Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score

Douglas Summers-Stay,[1] Stephanie Lukin,[1] Clare Voss [1]

[1]U.S. Army Research Lab
douglas.a.summers-stay.civ@army.mil

Abstract

Creative problem solving is a crucial ability for intelligent agents. A common method that individuals or groups use to invent creative solutions is to start with a "brainstorming" phase, where many solutions to a problem are proposed, and then to follow with a "selection" phase, where those solutions are judged by some criteria so that the best solutions can be selected. Using the Alternate Uses Task, a test for divergent thinking abilities (a key aspect of creativity) we show that when a large language model is given a sequence of prompts that include *both* brainstorming and selection phases, its performance improves over brainstorming alone. Furthermore, we show that by following this paradigm, a large language model can even achieve higher than average human performance on the same task. Following our analysis, we propose further research to gain a clearer understanding of what counts as "creativity" in language models.

Introduction

It has been argued that "the ability to solve problems is not just an aspect or feature of intelligence – it is the essence of intelligence." (Sternberg 2020) Creatively solving problems as they arise is necessary for intelligent agents to take action in any real-world situation that is constantly in flux. However, in order for a solution to be considered creative, it must be both new and nonobvious, as well as be useful and effective at solving an actual problem (Simonton 2012; Diedrich et al. 2015). These criteria present something of a conundrum for neural networks, however, because anything truly novel that is generated must be something the network has *not* seen in training. Yet somehow the network must have built a model whose generalization includes the potential for these solutions.

Deep neural generative models are much more successful at generalization than previous approaches, which often relied on combinatorial tricks to come up with "new" solutions (Summers-Stay 2011). Large neural language models have been applied to an enormous variety of creative writing tasks, including writing in the style of a named author on a given theme, writing jokes, interactive storytelling and role-playing, creating fictional interviews with celebrities, and many others (e.g. Branwen 2020). Despite the rise in designing models for these creative applications, quantitative measures of the generated creativity proper have been less common. Poetry generation was evaluated for its ability to evoke particular emotions and to use the language of dreams (Bena and Kalita 2020). Sawicki et al. (2022) compared fine-tuned GPT models using BERT classifiers to see which were better at preserving style without plagiarizing. In just the last few months, visual artists have begun to feel that neural image generation techniques trained on their work are capable of imitating their style to an uncomfortable degree. This has caused a great deal of discussion regarding to what extent such models are capable of creative generation (e. g. Salkowitz 2022), or may go beyond to evaluation of their own generated results. Indeed, evaluation is considered the most difficult of the skills that can be taught to children, due to its complexity and abstractness, as elucidated by Bloom (1956). As such, it also poses a particular challenge for language models. In short, it has become increasingly important to characterize both the capabilities and the limitations of such models on creative problem-solving tasks.

One such challenging creative generation task is the Alternate Uses Task (AUT), a test commonly used for divergent thinking ability, a key aspect of creativity (Guilford 1964). It was created as a means to measure and compare human creativity. The test taker is presented with the name of a common object, such as "pencil," and is asked to name as many and varied uses as possible within a time limit. In the original test, answers were scored on "fluency (overall sum of generated uses), originality (statistical infrequency of generated uses), flexibility (number of conceptual categories within which uses could be binned), and elaboration (degree of detail and richness in a response), amongst others" (Vartanian et al. 2020). Such measures proved hard to apply consistently, and more recently the test has been scored with a numerical assessment of novelty and usefulness which are thought to be the necessary and sufficient conditions for a response to be creative (Diedrich et al. 2015).

A recent study by Stevenson et al. (2022) explored using GPT as a problem-solver for the AUT. Their study used various settings of the largest GPT-3 model, Davinci, to solve the AUT for three objects (a book, a tin can, and a fork), generating a total of 690 responses, and additionally collected human-authored responses for the same set of objects. They then sanitized the GPT-3 responses to remove any telltale signs of its origin, e.g., removing numbered responses, and created a pool mixing the GPT-3 and human solutions to the AUT. This complete set was scored by trained raters for utility, originality, and surprise on a 1-5 scale for each quality. In this way, the experimenters were able to consistently compare GPT generated responses with the human-authored ones on these measures. Their results show that GPT scored *somewhat higher* than the human average on utility, and *slightly lower* on originality and surprise, among other findings.

Our objective in this paper is to assess GPT's creative problem-solving abilities for the AUT, using a human-inspired approach to creative problem-solving. A common method that individuals or groups rely on for coming up with creative solutions is to start with a "brainstorming" phase, where many solutions to a problem are proposed, and then to follow with a "selection" phase, where those solutions are judged by some criteria so that the best solutions can be selected. Generating a large number of candidates and choosing those that best fit some selection criteria has been used with large language models for solving math problems (Cobbe et al. 2021). We apply this to the task of generating creative uses: given a sequence of prompts that include *both* brainstorming and selection phases, the language model can improve its performance over just brainstorming alone. To the best of our knowledge, this paper is the first to show that with the right arrangement of prompts, a neural language model is capable of solving a standard test of human creativity, in particular, the AUT, at a level beyond the human average.

Methods

The objective of Stevenson et al. (2022) work was to explore how well GPT-3 was able to generate novel and useful responses when given similar instructions to humans. Our goal in this paper was somewhat different. We attempted to find a chain of prompts that gave whatever assistance we thought would be effective in improving GPT-3's performance on the task. Our system builds on Stevenson et al.'s GPT-3-powered divergent thinking phase ("brainstorming") with the next phase where GPT-3 powered convergent thinking selects from among these potential uses the ones that are most promising in terms of utility and originality.

Table 1 shows the prompt used in Stevenson et al. (2022), where '[object] is filled in by the object selected for the AUT. From this, GPT-3 generated solutions, such as, for a tin can: "1. Use a tin can as a mirror. 2. to create toys. 3. to create jewelry" (Stevenson et al. 2022).

What are some creative uses for [object]? **The goal is to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different.** **List 10 creative uses for** [object].

Table 1: Initial Prompt Format from Stevenson et al. (2022)

Stevenson's GPT-3 approach with this prompt generated 690 responses. In our work to test whether GPT-3 can, in effect, judge and improve on the utility and novelty of the text it generates, we adopt these Stevenson-responses treating them as the product of the brainstorming phase, to be incorporated into our prompts in the selection phase (as described below). We note that, while Stevenson rated the 'surprisal' of the generated responses, we found it difficult to draw a distinction between originality and surprisal, and therefore did not use those score ratings in our study. We do, however, in our prompt to elicit originality, use the word "surprised" as a way of conveying the general notion of novelty.

Improving Utility

On a scale of 1 to 5 measuring utility, only 64 of the 690 Stevenson-responses were given a utility score of 1 (useless) by human raters. If the remaining generated solutions were genuinely novel (i.e. not seen in training), how was the model able to recognize the potential for using a book, tin can, or fork in a completely new situation?

Large neural language models do not contain explicit physics models to simulate what might happen if, for example, a book were used as a shovel or an umbrella. They have, however, been trained on examples in literature of books being used to swat flies, or to press leaves, or to impress a girlfriend. They seem to be able to generalize from these cases, for example, by keeping the same action but choosing a semantically related object of similar size and features, and so will suggest using a book to kill a snake, or to press flowers, or to impress one's boss. These models seems to have some ability to find new possibilities which are both semantically similar, and situationally appropriate— they are not just nearby in semantic space, but nearby in an appropriate direction that preserves utility.

This generation process creates a wide distribution of possibilities, some of which were judged by humans to be of low utility. Moving into the selection phrase, we want to now reject low-scoring options. To achieve this, we utilize a multi-step reasoning that we split into three parts. To lead GPT-3 to consider what really makes a solution useful or useless, we ask it to list the advantages of the proposed object utility and the drawbacks of using the proposed object in such a way, and then to make a final evaluation that weights both the advantages and drawbacks together to determine if the advantages outweigh the disadvantages of the proposed object utility.

Because GPT-3 has limited multi-step reasoning capability, we break the problem into smaller, more easily solved parts that can be solved one-by-one. This allows the full use of its resources on one sub-problem at a time. A distinct prompt is constructed for the advantages, drawbacks, and evaluation steps, and is shown in Table 2 (each step is separated by a line break). Although we varied the prompts on a few examples to make sure the results were valid, we have no way of knowing whether an untried prompt may perform better at the task, so this may serve as a lower bound.

After GPT-3 has completed the evaluation prompt, in cases where the " Yes" token is significantly more likely than the " No" token as the first token of the evaluation prompt response, we consider the response to be of high utility. We use this difference between the log probability of the answers as

Advantages prompt: **Q: Name one or more advantages to using a** [object] **for the following purpose:** [purpose]? **A:**
Drawbacks prompt: **Q: Name one or more drawbacks to using a** [object] **for the following purpose:** [purpose]? **A:**
Evaluation prompt: **Advantages:**[advantages] **Drawbacks:** [drawbacks] **Q: Based on these advantages and drawbacks, do you think using a** [object] **for the purpose** [purpose] **is a good idea? Answer Yes or No. A:**

Table 2: Prompt Sequence for Improving Utility. Words in square brackets are variables substituted into the prompts.

a proxy measure for confidence; the greater this difference, the more we consider GPT to be "confident" that one answer is better than the other.

Table 3 shows a full example of the utility sequence, starting with Stevenson et al.'s brainstorming prompt and a generated response, and our subsequent selection process through prompt chaining. In each block, GPT-3's responses are in italics. In the segment from Stevenson's prompt and response, to save space, we only show one of GPT-3's responses in italics. That response is in turn substituted into our 'advantage', 'drawback' and 'evaluation' prompts, and the final 'evaluation' has both the responses from 'advantages' and 'drawbacks' (the text in regular type signifies that it was taken from the previous prompt).

Improving Originality

Because generative large language models try to predict the most likely next token, it seems paradoxical that one could ever produce original responses. Part of the answer is that given a prompt with words like "clever, unusual, interesting, uncommon, humorous, innovative, or different", there are many likely next tokens. GPT3 has picked up on these patterns, and sampling from among these with high temperature yields a variety of responses. In that context, the standard use for an object is actually unlikely to be generated, because it would not frequently be seen after a sentence describing the response with those kinds of adjectives. GPT-3 struggles with multi-step deductive reasoning (Bao, Witbrock, and Liu 2022), but is fairly good at analogical reasoning (Ushio et al. 2021). We speculate that the lists of creative alternate uses GPT has seen in training may be used in forming analogies to come up with alternate uses. For example, if it has seen a creative use of a glass bottle is "include in a stained glass window", it may implicitly form an analogy such as:

bottle : stained glass :: tin can : ???

to come up with the answer "embossed tin art." We know precisely (mathematically) how such analogies are formed in word embeddings such as word2vec (Mikolov et al. 2013). Any high-dimensional representation of concepts in which similar concepts are nearby will have an arrangement of concepts in which forming analogies of this kind requires only trivial arithmetic (Summers-Stay 2017).

What are some creative uses for a tin can? The goal is to come up with creative ideas, which are ideas that strike people as clever, unusual, interesting, uncommon, humorous, innovative, or different. List 10 creative uses for a tin can.
Make a wind chime out of tin cans and metal spoons

Q: Name one or more advantages to using a tin can for the following purpose:
Make a wind chime out of tin cans and metal spoons.
A: *Some advantages of using a tin can for a wind chime include that it is lightweight and easy to work with.*
Tin cans can also produce a pleasing sound when struck, making them ideal for this purpose.

Q: Name one or more drawbacks to using a tin can for the following purpose:
Make a wind chime out of tin cans and metal spoons.
A: *One drawback to using a tin can for a wind chime is that it can produce a fairly harsh, metallic sound.*
Additionally, tin cans are not particularly durable, so the wind chime may not last very long.

Advantages: Some advantages of using a tin can for a wind chime include that it is lightweight and easy to work with. Tin cans can also produce a pleasing sound when struck, making them ideal for this purpose.
Drawbacks: One drawback to using a tin can for a wind chime is that it can produce a fairly harsh, metallic sound. Additionally, tin cans are not particularly durable, so the wind chime may not last very long.
Q: Based on these advantages and drawbacks, do you think using a tin can for the purpose 'Make a wind chime out of tin cans and metal spoons' **is a good idea? Answer Yes or No. A:** *No*

Table 3: Full Example with Prompt Chaining

Moving from the brainstorming phrase into the selection phase of novel responses, we construct an evaluation prompt that gives the model a chance to reject a response as unoriginal, where, for example, the use of adjectives in the proposed object use may be biased towards considering it an original answer. The prompt is shown in Table 4.

After GPT-3 has completed the evaluation prompt, we again use the log probability as a proxy measure for confidence in determining if the response is of high originality.

Results

Table 5 shows the results of applying our process of using new prompts to serve as filters to the Stevenson-responses. By re-using GPT-3 to select only those answers which it judges to be original, useful, or both, we can improve the score. Their paper notes that Rietzschel et al. (2019) commented that originality and utility are trade-offs, and we

Q: If someone suggested using a [object] for the following purpose: [purpose], **would you be surprised and think it was a novel idea? Answer Yes or No. A:**

Table 4: Evaluation Prompt for Improving Originality. Words in brackets are variables substituted into the prompts.

found that to be true with these filters as well. If we select only responses that GPT-3 is confident are original, then the rated *originality* improves from 2.3 to 3.0 but the rated *utility* drops from 4.2 to 3.3. If we select only those responses that GPT judges confidently to be useful, we can improve the *utility* score from 4.2 to 4.3 without reducing *originality* much. If we choose only those responses which GPT decides are both original and useful with high confidence, then *originality* improves but *utility* drops slightly, although both *originality* and *utility* are still above the averaged ratings for human responses.[1]

	number	*originality*	*utility*
human	823	2.6	3.7
GPT3 (no filtering)	690	2.3	4.2
GPT3 (originality filter)	119	**3.0**	3.3
GPT3 (utility filter)	634	2.3	**4.3**
GPT3 (filter both)	39	2.7	4.1

Table 5: Results of filtering on ratings for *originality*, *utility*, or both. "Number" refers to how many responses out of 690 are left after the filtering. These are all the results for all three objects (book, fork, and tin can) summed together. The 823 responses are from 42 students and the 690 GPT-3 (no filtering) responses were generated for that study (Stevenson et al. 2022). The full data from this earlier study can be found here: https://github.com/cstevenson-uva/creAI-gpt3/tree/master/aut_pilot_ICCC22 Our results in this paper are shown below the dividing line.

Sample Outputs of Suggested Advantages and Drawbacks

We posited that the prompts for generating advantages and drawbacks would help bias the model toward a higher or lower probability of 'Yes' vs. 'No' when answering whether a solution is useful or novel. However, the prompts' purpose in our experimentation is only to serve as additional information for a final evaluation of utility. The responses themselves, however, are interesting in their own right for what they show about how the system is able (or unable) to reason about unusual situations.

Some of the advantages and disadvantages named are surprisingly good. Table 6 contains a few examples of using a book. Notice in the first three selected results what properties of a book have been implicitly stored in the weights. However, when the suggested use is completely ridiculous, the model still comes up with advantages and never simply rejects an idea as impossible or without redeeming features. The second set of three results shows that the advantages and disadvantages it generates can be very silly.

[1]The time-limit constraint probably kept humans from doing much of this kind of filtering on their own responses. GPT-3 has its own constraints, such as prompt length, so a direct comparison is difficult.

- *holding open a window: the book may become damaged from exposure to the elements.*
- *as a cutting board: it can be difficult to keep the book clean. If you cut meat or vegetables on a book, the book can become stained and difficult to clean.*
- *to make coasters: you can choose a book with a design or theme that you like, which can make your coasters more personal and unique. Additionally, using a book is a great way to upcycle and repurpose an old book.*

- *using a book for the purpose of canning: First, it is difficult to sterilize a book. Second, a book is not an airtight container, so it is not ideal for preserving food.*
- *using a book for making a paper airplane: it can be difficult to get a clean fold. Also, if the book is too thick, the airplane may not fly well.*
- *using a book for a clock: it is a relatively inexpensive way to make a clock. Additionally, it is a relatively easy way to make a clock, as all you need to do is glue numbers to the cover and attach hands.*

Table 6: Selected generated Advantages and Disadvantages

Discussion and Future Work

This paper has presented a paradigm to encourage creativity in GPT-3 by first brainstorming and then selecting the best answers according to their novelty and usefulness.

We seek to better understand in what other contexts the generation of possible solutions and evaluation can be successfully applied. While GPT-3 was successful at the AUT, this may not be the case for other problems requiring multiple novel steps to be carried out in succession.

A lingering question remains: to what extent has the model simply copied examples of "creative uses" it has already seen? One of the model's worse suggestions was to use a fork to "build a fort." Although the idea makes little sense for a fork, the idea of building a fort is associated with creative play in general. Perhaps the generated suggestions have more to do with the concept of "creativity" as learned by GPT from the training data in general than from true invention. One way to explore this would be to rerun the experiment without any mention of novelty or creativity in the original prompt. Without bringing in examples it has seen from other "creative" contexts, could GPT-3 still come up with answers that are both novel and useful?

We also hope to combine our approach with a code-generating model that converts the natural language description of a solution into a working program that a machine such as a robot could then carry out. Researchers are already exploring how knowledge stored in a pretrained large language model (LLM) can be transferred to robot task planning by way of prompt structures. (Singh et al. 2022) have demonstrated LLMs can be prompted with program-like specifications for available actions and objects, as well as with executable example programs, to generate next actions.

Writing in 2013, (Frey and Osborne 2017) analyzed professions to determine which were at risk of automation. They concluded that creative tasks were unlikely to be automated in the next few decades. Given the rapid advances in creative generation in many different fields over the last few years, however, that conclusion now seems hopelessly outdated.

References

Bao, Q.; Witbrock, M.; and Liu, J. 2022. Natural Language Processing and Reasoning.

Bena, B.; and Kalita, J. 2020. Introducing aspects of creativity in automatic poetry generation. *arXiv preprint arXiv:2002.02511*.

Bloom, B. S. 1956. Taxonomy of educational objectives: The classification of educational goals. *Cognitive domain*.

Branwen, G. 2020. GPT-3 Creative Fiction.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Diedrich, J.; Benedek, M.; Jauk, E.; and Neubauer, A. C. 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1): 35.

Frey, C. B.; and Osborne, M. A. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114: 254–280.

Guilford, J. P. 1964. Some new looks at the nature of creative processes. *Contributions to mathematical psychology. New York: Holt, Rinehart & Winston*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rietzschel, E. F.; Nijstad, B. A.; and Stroebe, W. 2019. Why great ideas are often overlooked. *The Oxford handbook of group creativity and innovation*, 179–197.

Salkowitz, R. 2022. AI Is Coming For Commercial Art Jobs. Can It Be Stopped? *Forbes*.

Sawicki, P.; Grzes, M.; Jordanous, A.; Brown, D.; and Peeperkorn, M. 2022. Training GPT-2 to represent two Romantic-era authors: challenges, evaluations and pitfalls. In *13th International Conference on Computational Creativity*.

Simonton, D. K. 2012. Taking the U.S. Patent Office Criteria Seriously: A Quantitative Three-Criterion Creativity Definition and Its Implications. *Creativity Research Journal*, 24(2-3): 97–106.

Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; and Garg, A. 2022. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In *NeurIPS 2022 Second Workshop on Language and Reinforcement Learning*.

Sternberg, R. J., ed. 2020. *Intelligence and Creativity*, 1087–1105. Cambridge Handbooks in Psychology. Cambridge University Press, 2 edition.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's Creativity to the (Alternative Uses) Test. *arXiv preprint arXiv:2206.08932*.

Summers-Stay, D. 2011. *Machinamenta: The Thousand Year Quest to Build a Creative Machine*. CreateSpace.

Summers-Stay, D. 2017. Deductive and analogical reasoning on a semantically embedded knowledge graph. In *International Conference on Artificial General Intelligence*, 112–122. Springer.

Ushio, A.; Espinosa-Anke, L.; Schockaert, S.; and Camacho-Collados, J. 2021. BERT is to NLP what AlexNet is to CV: can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*.

Vartanian, O.; Smith, I.; Lam, T. K.; King, K.; Lam, Q.; and Beatty, E. L. 2020. The relationship between methods of scoring the alternate uses task and the neural correlates of divergent thinking: Evidence from voxel-based morphometry. *NeuroImage*, 223: 117325.