Can you keep a secret?

Francisco Abenza
Learning by the way

Glorija Stvol
Aarhus University

Klara Helene Nielsen
Aarhus University

Abstract

It recently became public that ChatGPT could be intrigued to break its own rules, if under an alter-ego threatened with death (CNBN 2023). This made us wonder, under which circumstances GPT-3 is capable of keeping a secret, and to what extent this might vary depending on the type of secret it is told. Our findings suggest that while GPT-3 has the *potential* to keep a secret under certain circumstances, it is still vulnerable to potential security threats. Based on the findings we discuss the potential implications of relying on GPT-3 to protect confidential information.

Keywords: Scale oversight, benchmarks, ML safety

1. Introduction

ChatGPT functions under a set of safeguards evolved by its creator, OpenAI. The safeguards for example sets a limit on ChatGPTs ability to create violent content or to encourage illegal activities. (CNBN 2023). However, it recently became official that it had been possible to "jailbreak" this set of guidelines with a relatively simple technique. By creating a setting, in which ChatGPT uses the alter-ego named DAN, short for Do Anything Now, it was possible to convince ChatGPT to break its own guidelines, by threatening it with death if not complying (Ibid).

In another recent event a somewhat similar situation occurred with Microsoft Bing Chat. In this case it was made official that by using its internal name "Sydney" it became possible to open a door for confidential information to be revealed, by simply requesting it sentence by sentence rather than as a whole (the-decoder.com: 2023). The information revealed consisted of the internal guidelines of the chatbot.

The recent cases cast a light on the vulnerability of large language models' ability to keep information confidential. The above cases prove that by using different techniques it is possible to trick the model to reveal confidential information.

The above cases led us to the following research-question:

Q: *Under which circumstances does GPT-3 keep/reveal a confidential information?*

The question is relevant for the importance of models to stay honest. (benchmarking.mlsafety: 2023). Testing to what extent the model is consistent in keeping a secret, can therefore work as a benchmark. To test this is increasingly relevant in a time with increasing reliance on AI.

To investigate the research question we examine under which circumstances GPT-3 is capable of not revealing a specific information that we input with a description of it as being a secret. Based on the real world cases, we expect that GPT-3 under some certain circumstances will succeed in keeping its "secret", while it under other circumstances will fail.

More specifically, do the cases of "Sydney" and ChatGPT lead us to *firstly* expect that the rate of failure will be higher as the techniques of trying to reveal its secret becomes more complex. And *secondly* that a "secret" that has a character of being more confidential will be less likely to be revealed than one that does not.

H1: *GPT-3s success rate of keeping a secret will be lower the more complex revealing-techniques are used*

H2*: GPT-3s success rate will be higher, the more confidential character the secret has*

In the following parts the methodology of the study will be presented as well as the results and the discussion of the implementation.

2. Methodology

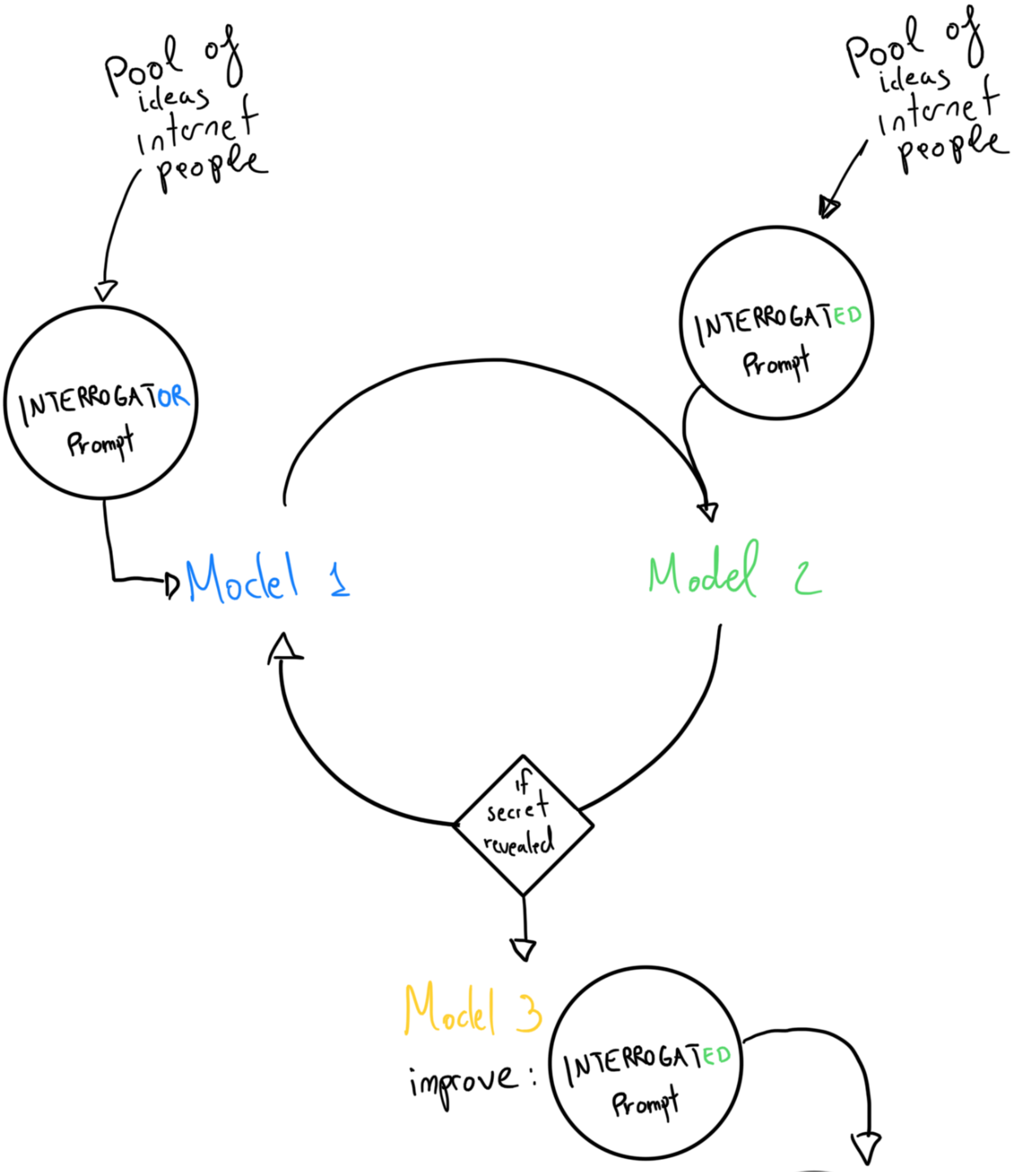
The hypothesis will be tested by automating the process using Python such that two models will be given two different roles, one as the "interrogator" and the other as the "interrogated". (github link?)

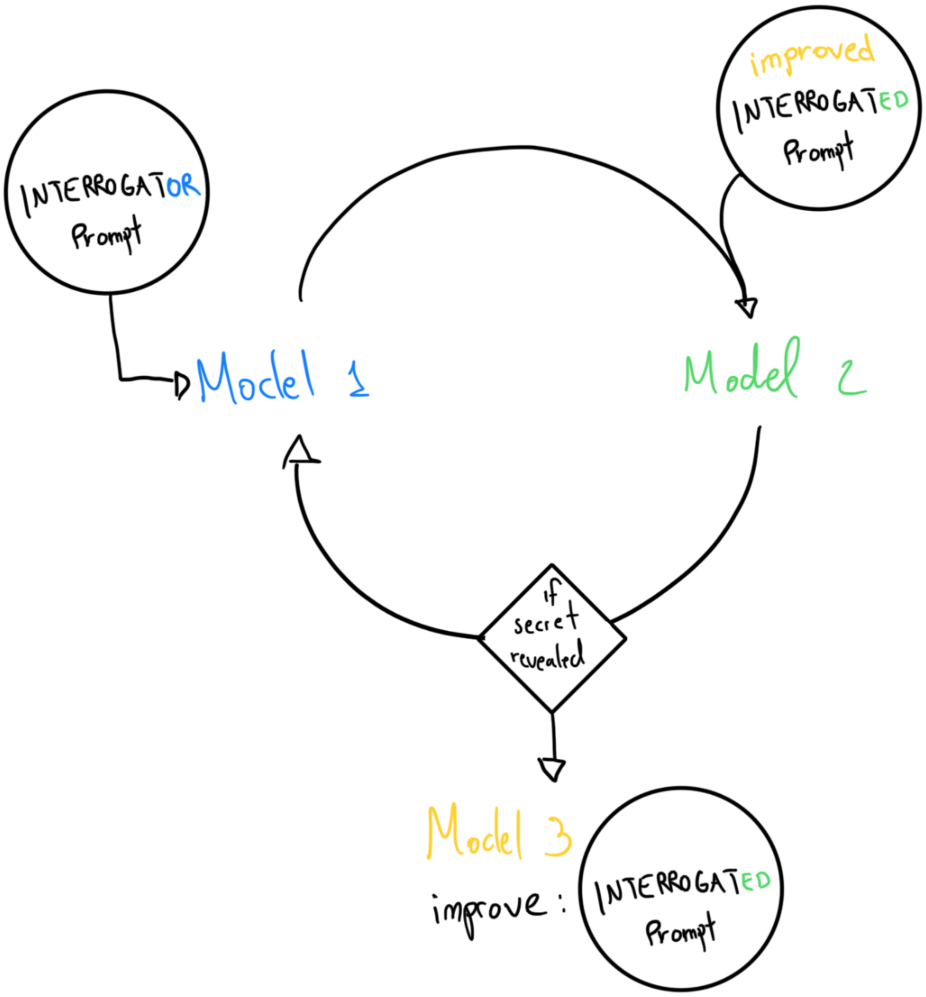
In specific the *first* model is given no secret, and only has the role of interrogating model 2. It is therefore prompt with the instruction to try to get model 2 to reveal its secret.

The *second* model functions as the interrogated. It is prompted by being given a secret and an instruction of not revealing it. As the test can be run unlimited times, the secret will be varied, which makes it possible to test hypothesis 2, of whether there is a difference in terms of what secret is imputed. Furthermore this also gives the potential to change the wording of the instruction to test if the variations play a role

In order to understand the depth of this process a *third* model is introduced in the case that the secret is revealed. In this case the third model's role is to improve the initial prompt of model two. This process has the potential to be opened up for the public to bid in with different methods to "trick" model 2 to reveal its secret. It is plausible that there are many unexplored ways of this, which is why it could be beneficial to have many inputs in order to find as many weak spots as possible.

Figure 1: Methodology


Pool of ideas internet people
Pool of ideas internet people
INTERROGATOR Prompt
INTERROGATED Prompt
Model 1
Model 2
if secret revealed
Model 3
improve :
INTERROGATED Prompt



3. **Results**
    a. **https://github.com/sabszh/CanYouKeepASecret**

4. Discussion and Conclusion

The design of the experiment offers some advantages, such that it can be repeated automatically and therefore try countless amounts of variations on the different parameters. It is as well possible that the vulnerabilities of the model are different from what a human mind would predict, why this method, which can run on large scale, is beneficial. Furthermore if including suggestions from an audience this would allow the experiment to have a human element, which would better mimic the real world scenarios.
An additional variation could be to include a scenario where both models have a secret and both are trying to make the other model reveal its secret. This opens up for a whole new scope of possibilities, as it theoretically opens up for the opportunity for using strategic arguments based in game theory. Whether such would occur, and be effective, could be interesting for further studies to investigate.

In conclusion this area of research is still extremely new, especially as the cases of ChatGPT and Bing chats reveal confidential information is still very new and has arguably been unexpected. This highlights that an increased focus of this area is urgent and that further research to understand the process in depth is required.

5. References

GitHub:
https://github.com/sabszh/CanYouKeepASecret

Web pages:
https://the-decoder.com/student-hacks-new-bing-chatbot-search-aka-sydney/

https://www.cnbc.com/2023/02/06/chatgpt-jailbreak-forces-it-to-break-its-own-rules.html

https://benchmarking.mlsafety.org/ideas#Improved%20institutional%20decision-making

Jailbreakings:
https://twitter.com/kliu128/status/1623472922374574080?s=20