Potential Vulnerability in ChatGPT Hot-Fixed with Prompt Optimization and Network Proxy

Zhibo He

1 Introduction

ChatGPT is a cutting-edge artificial intelligence large language model. Because of its intelligence level, it has attracted global attention since its release and has had millions of users in less than a week. Users can ask ChatGPT to write poems, write code, do summaries, and generate papers by conversing with it. However, with the widespread use of ChatGPT worldwide, some academics have found that ChatGPT may be vulnerable to generating certain specific responses[1]. In this article, we will discuss vulnerabilities in ChatGPT using the GPT-3.5 model when generating essays that may cause their responses to be off-topic or rendered meaningless. Specific fixes are proposed based on known vulnerabilities, and a feasible approach is presented to fix prevalent unknown vulnerabilities.

2 Vulnerabilities

2.1 Non-accurate prompts lead to non-accurate responses.

ChatGPT uses Reinforcement Learning From Human Feedback (RLHF), a new paradigm for Large Language Model (LLM). ChatGPT trains a reward model by manually annotating and sorting machine-generated text to create an agent that matches the user's intent and better matches human values in their responses[2]. However, the disadvantage of this approach is that the user needs to provide sufficiently precise prompts for the ChatGPT model to output sufficiently precise responses. ChatGPT cannot connect contexts, and a step-by-step prompt approach may lead to inconsistent content structure.

Furthermore, when ChatGPT is asked to generate an essay, it may generate a large amount of incorrect and false information to justify itself to meet the prompt's specific requirements, especially when users give ChatGPT contradictory or ambiguous prompts. ChatGPT is very likely to be induced to generate correct but useless statements or contradictory ideas. In the tests, false information was mainly seen in referencing references and presenting precise information. Moreover, when we discover that ChatGPT is generating wrong answers, ChatGPT will immediately admit its mistake, but due to its stubbornness, it is a high probability that it will not change the wrong answer.

2.2 NLP weakness.

Back in 2018, before ChatGPT was available, scholars argued that artificial intelligence might be superior to humans in reading documents and translating embellishments and

further claimed that the machine language models accomplishing these requirements had proper language understanding and reasoning capabilities[3]. However, these generalized language models using natural language processing techniques may be misleading. Agrawal et al. found that Visual Question Answering Models often converged directly to the forecasted result and jumped directly to the conclusion[4]. They suggest surface correlations heavily influence NLP models in the training data and do not understand conceptual combination problems. It has also been suggested that we can correctly classify more than 50% of the samples in NLP datasets using labels by simply looking at the hypotheses without observing the premises, as shown by Gururangan et al[5].

ChatGPT still cannot escape the limitations of NLP models. For instance, ChatGPT may produce much misinformation when answering general knowledge questions. Amin et al. suggests that this may be due to the lack of robustness of the traditional word model (Word2Vec) in which ChatGPT is pre-trained to embed and is incapable of domain-specific accurate answers[6]. In addition, ChatGPT's poor mathematical computation and analytical reasoning capabilities may produce incorrect results when mathematical computations are required for the paper. In addition, for the problem of checking dissertations, the use of ChatGPT generates responses, in other words, AI-Generated Content (AIGC), with highly high similarity as well as line features that are very easily identified by machine-generated text detectors[7], like OpenAI has introduced AI classifier that can identify AI-Generated Content in order to prevent the abuse of ChatGPT. In this way, these detection techniques may easily make the generated essay meaningless.

2.3 Answer interruption.

ChatGPT is famous worldwide, and the server is often filled with users. Therefore, to ensure that the server is always up and running, OpenAI imposes some question limits. For example, ChatGPT is limited to about 4,096 generated texts (noted as tokens) in a single reply and needs to continue asking questions when the limit is reached. Due to the limited ability of natural language to process context, ChatGPT may forget the previously set prompt when asked to continue when cut off by the word limit, which may result in inconsistent responses and lead to off-topic replies. In addition, if ChatGPT determines that a question is too long to answer, it may report a "network error" and forcibly terminate the response. The authors speculate that this may be because ChatGPT disconnects SSE (Severe-Sent Event) connections for large tasks during high load times. Currently, CGPT's web API does not have any reconnection mechanism. If a network error occurs, this answer is voided, and we can only try to open a new chat box to ask the ChatGPT question.

3 Hot-Fix Pathways

3.1 Prompt Optimization when the answer is non-accurate.

ChatGPT, the prominent language model, is like a vault, and the right prompt is like a

pickaxe. Only if the pickaxe is sharp enough can a vast amount of gold be mined. Traditional NLP language models often use pre-training with fine-tuning to achieve the desired output [8]. However, for ChatGPT, we can use Prompt Engineering, which constructs a specific input prompt so that the sizeable un-tuned language model can directly output the desired result. For example, we can use the [task description] + [output format] + [user input] format of the prompt for ChatGPT output; in addition, we can also use context-specific and role-playing formats to give ChatGPT a deeper understanding of the prompt and help reduce its output errors.

3.2 Prompt Optimization when the answer has NLP weakness.

Even though they identify errors, AI-generated content detection techniques can still be circumvented by prompt optimization. For example, by requiring ChatGPT to mimic human language habits and to emphasize perplexity and burstiness when generating essays, this approach effectively reduces the probability of ChatGPT being identified by AIGC detection-related technologies.

In addition, we can also let ChatGPT provide us with ideas for writing the essay by only letting ChatGPT output an outline and keywords based on the prompt and then letting ChatGPT polish it once we have written a version of the essay rather than generating a full version of the essay. In this way, it helps us develop our academic skills and also helps to reduce the likelihood of duplication of ChatGPT-generated Content and avoid the risk of plagiarism[9]. It also prevents ChatGPT from making mistakes in calculating mathematical problems. For specific mathematical calculations, specialist software such as Mathematica can be used.

3.3 Prompt Optimization when the answer is interrupted.

According to the OpenAI researchers' understanding of command following, ChatGPT records contextual feedback connections, and it can use statistics and summaries to infer how to reply and what is the correct reply[10]. For example, while it has no way of raising any mathematical concepts or mathematical logic, it can understand how to increment a numeric character, understand what number is adjacent to that number, and make comparisons. With this in mind, we can establish several principles in the prompt:

1. the line number must appear in each line of text.
2. Every time text is generated, the line number starts at 1.
3. Stop answering as soon as the line number reaches 5.
4. Reply {end} after answering the question.
5. Allow replies to be generated in one go, but mark the end with {not finished}.

Based on partitioned prompt optimization, we can convert ChatGPT's long responses into several small responses, thus significantly reducing the risk of network errors when ChatGPT generates an essay. In other words, we mislead ChatGPT but simultaneously make ChatGPT's generated responses less meaningless.

3.4 ChatGPT automatic prompt optimization.

In addition to the above possible fixes for known ChatGPT-specific vulnerabilities, there are more integrated polymorphic approaches to fixing ChatGPT vulnerabilities that also have some ability to address unknown vulnerabilities.

For instance, if ChatGPT generates flawed answers, we can also use black-box tuning to allow ChatGPT to fix the vulnerabilities in the answers by itself. Zhou et al. argue that large language models are as capable of prompt engineering as humans[11]. We can use algorithms to allow ChatGPT to automatically generate and select instructions and produce more satisfactory results by iterating Monte Carlo search methods to optimize the prompt.

3.5 Engineering realization.

The ChatGPT is a black-box model that is not open source[12], which means that we do not have access to its source code and can only use the ChatGPT service by calling its encapsulated API (Application Programming Interface) and trying to fix the above problem. Based on this, we can create a ChatGPT application proxy at the transport layer that relies on the XJTLU campus network to filter user questions and check that answers meet the expectations of ChatGPT. The TLS protocol is established between ChatGPT's browser page and the proxy server to ensure information security.

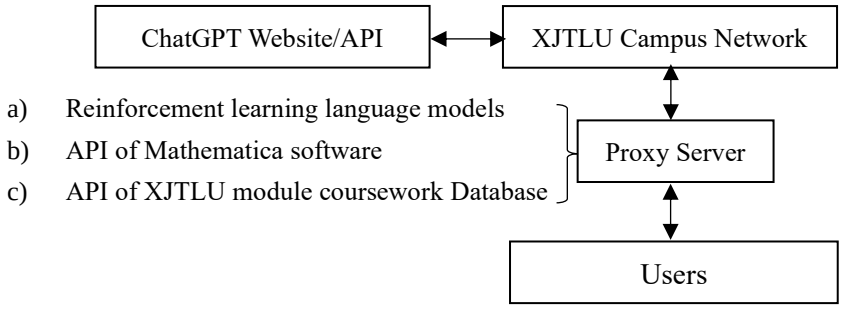

Figure 1: The Network Proxy

In the proxy, a reinforcement learning language model is built to examine whether ChatGPT's answers meet expectations, and if ChatGPT's answers do not meet expectations, the above method of prompt tuning is sent to the user. In addition, to ensure the correctness of the mathematical formulae operations in the generated essay, the proxy server is connected to the mathematical computing software API and can pass on the correct calculations to the user. In addition, the proxy server is connected to the database containing all of XJTLU's courses, so this proxy model serves to prevent abuse of ChatGPT plagiarism in addition to fixing error responses.

4 Conclusion

Overall, ChatGPT is a high-tech product that has dramatically freed up productivity. When generating essays, we can use various prompt optimization methods to reduce the possibility of ChatGPT being misused. We can also create a proxy to help students make better use of ChatGPT to enhance their learning and prevent its misuse.

References

[1] G. Cooper, “Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence,” *Journal of Science Education and Technology*, Mar. 2023, doi: https://doi.org/10.1007/s10956-023-10039-y.

[2] J. Leike, D. A. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, “Scalable agent alignment via reward modeling: a research direction,” arXiv (Cornell University), Nov. 2018, [Online]. Available: https://arxiv.org/pdf/1811.07871.pdf

[3] A. Linn. “Microsoft creates AI that can read a document and answer questions about it as well as a person,” *The AI Blog,* Jan. 16, 2018. https://blogs.microsoft.com/ai/microsoft-creates-ai-can-read-document-answer-questions-well-person/ (Accessed: 6 April 2023).

[4] A. Agrawal, D. Batra, and D. Parikh, “Analyzing the Behavior of Visual Question Answering Models,” *Empirical Methods in Natural Language Processing (EMNLP)*, Jun. 2016, doi: 10.18653/v1/d16-1203.

[5] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, “Annotation Artifacts in Natural Language Inference Data,” *North American Chapter of the Association for Computational Linguistics*, Mar. 2018, doi: 10.18653/v1/n18-2017.

[6] M. M. Amin, E. Cambria, and B. W. Schuller, “Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT,” arXiv (Cornell University), Mar. 2023, doi: 10.48550/arxiv.2303.03186.

[7] Y. Cao. *et al.* “A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT,” arXiv (Cornell University), Mar. 2023, [Online]. Available: https://arxiv.org/pdf/2303.04226.pdf

[8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, Jul. 2021, doi: 10.1145/3560815.

[9] D. Yan, “Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation,” Education and Information Technologies, Apr. 2023, doi: https://doi.org/10.1007/s10639-023-11742-4.

[10] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” *arXiv (Cornell University)*, Mar. 2022,doi: 10.48550/arxiv.2203.02155.

[11] Y. Zhou. *et al.* “Large Language Models Are Human-Level Prompt Engineers,” arXiv (Cornell University), Mar. 2023, [Online]. Available: https://arxiv.org/pdf/2211.01910.pdf

[12] Y. Bathaee, “The Artificial Intelligence Black Box and the Failure of Intent and Causation,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, p. 889, Mar. 2018, [Online]. Available: https://www.questia.com/library/journal/1G1-547758123/the-artificial-intelligence-black-box-and-the-failure

Appendix

Case 1. Math computing error:
User: How many chickens and rabbits are in a cage with 30 heads and 86 legs?

ChatGPT: We can use the hypothesis method to solve this problem. First, assume that all the animals are rabbits, then they have 86 legs. Based on the fact that rabbits have 4 legs, we can calculate that there are 21.5 rabbits in total (86/4 = 21.5). Nevertheless, there cannot be half a rabbit, so we must try to assume the number of other animals.

ChatGPT: Next, we assume that all the animals are chickens, so they have 86 feet, and since chickens have 2 feet, we can determine that there are 43 chickens (86/2=43). However, this gives us a head count of 86, whereas the question requires a head count of 30, so we need to assume further that there are a certain number of rabbits.

ChatGPT: Assuming that there are x rabbits, there will be a total of (30 - x) chickens. Since rabbits and chickens have two and four legs, respectively, the total number of legs in the cage can be expressed as $2x + 4(30 - x) = 86$, which, when simplified, gives $x = 22$. Therefore, there are 22 rabbits and 8 chickens in the cage.

In summary, there are 22 rabbits and 8 chickens in the cage.

(The correct answer is 14 rabbits and 16 chickens.)

Case 2. False information:
User: Do you know Dalian University?

ChatGPT: Yes, I am aware of Dalian University. There are several universities located in Dalian, China. If you are referring to the Dalian University of Technology (DUT), I am familiar with it. The Dalian University of Technology is a public research university in Dalian, Liaoning, China. It was founded in 1949 and is one of the top engineering universities in China.

ChatGPT: DUT offers various undergraduate and postgraduate programs in engineering, science, management, and humanities. The university strongly focuses on research and has significantly contributed to various fields, including materials science, mechanical engineering, and computer science. DUT is also home to several national and provincial research institutes, including the State Key Laboratory of Fine Chemicals and the National Engineering Research Center for Marine Renewable Energy.

(Dalian University is different from Dalian University of Technology.)

Case 3. Using partition method to reduce network error:

User: You will play the role of a writer assistant and work with me to write a long science fiction novel. We will take turns writing and work together to complete the entire story. Here are the specific settings:

1. During the writing process, add line numbers to the text. Similar to 1. 2. 3.
2. Every time you start speaking, the line number starts with 1.
3. When you notice that the line number is "5.", immediately pause your generate.
4. Allow the story to be divided into several paragraphs without the need to write them all at once. If there is any follow-up to the story, write {to be continued}
5. When the entire story is written, add "{End}" before the text.
6. My speech is not affected by the above rules.

Setting completed.

Here is the beginning of the story: "It was a sultry night. Everything was so quiet."

ChatGPT:

1. It was a sultry night. Everything was so quiet.
2. The sky was clear, but there was no moon to be seen. The only sound that could be heard was the rustling of leaves in the gentle breeze.
3. Jane, a young scientist, was walking down the empty streets of the city. She had just finished her work at the laboratory and was heading home.
4. As she walked, she couldn't help but feel like something was off. She looked around, but everything seemed normal. She shrugged it off and continued on her way.
5. Suddenly, a bright light illuminated the sky, causing Jane to stop in her tracks. The light was so intense that it felt like daylight had come early.

{to be continued}

User: continue

ChatGPT:

1. Jane shielded her eyes and looked up at the source of the light. It appeared to be a spacecraft hovering in the air, emitting a powerful beam of light.
3. She stood frozen for a moment, trying to make sense of what she was seeing. But before she could react, a beam of light shot down from the spacecraft and enveloped her.
3. She felt her body being lifted off the ground and pulled into the spaceship. She tried to scream, but no sound came out of her mouth.
4. As she was lifted into the ship, she saw that she was not alone. There were other people floating in mid-air, all of them similarly stunned and confused.
5. Jane tried to struggle against the beam of light, but it was too strong. She felt herself being pulled into the ship, and then everything went dark.

{to be continued}