865
Llama-3	3.630	0.800
Mistral-7B	3.086	0.574
Qwen-1.5	3.802	0.762
Fork		
Hermes-2-Pro-Llama	4.378	0.610
Llama-2	2.865	0.908
Llama-3	3.600	0.816
Mistral-7B	3.060	0.523
Qwen-1.5	3.394	0.652
Tin can		
Hermes-2-Pro-Llama	4.622	0.592
Llama-2	3.157	1.016
Llama-3	3.950	0.833
Mistral-7B	3.506	0.749
Qwen-1.5	3.347	0.814
All models	3.661	0.903

Table 4: Large language model evaluations on the book, fork and tin can objects. Mean and standard deviation of the AUT scores are given for each model on each object, then the overall mean and standard deviation for all LLMs on all objects.

Object	Mean	Std. dev.
Book	1.423	0.606
Fork	1.559	0.602
Tin can	1.731	0.590
All objects		
	Mean	Std. dev.
	1.571	0.612

Table 5: Human evaluations on the book, fork and tin can objects. Mean and standard deviation of the AUT scores are given for each object, then the overall mean and standard deviation on all objects.

The mean score over all objects for the large language models was 3.66, versus the human score of 1.57. We can reasonably conclude that, on the alternate uses test, open-source LLMs, locally processed on a consumer GPU, perform considerably better than humans.

7 Conclusion

We have shown that, on the alternate uses test, large language models running on a consumer GPU are broadly equivalent to, if not better than, human levels of performance. The alternate uses test is considered to be associated with creative qualities. This work shows that LLMs are capable of creating work that can be considered, according to the AUT scoring rubric, creative, fluent, and original.

This study has focused on the alternate uses test, but it takes only a little imagination to consider other, related tasks which might be considered creative. For instance, the AUT could be replaced with a task such as generating marketing, advertising or political slogans [7]. LLMs have already been put to work producing short PR or news digests, so such extensions would be trivial.

[7] Some may argue that such tasks require little creativity in the first place.

There are other qualitative tests of creativity, however one defines it, which could be explored, and further comparisons between human and LLM performance possible. Using a different model to evaluate the responses, or possibly using a number of different evaluators, and in turn comparing their performance and discrepancies between the evaluated scores, would likely be a fruitful direction for further research - do LLMs which are trained to evaluate human and machine-generated text vary in their scores? To what extent? What does this tell us about how they have been trained?

Perhaps a more useful research direction would be to ask: how can we push the language models to be more creative? Is there a way to increase the original and divergent creativity of their output? Similarly, can this be done without the resource implications of retraining or fine-tuning the large language model?

Acknowledgements

Thanks to Thurston Park for many stimulating conversations.

References

R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, L. Rosias, S. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. Behbahani, A. Faust, and H. Larochelle. Many-Shot In-Context Learning, 2024. URL https://arxiv.org/abs/2404.11018.

D. Birss, 2024. URL https://davebirss.com/altuses/.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners, 2020. URL https://arxiv.org/abs/2005.14165.

C.-H. Chiang and H.-y. Lee. Can large language models be an alternative to human evaluations? In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870.

CreativeHuddle, 2024. URL https://www.creativehuddle.co.uk/post/the-alternative-uses-test.

Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, and Z. Sui. A Survey on In-context Learning, 2024. URL https://arxiv.org/abs/2301.00234.

Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback, 2023.

M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan. LLM-based NLG Evaluation: Current Status and Challenges, 2024. URL https://arxiv.org/abs/2402.01383.

L. F. Góes, M. Volpe, P. Sawicki, M. Grses, and J. Watson. Pushing GPT's creativity to its limits: Alternative uses and Torrance tests. 2023.

J. Guilford, P. Christensen, P. Merrifield, and R. Wilson. Alternate Uses: Manual of instructions and interpretations. Palo Alto. *CA: Mind Garden*, 1978.

J. Haase and P. H. Hanel. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, 33(3):100066, Dec. 2023. ISSN 2713-3745. doi: 10.1016/j.yjoc.2023.100066. URL http://dx.doi.org/10.1016/j.yjoc.2023.100066.

S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models, 2024a. URL https://arxiv.org/abs/2310.08491.

S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models, 2024b. URL https://arxiv.org/abs/2405.01535.

A. Kozbelt, R. A. Beghetto, and M. A. Runco. Theories of creativity. *The Cambridge handbook of creativity*, 2:20–47, 2010.

S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu. Prompting Large Language Models for Zero-shot Essay Scoring via Multi-trait Specialization, 2024. URL https://arxiv.org/abs/2404.04941.

S. Leonov. Prometheus-7b-v2.0, 2024. URL https://huggingface.co/vsevolodl/prometheus-7b-v2.0-GGUF.

Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, and S. Ma. Leveraging Large Language Models for NLG Evaluation: Advances and Challenges, 2024. URL https://arxiv.org/abs/2401.07103.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

G. Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15:625–632, 2010.

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

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

J. Plucker. *Assessment of creativity*. The Cambridge Handbook of Creativity/Cambridge University Press, 2010.

A. Raventós, M. Paul, F. Chen, and S. Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression, 2023. URL https://arxiv.org/abs/2306.15063.

M. A. Runco and R. S. Albert. *Theories of creativity*, volume 990. Sage Publications London, 1990.

M. A. Runco and G. J. Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

D. K. Simonton. Defining creativity: Don't we also need to define what is not creative? *The Journal of Creative Behavior*, 52(1):80–90, 2018.

C. Stevenson, I. Smal, M. Baas, R. Grasman, and H. van der Maas. Dataset for: Putting GPT-3's Creativity to the (Alternative Uses) Test, 2022a. URL https://osf.io/vmk3c/.

C. Stevenson, I. Smal, M. Baas, R. Grasman, and H. van der Maas. Putting gpt-3's creativity to the (alternative uses) test, 2022b. URL https://arxiv.org/abs/2206.08932.

G. M. Sullivan and A. R. Artino. Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, 5(4):541–542, Dec. 2013. ISSN 1949-8349. doi: 10.4300/jgme-5-4-18. URL http://dx.doi.org/10.4300/JGME-5-4-18.

L. Sun, H. Gu, R. Myers, and Z. Yuan. Cambridge-aut-dataset, 2023. URL https://github.com/ghydsgaaa/Cambridge-AUT-dataset.

L. Sun, H. Gu, R. Myers, and Z. Yuan. A New Dataset and Method for Creativity Assessment Using the Alternate Uses Task. *Intelligent Computers, Algorithms, and Applications*, page 125–138, 2024. ISSN 1865-0937. doi: 10.1007/978-981-97-0065-3_9. URL http://dx.doi.org/10.1007/978-981-97-0065-3_9.

Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, W. Ye, S. Zhang, and Y. Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization, 2024. URL https://arxiv.org/abs/2306.05087.

R. W. Weisberg. *Creativity: Understanding innovation in problem solving, science, invention, and the arts*. John Wiley & Sons, 2016.

C. Xiao, W. Ma, Q. Song, S. X. Xu, K. Zhang, Y. Wang, and Q. Fu. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs, 2024. URL https://arxiv.org/abs/2401.06431.

Y. Yang, Z. Li, Q. Dong, H. Xia, and Z. Sui. Can Large Multimodal Models Uncover Deep Semantics Behind Images?, 2024. URL https://arxiv.org/abs/2402.11281.

Y. Zhao, R. Zhang, W. Li, D. Huang, J. Guo, S. Peng, Y. Hao, Y. Wen, X. Hu, Z. Du, Q. Guo, L. Li, and Y. Chen. Assessing and Understanding Creativity in Large Language Models, 2024a. URL https://arxiv.org/abs/2401.12491.

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

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. URL https://arxiv.org/abs/2306.05685.