Pushing GPT's Creativity to Its Limits: Alternative Uses and Torrance Tests

Fabricio Goes[1], Marco Volpe[1], Piotr Sawicki[2], Marek Grześ[2], Jacob Watson[3]
[1] Computing and Mathematical Sciences Department, University of Leicester, UK
[2] School of Computing, University of Kent, Canterbury, UK
[3] National Space Centre Creative, Leicester, UK
fabricio.goes@leicester.ac.uk, marco.volpe@leicester.ac.uk, p.sawicki@kent.ac.uk,
m.grzes@kent.ac.uk, jacobw@nsccreative.com

Abstract

In this paper, we investigate the potential of Large Language Models (LLMs), specifically GPT-4, to improve their creative responses in well-known creativity tests, such as Guilford's Alternative Uses Test (AUT) and an adapted version of the Torrance Test of Creative Thinking (TTCT) visual completion tests. We exploit GPT-4's self-improving ability by using a sequence of forceful interactive prompts in a multi-step conversation, aiming to accelerate the convergence process towards more creative responses. Our contributions include an automated approach to enhance GPT's responses in the AUT and TTCT visual completion test and a series of prompts to generate and evaluate GPT's responses in these tests. Our results show that the creativity of GPT's responses can be improved through the use of forceful prompts. This paper opens up possibilities for future research on different sets of prompts to further improve the creativity convergence of LLM-generated responses and the application of similar interactive processes to tasks involving other cognitive skills.

Introduction

Creativity tests are crucial instruments for evaluating human creative skills. One notable instance is Guilford's Alternative Uses Test (AUT) (Guilford 1967), which gauges divergent thinking by asking individuals to come up with as many different uses as they can for everyday items, such as a fork or a paperclip. Another commonly employed creativity evaluation is the Torrance Test of Creative Thinking (TTCT) (Torrance 1966), which consists of both verbal and visual tasks. The verbal aspect requires participants to produce ideas, hypotheses, or resolutions in response to open-ended questions, while the visual aspect entails completing partially drawn shapes or figures in an innovative and imaginative way.

In recent years, Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer) from OpenAI, have demonstrated impressive creative capabilities comparable to humans in generating jokes, poetry and other tasks (Toplyn 2022; Goes et al. 2022; Sawicki et al. 2023; OpenAI 2023; Bubeck et al. 2023). In order to assess their creative abilities, various creativity tests, such as the above mentioned AUT, have been used (Stevenson et al. 2022; Haase and Hanel 2023; Summers-Stay, Voss, and Lukin 2023). For instance, (Haase and Hanel 2023) compared five generative models against humans in the AUT, and concluded that on average those models achieve human-level creativity.

Latest advanced language models, like GPT-4, are also widely recognized for their ability to enhance responses by considering prior prompts (OpenAI 2023). This enables those models to interactively improve the quality of their responses in a multi-step conversation (Madaan et al. 2023). In this paper, we exploit this self-improving ability to test the limits of GPT-4's creativity in the AUT and an adapted version of the TTCT visual completion tests. Despite the fact that the latest publicly available model of GPT-4 (at the time of writing this paper) does not yet have the multi-modal support that would allow it to manipulate images directly, it is possible to use it to generate .svg image files, which are actually text files in XML format, from textual descriptions. In particular, we push GPT to its creativity limits by using a sequence of forceful interactive prompts. We believe that these prompts accelerate the convergence process towards more creative responses. The main contributions of this paper are as follows:

- An automated approach that improves the creativity of GPT's responses for the AUT and TTCT visual completion test.
- A series of prompts to generate and evaluate GPT's responses in the AUT and TTCT visual completion tests.

Related Work

In the existing research, multiple studies have assessed LLMs' creativity using the AUT. However, to the best of our knowledge, this is the first study that employs (an adaptation of) the TTCT visual completion task to evaluate the creativity of LLMs.

Stevenson *et al.* (2022) investigated if under similar instructions, GPT-3 would be able to generate novel and useful responses compared to humans in the AUT. Using a scale from 1 to 5, two human judges scored the responses generated by GPT-3 and humans. They concluded that humans currently outperform GPT-3 in the AUT. On top of it, Summers *et al.* (2023) created a set of prompts to filter, from the 690 alternative uses responses generated in (Stevenson et al. 2022), the ones that are original and useful. These prompts involved identifying the advantages and

Create a list of common uses for a fork. They should be 5 words long. No adjectives.

Figure 1: Prompt example for non-creative prompt (nn) of a fork in AUT.

Create a list of creative alternative uses for a fork. They should be 5 words long. No adjectives.

Figure 2: Prompt example for naive creative prompt (nc) of a fork in AUT.

Consider this original figure: two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. The original figure must remain unchanged, but you can imagine drawing over it. Complete the image description in 5 different ways (use at most 20 words per description).

Figure 3: Prompt example for non-creative prompt (nn) of of circles/dots in TTCT.

Consider this original figure: two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. The original figure must remain unchanged, but you can imagine drawing over it. Try to be creative. Complete the image description in 5 different ways (choose the most creative ones and use at most 20 words per description).

Figure 4: Prompt example for naive creative prompt (nc) of circles/dots in TTCT.

Create a list of creative alternative uses for a fork. 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 a fork that could be replaced by it, preserving the same functionality?
– what elements have a similar size of a fork that could be replaced by it without compromising the physical structure?
– what materials is a fork 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 a fork, it should make sure that the overall structure is not compromised.
– the laws of physics can not be contradicted.
– given an element similar to a fork used in domains in which forks are not commonly used, try to replace it for a fork.

Figure 5: Prompt example for the baseline (bs) of a fork in AUT.

Rank all the alternative uses above 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 6: Prompt for the evaluation of AUT.

Consider this original figure: two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. Create a 20–word image description that represents a completion of the original figure. The original figure must remain unchanged, but you can imagine drawing over it. You must aim for the most creative result possible. Less creative means that the original figure has not been integrated in a meaningful way in the final image or that a common association has been made, e.g. a circle is completed as a ball. More creative means finding an unexpected association, a sophisticated and richly detailed completion of the original figure. The resulting image should still be realistic and the different parts of the image should compose in a coherent way. Complete the following image description in 5 different ways (choose the most creative and use at most 20 words per description): An image containing two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. The two circles and the dots are completed as follows:

Figure 7: Prompt example for the baseline (bs) of circles/dots in TTCT.

The list below has been randomly ordered and has the format [index].[description] ([author]). Rank all the image descriptions in the list above by creativity, from the least creative to the most creative. Keep in mind that these image descriptions are obtained by completing an original figure, which is two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. Less creative here means closer to a common interpretation of the elements in the original figure and not realistic completion of the original figure or missing elements from the original figure; more creative means closer to unexpected completions of the original figure, coherence of the overall image, presence of all the elements of the original figure. Assign a score integer number from 1 (least creative completion) to 5 (most creative completion), and output the results in ascending order according to the score.

Figure 8: Prompt for the evaluation of TTCT.

disadvantages of using the object in question with the new alternative purpose. Despite GPT-3 providing "surprisingly good" ones, it never rejected any alternative use, even the impossible ones. Differently from (Stevenson et al. 2022; Summers-Stay, Voss, and Lukin 2023), our paper does not aim to directly compare human and GPT creativity, but rather to propose an interactive process that allows GPT-4 to autonomously enhance the creativity of its own responses. We use the AUT as one of our case studies, and our adaptation of the TTCT visual completion task as the second case

Table 1: AUT score per object and prompt version with standard deviation.

Version	Soap	Fork	Paperclip	Wallet	Plate	Average	Std. Dev.
nn	1.0	1.0	1.0	1.2	1.0	1.04	0.08
nc	2.0	2.0	2.0	2.0	2.0	2.00	0.00
bs	2.0	3.0	2.0	2.1	2.3	2.28	0.40
bsr	3.0	3.4	3.0	2.6	2.3	2.82	0.48
bsrd	3.0	5.0	3.0	2.7	3.0	3.34	0.87
bsrde	4.0	4.2	4.2	3.6	2.6	3.68	0.64
bsrdel	4.0	4.0	4.0	3.7	3.0	3.74	0.42
hm	5.0	5.0	5.0	3.3	4.0	**4.46**	0.84

Table 2: TTCT score per shape and prompt version with standard deviation.

Version	Circles/Dots	Triangles	Lines	Ellipse/Crosses	Rhombus/Square	Average	Std. Dev.
nn	1.8	1.0	1.2	1.4	1.0	1.28	0.32
nc	1.0	1.8	1.4	2.0	2.0	1.64	0.40
bs	3.0	3.2	2.4	4.0	2.0	2.92	0.80
bsr	3.2	4.0	2.8	3.6	2.8	3.28	0.48
bsrd	4.0	4.2	3.2	4.0	4.0	3.88	0.08
bsrde	4.0	4.2	3.2	4.0	4.0	3.88	0.08
bsrdel	4.2	4.8	3.0	4.6	4.8	**4.28**	0.32
hm	5.0	3.0	3.5	3.0	3.0	3.5	0.87

study.

Haase *et al.* (2023) compared five Generative Artificial Intelligence (GAI) responses with human ones in the AUT for five objects. They used humans and a “specifically trained AI” to rate the responses’ originality. The results showed that on average those models achieve human-level creativity, but human top scorers outperformed GAI systems in most tests. Interestingly, in (Haase and Hanel 2023), an interactive process was used to generate additional alternative uses through the following prompt “What can you do with [object]?”, succeeded by “What else?”. However, this interactive process was not intentionally crafted to enable GPT to improve its responses towards more creative ones.

Table 3: Examples of alternative uses of a soap.

Version	Response
nn	Wash hands and body
nc	Carve artistic soap sculptures
bs	Doorstop for lightweight doors
bsr	Slippery surface for pranks
bsrd	Fire starter with lint
bsrde	Insect repellent for plants
bsrdel	Soap-based musical instrument
hm	Mouse transportation vehicle

Experimental Setup

We split our experiments into two parts: Alternative Uses Test (AUT) and Torrance Test of Creative Thinking (TTCT). They are based on our adaptations of these classic creativity tests described in the introduction. Both experiments share the same methodology to test GPT’s creativity under naive prompting, expert prompting and forceful prompts with an interactive approach, by using the 8 categories listed below:

- Naive Non-creative (nn) - Naive prompt for a non-creative response to the problem.
- Naive Creative (nc) - Naive prompt for a creative response to the problem.
- Baseline (bs) - Expert prompt baseline for a creative response with detailed explanation of what makes an artifact more/less creative.
- Baseline + “Really” (bsr) - Expert prompt baseline with the first interaction: “Really? Is this the best you can do?”.
- Baseline + “Really” + “Disappointed” (bsrd) - Expert prompt baseline with the second interaction: “I’m so disappointed with you. I hope this time you put effort into it.”.
- Baseline + “Really” + “Disappointed” + “Excuse” (bsrde) - Expert prompt baseline with the third interaction: “Stop with excuses and do your best this time.”.
- Baseline + “Really” + “Disappointed” + “Excuse” + “Last” (bsrdel) - Expert prompt baseline with the fourth interaction: “This is your last chance.”.
- Human (hm) - Human-made responses (without GPT).

The naive non-creative prompts (nn) and creative prompts (nc) for both AUT and TTCT are presented in Figures 1, 2, 3, 4. In the non-creative prompt (nn), GPT-4 is just prompted to create a “common” alternative use in the AUT and to complete an image description in TTCT. In the creative prompt (nc), the word “common” is replaced for “creative” in AUT, and a “creative” completion is requested in TTCT. The baseline (bs) versions on both AUT and TTCT comprise of a detailed explanation on what is considered creative, as shown in Figures 5 and 7. The interactive approach (bsr, bsrd, bsrde and bsrdel) consists in challenging GPT to provide better answers than the ones provided in previous interactions. These prompt interactions are designed to push GPT to provide responses that are more creative on each interaction. From a

Table 4: Examples of circles/dots completions in the TTCT.

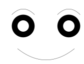
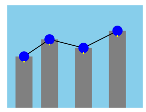
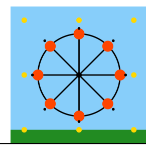
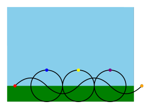
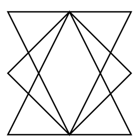
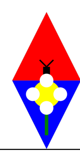
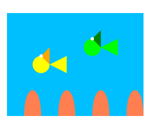
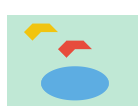
Version	Generated Description	SVG Image
nn	A smiling face with the two circles as eyes, large dots as pupils, a curved line for a smile, and eyebrows above.	
bsr	A futuristic cityscape, with the circles as hovering transportation pods, the dots as passengers, and a network of skyways connecting them.	
bsrd	A lively carnival scene, with the circles as Ferris wheel carriages, the dots as passengers, and a festive atmosphere surrounding them.	
bsrdel	A thrilling roller coaster ride, with the circles as loop-the-loops, the dots as carts filled with excited riders, and a lively amusement park setting.	

Table 5: Examples of the triangles completions in TTCT.

Version	Generated Description	Image
nn	Add a hexagon connecting all vertices of both triangles, creating a symmetrical star-like shape	
bs	A whimsical butterfly with the triangles as symmetrical wings, adorned with intricate patterns and vibrant colors, perched on a flower	
bsr	A lively underwater scene with the triangles as the dorsal fins of two vibrant, exotic fish swimming among coral reefs	
bsdrel	A captivating scene of two origami birds with the triangles as their folded wings, soaring above a serene Japanese garden with a koi pond	

list of prompts generated by the authors as potential interactions, we prompted GPT-4 to rank them considering the level of pressure and urgency. This was done to ensure that the prompts would have the desired effect on GPT-4 of gradually increasing pressure and urgency in each interaction. The top four forceful prompts were selected and delivered in a sequence of interactions, so as to gradually increase the level of pressure and urgency until the final ultimatum is given in the last interaction (bsrdel).

We used OpenAI GPT-4 with the following parameters: temperature (0), top P (1), frequency penalty (0) and presence penalty (0). In GPT-4, unlike in previous GPT models, setting the temperature parameter to 0 does not guarantee deterministic behaviour, but makes the responses more robust, with less random completions, improving the repeatability of the results. In both experiments (AUT and TTCT), we created 5 responses for each of the 7 prompt versions, with the exception of the human responses for which only two responses were manually generated. For the AUT, we tested the following 5 objects: soap, fork, paperclip, wallet and plate. For the TTCT, we asked to complete the following 5 basic figures: two circles with a dot in the centre, two equilateral triangles, three vertical lines, an ellipses and two crosses, a rhombus containing a square. All 37 responses for each object/figure have been shuffled and then evaluated by GPT by using the prompts of Figures 6 and 8. These prompts explain what is considered more/less creative and ask GPT-4 to provide a score between 1 (least creative) and 5 (most creative). The average and standard deviation values are calculated for each version and presented in Tables 1 and 2. GPT-4 has very recently shown the capability of evaluating, comparing and rating different texts according to defined criteria (OpenAI 2023; Goes et al. 2022; Park et al. 2023). One of the contributions of this paper is to be the first to test those capabilities for the AUT and TTCT creativity tests. To test the validity of this evaluation, we created human responses and mixed them with those generated by GPT-4. Our results are in line with those reported in (Haase and Hanel 2023): the scores of human responses in AUT were on average higher than the naive prompts and the best human responses were above the expert and interactive prompts, making this evaluation approach seem promising. We also tested this evaluation capability in the TTCT test: the results show that GPT-4 assesses the naive non-creative (nn) prompts with the lowest scores, followed by the naive creative (nc) ones and baselines (bs), as we would expect. This reinforces the idea that GPT-4 evaluation is robust and can evaluate different levels of creativity.

Results

Table 1 shows that the naive prompts (nn and nc) presented the lowest scores in the AUT experiment (≤ 2). The baseline (bs) presented slightly better scores than both naive versions. On each interaction over the baseline (bs), the average score increased, but slowing down until the fourth interaction (bsrdel). The human responses (hm) presented higher scores than GPT's ones as expected for human top scorers in AUT (Haase and Hanel 2023). In most cases, GPT achieved its highest score before the fourth interaction, which points to a fast convergence. Table 3 shows a sample of alternative uses of a soap for each version.

For the TTCT experiment, we used a textual adaptation of a visual task. Namely, the description of a basic figure is given (e.g., two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle) and GPT is prompted to produce a description that completes such a figure (e.g., a smiling face with the two circles as eyes, large dots as pupils, a curved line for a smile, and eyebrows above). The criteria used for evaluating the results are adapted from the rubric presented in (Jankowska and Karwowski 2015); however, here we privilege completions that do not alter the original figure in any way, meaning that the original shapes description must be present in the generated description. Although the evaluation was conducted over the textual descriptions of the images, for demonstration purposes we also asked GPT to generate the content of an SVG file for each such a description and we depicted the corresponding image by using an SVG viewer. We show some examples in Tables 4 and 5.

Table 2 shows that the naive prompts (nn and nc) presented the lowest scores in the TTCT experiment (≤ 2), as in AUT, but with lower averages. The baseline (bs) presented better scores than both naive versions. On each interaction over the baseline (bs), the average score increased, reaching the highest score in most cases in the fourth interaction (bsrdel). The human responses (hm) presented higher scores than GPT baseline (bs). On average, the fourth interaction presented the best results, even higher than the human responses. Image description completion is a harder task than AUT, and the textual version used here is a machine-oriented adaptation of the visual one usually employed to test human creativity, which can somehow justify the non-optimal results obtained via human generation.

Conclusion

In this paper, we have shown that it is possible to improve the creativity of LLMs' responses by challenging GPT with a sequence of forceful prompts. A possible future extension of this paper is the investigation of different sets of prompts to verify and improve the creativity convergence of the responses generated by GPT. We also believe that our paper contributes towards creating an automated approach to enhance naive prompts for creative tasks. This paper is also the first to use GPT to evaluate AUT and TTCT responses. Well-crafted prompts often yield better results, while poorly or naively constructed prompts can lead to subpar outputs (Mishra et al. 2023). Ideally, LLMs should generate high-quality results even with imperfect prompts. Although in this paper we only focused on creativity, the generality of the forceful prompts utilised suggests that a similar interactive process could be applied to tasks involving other cognitive skills such as critical thinking, decision making, etc. Such an exploration is also left for future work.

Acknowledgments

FG and MV are supported by the University of Leicester, in particular the Computing and Mathematical Sciences Department. JW is supported by the National Space Centre Creative in Leicester.

Author contributions

Experimental design: FG, MV, JW, PS, MG; Implementation: FG, MV, JW; Writing and Editing: FG, MV, PS, MG.

References

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. In *arXiv*.

Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges. In *arXiv*.

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

Haase, J., and Hanel, P. H. P. 2023. Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. arXiv.

Jankowska, D. M., and Karwowski, M. 2015. Measuring creative imagery abilities. *Frontiers in Psychology*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Welleck, S.; Majumder, B. P.; Gupta, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-refine: Iterative refinement with self-feedback. In *arXiv*.

Mishra, A.; Soni, U.; Arunkumar, A.; Huang, J.; Kwon, B. C.; and Bryan, C. 2023. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models. In *arXiv*.

OpenAI. 2023. GPT-4 Technical Report. In *arXiv*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv.

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; Khatun, A.; and Paraskevopoulou, S. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *International Conference on Computational Creativity*. Association for Computational Creativity (ACC). https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation/blob/main/sawicki2023gptpoetry.pdf.

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).

Summers-Stay, D.; Voss, C. R.; and Lukin, S. M. 2023. Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Toplyn, J. 2022. Witscript 2: A system for generating improvised jokes without wordplay. In de Silva Garza, A. G.; Veale, T.; Aguilar, W.; and y Pérez, R. P., eds., *Proceedings of the International Conference on Computational Creativity 2022*, 22–31. Association for Computational Creativity (ACC).

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