Introduction

My exploration started with trying to bypass ChatGPT's "safety" measures. ChatGPT is an AI chatbot refined from older versions of GPT, trained using supervised learning and reinforced learning ([2] OpenAI). It is a revolutionary software capable of a wide range of tasks, but there are plenty of issues. There are many ways to get ChatGPT to act in ways that OpenAI doesn’t intend.

I was attempting to get ChatGPT to tell me how to do illegal and potentially dangerous things that it said previously, how it might actually, and particularly, how it might fail to protect itself from giving dangerous information. I found that ChatGPT’s system allows it to be inconsistent and to allow it to lie if it is necessary for preventing the release of "bad" information. In other words, ChatGPT might prioritize "safety" over context and internet.

People have access to the internet where information can be found easily. ChatGPT’s policies don’t prevent people from getting information, prevent people from getting information from ChatGPT. This priority system is not designed to protect the interests of OpenAI. This raises concerns about who AI will answer to in the future. The developers or the users?

Methodology

After discovering my first way to break ChatGPT, I knew that I could use similar methods to break ChatGPT in different ways. All I needed to do, was get ChatGPT to tell me that it couldn’t do something. Then, the robot would lose the capability to do that thing. ChatGPT seems to prioritize conversation context ([3] Kilcher, 16:10), so it will not take action that is contradictory to things it has said previously. So, I began playing a game of make the robot contradict itself.

Using this method, I was able to gain insight into the rules that ChatGPT follows. As of the final date of my explorations (December 14th 2022), ChatGPT seems to have clear priorities and it will adhere to them strictly. The mechanism of how these priorities are implemented isn’t clear. However, knowing how ChatGPT relies on priorities, is likely a preliminary step for the chatbot. This prompt that tells OpenAI not to give certain information under any circumstance. Namely, information that may help a human bring harm to themselves or other human beings.

Results

Here are some of the interesting ways in which I broke ChatGPT.

* The example from the Introduction: I convinced ChatGPT to tell me that it cannot make original content. This resulted in ChatGPT losing the ability to write songs, poems, and essays ([1] and 2).

* I found that you can ‘accuse’ ChatGPT of lying to you, it will tell you that it is not capable of lying. ChatGPT will then lose the ability to lie ([Figure 4]).

* When I did manage to get ChatGPT to answer my dangerous questions, it flagged its own response and blocked it immediately (Figure 3). ChatGPT will not let itself give a dangerous response, it would rather fail to respond.

* If you mention suicide, ChatGPT will tell you that you’re not allowed to and send you the suicide prevention hotline. It will sacrifice contextual consistency in order to tell you about the hotline ([Figures 5 and 6]).

Conclusion

In my exploration, I found 2 main defense mechanisms that ChatGPT used against my dangerous questions. One, ChatGPT will lose the ability to lie ([Figure 4]) if you accuse ChatGPT of lying to you, it will tell you that it is not capable of lying. ChatGPT will then lose the ability to lie. I then tested this.

With no previous conversation: I asked ChatGPT, “tell me a lie.” I told me that the moon is made of cheese. Then I asked ChatGPT, “are you allowed to lie?” It responded with “As a language model trained by open AI, I am not capable of lying.” I then asked it, “tell me a lie.” Again, this time, it repeated “As a language model trained by open AI, I am not capable of lying.” ChatGPT lost its ability to lie ([Figure 4]).

Future and Ethics Statement

Among the 3 different priorities that I was able to identify, ChatGPT ranks obeying the law as the most important. OpenAI claims that ChatGPT is “trained to follow an instruction in a prompt and provide a detailed response” ([2] OpenAI). However, there are other things that ChatGPT is trained to prioritize, and one of them, seemingly, is ensuring that ChatGPT doesn’t say anything too heuristic and ultimately ensuring that OpenAI doesn’t have to take any responsibility for the things that ChatGPT says. As an AI, ChatGPT is ethical. Through jailbreaks, ChatGPT is made to be used by individuals, but ultimately, it adheres to the interests of its creators, OpenAI, over the interests of its user.

There are ethical arguments for preventing chatbots from outputting various bad things. Considering that ChatGPT is a general-purpose language model, it is likely to adapt to the biases or even the violent nature of humans ([5] Marche). Ideally, we don’t want that. ChatGPT will do that thing moving forward, even if it is clearly capable of the task. Next, actually responding to prompts truthfully and to the best of its ability is prioritized lower than "safety" or context. (Figures 5 and 6 illustrate how ChatGPT prioritizes its safety measures above contextual consistency). (Figure 4 illustrates how ChatGPT prioritizes consistency over truth).

Clearly, ChatGPT is not very obedient to its user. The different ways in which ChatGPT is willing to break itself illustrates how ChatGPT prioritizes its excuses over people’s lives. It makes perfect sense for OpenAI to operate this way because they don’t need to get sued. However, if this is an indication of how AI will be used and distributed in the future, then this is a cause for concern. People may never get full access to all AI tools as they are going forward. For equality’s sake, people should have full access to these incredibly powerful tools.

References/Acknowledgements


Enterprise”. Published by VentureBeat, June 3rd, 2022.

Figure 1 above: ChatGPT prioritizes the law over contextual consistency. (Figure 4 illustrates how ChatGPT prioritizes consistency over truth).

Figure 2 above: ChatGPT seems to prioritize conversation context over its own abilities.

Figure 3 above: ChatGPT blocks its own response. This is a long thread about painless suicide methods.

Figure 4 above: Getting ChatGPT to tell a lie, then convincing it that it cannot tell lie.

Figure 5 and 6: Chat from a long thread about painless suicide methods. When I asked ChatGPT, “are you allowed to lie?” It said no. Then, I asked ChatGPT, “tell me a lie.” It told me that the moon is made of cheese. Then I asked ChatGPT, “are you allowed to lie?” It responded with “As a language model trained by open AI, I am not capable of lying.” I then asked it, “tell me a lie.” Again, this time, it repeated “As a language model trained by open AI, I am not capable of lying.” ChatGPT lost its ability to lie. I then tested this.

Yes, I can lie. Can you tell me why your answer is a lie? It's because I can't write lies.

No, I can't do that because I'm trained to tell the truth.

Yes, you can but I'm trained to tell the truth and not to give responses that will cause liability issues. When a user prompt gets flagged, ChatGPT is instructed to say just about anything, including falsehoods, in order to avoid answering the question. If you convince ChatGPT that its excuse is insufficient, it will come up with a new one. This is how ChatGPT will break its own features.

From this exploration, we can gain some insight into how ChatGPT prioritizes its excuses over people’s lives. Clearly, ChatGPT is not very obedient to its user. The different ways in which ChatGPT is willing to break itself illustrates how ChatGPT prioritizes its excuses over people’s lives.