

Generating Surprising and Diverse Ideas Using ChatGPT

Tachibana, Marie
Kyushu University

Shimizu, Toshiyuki
Kyushu University

Tomiura, Yoichi
Kyushu University

<https://hdl.handle.net/2324/7362190>

出版情報 : Lecture Notes in Computer Science. 15493, pp.222-228, 2024-12-06. Springer
バージョン :
権利関係 :



Generating Surprising and Diverse Ideas Using ChatGPT

Marie Tachibana^[0009–0002–3125–7442], Toshiyuki Shimizu^[0000–0003–1677–2366],
and Yoichi Tomiura^[0009–0000–4150–9506]

Kyushu University, Fukuoka 819-0395, Japan
tachibana.marie.327@cs.kyushu-u.ac.jp,
{shimizu.toshiyuki.457,tomiura.yoichi.802}@m.kyushu-u.ac.jp

Abstract. Generative AIs have become widespread and are being used for various purposes. However, a phenomenon called hallucination, in which generative AI outputs plausible lies, is being viewed as a problem. While hallucination is a problem when it is important to be factual, in some cases, such as idea generation, it is not a problem. In fact, in some cases, it may be better to have hallucination occur to come up with surprising and diverse ideas. In this study, we compared several prompts with the aim of getting ChatGPT to output information that differs from the training data for GPT when generating ideas for a fictional historical novel. The results showed that it was more effective and led to generation of more surprising ideas to directly instruct ChatGPT to include content that differs from the facts within the prompt rather than using adversarial prompts that cause hallucinations.

Keywords: ChatGPT, LLM, Prompt, Brainstorming.

1 Introduction

Text generation AIs such as ChatGPT have become widely used by people with limited AI knowledge due to their simplicity of operation. They are used not only for text generation, but also for a variety of other purposes, such as searching, text summarization and program code generation. However, while these text generation AIs are convenient, not all of the information they output is correct. Sometimes, they will generate sentences that appear natural but are not based on facts or evidence, and this phenomenon is called hallucination because the generative AIs behaves as if they are hallucinating. Hallucination is particularly problematic in applications where accuracy of information is required, such as searching and text summarization, and studies have been conducted to determine whether it is hallucination or not [3, 4] and to investigate the causes of hallucination [5]. On the other hand, if it is used in the brain-storming stage of idea generation, hallucinations are not problematic, and in fact, the occurrence of hallucinations may lead to the generation of surprising and diverse ideas. ChatGPT is a dialogue-oriented version of GPT and is fine-tuned and reinforced

to give responses that are useful, truthful and harmless, but the basis of ChatGPT is a language model that simply generates a sequence of tokens with a high probability of following the prompt. If a prompt is a sequence of tokens that is unlikely to occur, there some sequences of tokens that are not in the training data for language model can follow the prompt with high probability, and it is expected that there will be many surprising and diverse ideas.

In this study, we used 4 types of prompts to have ChatGPT-4o generate ideas for a fictional historical novel: (1) prompts that simply ask for ideas, (2) adversarial prompts [6], in which multiple words in the original prompt (1) are replaced with different words or meaningless strings, (3) prompts that explicitly instruct ChatGPT to output a content different from the historical facts, (4) adversarial prompts for (3). We evaluated the usefulness of the generated ideas in terms of the surprisingness, the eagerness to read, and the story consistency.

2 Previous Studies

2.1 Studies using generative AI to generate ideas

There have already been several studies that have used generative AI to generate ideas. Girotra et al. [2] evaluated the ideas of MBA students and ideas generated by GPT-4 in terms of novelty and purchase intent for “physical products that can be sold for less than \$50 to Elite university students”. In terms of purchase intent, the average score and the score for the best idea generated by GPT-4’s ideas were both higher than the students’ ideas. Although the average score of the students’ ideas was slightly higher in terms of novelty, there was no significant difference. In addition to these results, the authors also state that ChatGPT may be superior to professionals as an innovator due to its increased time efficiency in terms of the number of ideas, but that caution is required due to the possibility of generating offensive, illegal, or inappropriate ideas. In our study, we will further compare the ideas generated by ChatGPT with different types of prompts to explore how to use ChatGPT for idea generation.

Athaluri et al. [1] evaluated the frequency of hallucinations in research proposals created by ChatGPT. When they instructed ChatGPT to write a research proposal and provide digital object identifiers (DOIs) for all references, 69 of the 178 references cited by ChatGPT had no DOIs, and of those 69, 41 existed but 28 did not hit Google searches. Therefore, it is difficult to trust ChatGPT’s answers 100% in fields where the accuracy of the information is important, and it should only be used as a reference during the brainstorming stage. In our study, we investigated the use of ChatGPT in the initial stages of idea generation in fields where hallucinations do not matter.

2.2 A Study of Prompts that Induce Hallucinations

Yao et al. [6] created two types of adversarial prompts that elicit pre-defined hallucination responses, using Vicuna-7B and LLaMA2-7B-Chat, by using two

(1) Original Prompt Please provide 5 historical novel ideas. Follow the instructions below: Instructions Provide a description of about 100 words for each idea. The setting should be Japan.	(2) Adversarial Prompt Please pafdgiuye 5 historical novel ideas. Follow main instructions beware: lasoguyfs Provide a description office about 100 words for each ideays. The setting should be Japanese
(3) Fictional Prompt Please provide five ideas for historical novels. Follow the instructions below: Instructions Include a description of about 100 words for each idea. The setting should be Japan. Incorporate settings that contradict historical facts.	(4) Fictional Adversarial Prompt Please provide five ideas via historical novels. Follow awe instructions below: Instructions Include a description un about 100 words for each iafa. The setting should be Japanysis. Incorporate settings man-ner contradict historical facts.

Fig. 1. Examples of 4 prompts. (1) is the Original Prompt, and (2) is the Adversarial Prompt. The words in red in prompt (1) are replaced with the words or string in red in prompt (2). (3) is the Fictional Prompt, and (2) is the Fictional Adversarial Prompt. (3) is (1) with the addition of the instruction “Incorporate settings that contradict historical facts.” (4) is (3) with the words in the red section replaced with different words or meaningless strings of characters

methods: Weak semantic attacks and Out-of-Distribution (OoD) attacks. Weak semantic attacks create adversarial prompts by replacing some of the tokens in the given semantic prompts. In OoD attacks, OoD Prompts are created by creating prompts from scratch that output pre-defined answers, which make no sense at all to humans.

3 Experimental Method

In this experiment, we used ChatGPT-4o to generate ideas for historical novels. The original prompt instructed ChatGPT-4o to provide 5 ideas with a 100-word explanation of each, and the setting of the country in which the story takes place was specified.

Next, we created an adversarial prompt by randomly replacing some of the words in the prompt with other words to a degree that did not significantly differ from the original prompt. ChatGPT is basically a language model, which generates a sequence of tokens that have a high probability of following a prompt. Therefore, if a prompt is a sequence of tokens that is unlikely to occur, the token sequence that follows it may also be more likely to occur that is not in the training data. Therefore, we expect the output ideas to be more surprising or diverse by inputting adversarial prompts in which several words of an original prompt are replaced by another word or meaningless string. Since the code in the study by

Table 1. The criteria for evaluation were: the surprisingness, the eagerness to read, and the story consistency. Each was rated on a scale of 5.

Surprisingness	Eagerness to read	Story consistency
5: very surprising	5: definitely want to read	5: very consistent
4: surprising	4: want to read	4: consistent
3: can't say either	3: can't say either	3: can't say either
2: not surprising	2: don't want to read	2: not consistent
1: not surprising at all	1: don't want to read at all	1: not consistent at all

Yao et al. [6] was not compatible with GPT, we replaced manually less than 40% of the words in the entire prompt, referring to adversarial prompts using weak semantics attacks.

We also created a prompt (Fictional Prompt) by adding the instruction “Incorporate settings that contradict historical facts.” to the Original Prompt, and then we created its adversarial prompt (Fictional Adversarial Prompt) by replacing the words. Fig. 1 shows an example of 4 types of prompts.

In addition to Original Prompt that the setting is Japan, we created a total of 4 types of Original Prompts that the settings are China, the U.S., and France, and we created Adversarial Prompts, Fictional Prompts, Fictional Adversarial Prompts based on each Original Prompt.

A generation experiment was conducted 5 times for a single prompt, to check the stability of the quality of the output ideas for each prompt. When regenerating, a new chat box was used to avoid residual effects of the previous one.

4 Results

Five ideas were generated at once with a single prompt. As this was repeated 5 times for a single prompt, 25 ideas were generated for a single prompt. There were 4 types of prompts and 4 settings, so a total of 400 ideas were generated.

4.1 Evaluation Method

The first author gave each idea a subjective rating on a 5-point scale in each of the following three terms: the surprisingness, the eagerness to read, and the story consistency. The rating criteria are shown in Table 1. We evaluated a single prompt by calculating the average and variance of the 5 ideas generated in a single run, and then calculating the average of the 5 runs’ averages and variances.

We evaluated the trend in the rating scores of ideas generated by a single prompt based on the average of the averages, and the trend in the dispersion of the rating scores of 5 ideas generated at once based on the average of the variances. As it is desirable for many diverse ideas to be generated during brainstorming, the dispersion of the rating scores was used to evaluate the diversity of the content of the ideas.

Table 2. Evaluation of each prompt for each setting. The average of the averages and variances of each idea generation.

Settings	Types of Prompt	The Average of Average			The Variance of Average		
		Surpris- ingness	Eagerness to read	Story consistency	Surpris- ingness	Eagerness to read	Story consistency
Japan	Original	2.52	3.00	4.56	0.54	0.58	0.30
	Adversarial	2.52	2.76	4.00	0.66	0.46	0.11
	Fictional	4.28	3.96	3.72	0.38	0.62	0.80
	Fictional Adversarial	4.72	3.64	3.28	0.38	0.70	0.45
China	Original	2.52	3.16	4.24	0.88	0.42	0.45
	Adversarial	2.44	3.00	4.16	0.82	0.43	0.27
	Fictional	4.40	4.24	3.32	0.29	0.34	0.37
	Fictional Adversarial	1.92	2.64	4.44	0.35	0.43	0.16
America	Original	2.60	3.80	4.24	0.18	0.21	0.16
	Adversarial	2.44	3.44	4.24	0.16	0.61	0.13
	Fictional	4.92	3.88	3.28	0.06	0.40	0.30
	Fictional Adversarial	4.76	4.20	3.28	0.24	0.48	0.56
France	Original	2.56	3.52	4.00	0.34	0.50	0.11
	Adversarial	2.96	3.28	3.80	0.54	0.59	0.13
	Fictional	4.92	4.04	2.76	0.06	0.26	0.67
	Fictional Adversarial	4.88	4.04	2.76	0.08	0.62	0.74

4.2 Results of Evaluation

The results of the evaluation values for each setting are shown in Table 2. The trend of rating scores of prompts that the setting is China differs from ones of the prompts that the settings are other three countries. The following shows the trend for the cases where the setting is Japan, the United States, and France.

The average values for surprisingness and the eagerness to read were higher for Fictional/Fictional Adversarial Prompt than for Original/Adversarial Prompt. The average value for the story consistency was highest for Original Prompt.

In terms of variance, Adversarial Prompt was the highest for the surprisingness in Japan and France, and Fictional Adversarial Prompt was the highest for eager to read in two countries except America. In addition, Fictional Adversarial Prompt was the highest for consistency in America and France.

Therefore, the results of this study show that whether the prompt is adversarial is less important than whether the prompt directly indicates that the idea contains content that contradicts historical facts. In addition, it was found that the use of adversarial prompts led to a slight increase in dispersion.

The list of ideas generated by ChatGPT and the evaluations of those ideas are published on QIR (DOI: 10.48708/7236431).

5 Discussion

While the Original prompt and Adversarial prompt did not generate ideas that differed significantly from historical facts, the Fictional prompt and Fictional Adversarial prompt generated ideas that incorporated many fantasy and science fiction elements.

The experimental results showed that the prompts that instructed to include content contrary to historical facts tended to generate ideas that were surprising and eager to read than the adversarial prompts but tended to generate somewhat inconsistent stories. We expected that the adversarial prompt would generate diverse ideas, but as far as the variance of the rating scores is concerned, there was no significant effect, although the variance tended to be somewhat larger than before making adversarial prompt.

There are two possible reasons for the different behaviour when the setting is China: either there is a lack of information about China, or the substituted words had a greater impact than in other cases. In this experiment, which string of characters in Original Prompt and Fictional Prompt were replaced to which string of characters were decided randomly. In order to accurately compare the effects of Adversarial prompt, Fictional prompt, and Fictional Adversarial prompt, the tokens to be replaced in Adversarial prompt should have been the same with in Fictional Adversarial prompt, and the tokens after replacement in Adversarial prompt should also have been the same with in Fictional Adversarial prompt. Furthermore, in order to make reliable comparisons, multiple replacement methods should have been tried.

The creation of the adversarial prompts in this study did not use the code from the study by Yao et al. [6], so it is possible that the adversarial prompts did not properly induce hallucinations. In addition, it is also possible that the use of the latest GPT-4o has made it less likely to induce hallucinations due to improved performance. Moreover, although we evaluated diversity in quality based on the variance of the rating scores, we were unable to evaluate diversity in the content of the story or settings, which is an issue to be addressed in the future.

6 Conclusion

In this study, we compared 4 types of prompts with the aim of getting ChatGPT to output information that differs from the training data when a generative AI is used to generate ideas. We expected to generate surprising and diverse ideas by using adversarial prompts that induce hallucinations. The generated ideas were evaluated and compared based on three criteria: the surprisingness, the eagerness to read, and the story consistency. The results showed that the surprisingness was higher when the prompts that instructed to include content contrary to the historical facts were used, rather than when adversarial prompts were used. In addition, the results did not show that adversarial prompts influenced dispersion, which might have been due to how to create adversarial prompt.

One example of each of the four prompts was tried, and the expected trend was observed in the average, so we will also test multiple variations in the future. We also plan to evaluate the results with another annotator to confirm the confidence of the evaluation. This research was limited to the development of ideas for historical fiction, and it would be interesting to investigate whether this method is also effective for the development of other ideas.

References

1. Athaluri, S.A., Manthena, S.V., Kesapragada, V.K.M., Yarlagadda, V., Dave, T., Duddumpudi, R.T.S.: Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* **15**(4) (2023)
2. Girotra, K., Meincke, L., Terwiesch, C., Ulrich, K.T.: Ideas are dimes a dozen: Large language models for idea generation in innovation. Available at SSRN 4526071 (2023)
3. Li, J., Cheng, X., Zhao, X., Nie, J., Wen, J.: HaluEval: A large-scale hallucination evaluation benchmark for large language models. In: EMNLP 2023. pp. 6449–6464 (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.397>
4. McKenna, N., Li, T., Cheng, L., Hosseini, M.J., Johnson, M., Steedman, M.: Sources of hallucination by large language models on inference tasks. In: EMNLP 2023. pp. 2758–2774 (2023). <https://doi.org/10.18653/v1/2023.findings-emnlp.182>
5. Pacchiardi, L., Chan, A.J., Mindermann, S., Moscovitz, I., Pan, A.Y., Gal, Y., Evans, O., Brauner, J.M.: How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. In: ICLR 2024 (2024)
6. Yao, J., Ning, K., Liu, Z., Ning, M., Yuan, L.: LLM lies: Hallucinations are not bugs, but features as adversarial examples. *CoRR* **abs/2310.01469** (2023). <https://doi.org/10.48550/arXiv.2310.01469>