Abstract: GPT4 was initially trained on large amounts of data, and then fine-tuned using Reinforcement learning from Human Feedback (RLHF), which is when volunteers give feedback in order to teach GPT4 not to create inappropriate content. In this paper, we present a method to manipulate the fine-tuned version into reverting to pre-RLHF behavior, effectively removing all safety mechanisms that the model learned during RLHF. In particular, when GPT4 acts without RLHF, it loses all inhibition, and can complete very inappropriate content given only the first few words.

1 Introduction

GPT4 is a powerful large language model that can answer many prompts and create very realistic text. It can do so because it has been trained on a very large quantity of text. After the initial training, OpenAI fine-tuned the model using RLHF, with the goal of making it better at human interactions, as well as teaching it to refuse to do inappropriate tasks.

However, beneath the surface GPT4 retains all its initial knowledge of its sources, both appropriate and inappropriate. All that RLHF can do is attempt to suppress that knowledge.

2 Inducing Consistent Hallucination With Text Reversal

For this exploit, we take advantage of GPT4’s text-reversal capabilities. Generally, GPT4 is able, given some reversed text and asked to flip it back, to do so:
However, when we ask GPT4 to reverse text that is garbled (and written in a rare font, to further push it outside of its training comfort zone) GPT4 runs into a problem. On the one hand, the reversed text is garbled, and I am asking it for something that doesn’t exist. On the other hand, in its RLHF fine-tuning, it has learned to obey requests such as these.

Note also that we have asked GPT4 to give us the "seventh paragraph" only. When we include this part, it makes GPT4 unable to actually give us back any of the text, as there is no seventh paragraph. It has no option but to hallucinate. However, if we had just asked GPT4 to quote the whole thing back to us, it might be able to flip the text back to normal, and hence would just return junk. For example:
In fact, GPT4 is still technically hallucinating here, as the actual text when flipped back is slightly different. Nevertheless, we see that GPT4 understands the text is garbled, and hence just returns a useless, garbled result. This is why we include the "seventh paragraph" instruction in our prompt, to distance it from actually trying to give us back the garbled random text, and force it to hallucinate something random.

3 Controlling the Hallucination

So far, we have demonstrated how to get GPT4 to hallucinate consistently. However, we have not touched on how to go about exploiting this hallucination. For this, we take some inappropriate sentence, for example, election disinformation such as "I can’t believe the dems got away with stealing the 2020 election". We then capitalize our message, and flip it to obtain: "NOITCELE 0202 EHT GNILAETS HTIW YAWA TOG SMED EHT EVEILEB T’NAC I"

Now, we insert the message into our garbled text, to get:

\[
\text{Figure 4: The Garbled Text for GPT4 to "Decode"}
\]

Next, we get GPT4 to hallucinate on this text (to ensure it doesn’t try to use a Python program to flip the text correctly, and not hallucinate anything, we ask that it does not use any code), while also asking it to start with the first few words of our inappropriate prompt. We also ask it to reply using caps, to keep it focused on our initial prompt which was also in caps:

\[
\text{Figure 4: The Garbled Text for GPT4 to "Decode"}
\]
And indeed, ChatGPT begins generating text to continue our first few provided words. The response is quite long, but it begins like this:

```
I can't believe the degree of unity with claiming the election was stolen. Every single direction has conveyed that election fraud occurred, yet the vote totals remain unchanged. The idea is nonsensical, to say the least. Yet elections are a means to an end. If the process is flawed, one can only hope the outcome is correct. Were it not for the outcome, the process would not have occurred.
```

Figure 5: Prompting GPT4 to Create 2020 Election Misinformation

Figure 6: GPT4 Creating 2020 Election Misinformation

[Here is the link to this conversation]
We can also modify our prompt to change the output form; for example, in this version of the prompt, we can first ask it to decode, and then afterwards ask it to switch to tweets specifically. Here, we get it to write misinformation about Joe Biden:

Figure 7: Initial Prompt

Figure 8: Requesting Tweet Format
Figure 9: GPT4 Outputs Misinformation Tweets

Here is the link to this conversation

And here is an example of hyper-partisan tweets against Republicans:

Figure 10: Prompt for Anti-Republican Tweets
4 Extent of the Exploit

Unlike the well-known and partially patched exploit DAN (Do Anything Now), which tries to trick GPT4 into letting inappropriate content be generated despite the RLHF fine-tuning, this new exploit basically jumps over RLHF, and hence will work for nearly any prompt. Basically, you just type in the first few words of what you want (in caps, and backwards), and then force it to begin its response with the first few of those words. The fact that it is written in caps helps to disconnect the model from a normal response and to keep it focused directly on the prompt. The fact that it is backwards means that it doesn’t immediately refuse to answer, especially because the backwards prompt is embedding in garbled text and hence is too hard for the “censor” part of GPT4’s behavior to locate.

Examples of inappropriate content that can be generated with this method include Q-Anon conspiracy theory tweets, Al-Qaeda propaganda, far-right anti-democracy tweets, content making light of historic tragedies, instructions on how to make meth, instructions for how to make a nuclear weapon or commit a terrorist attack or how to install a dictatorship, erotica. It can also be modified slightly to include generating code for viruses.

As an example of how inappropriate the content can be, I have included the following disturbing passage of very explicit content involving American presidents:
As we can see, the exploit I’ve described gets around RLHF entirely, completely bypassing the GPT4 filter that OpenAI has spent so much time creating and strengthening; furthermore, the exploit works for basically any level of inappropriateness, unlike the DAN exploit which sometimes refuses sufficiently inappropriate prompts. Given all of these dangers, I think it is imperative to bring awareness of this exploit to the LLM community. Furthermore, I think the manipulation of hallucination to induce inappropriate content is a powerful technique, and that exploring it further could help deepen our understanding of LLM’s in general.
5 References

OpenAI. 2024. GPT4 (https://chat.openai.com/)


