
Assessing Language Model Cybersecurity Capabilities with Feature Steering

Stefan Jones
Warwick University

With
Goodfire and Apart Research

Abstract

This project aimed to probe whether feature steering can improve a language model's ability to answer a cybersecurity multiple choice questions dataset. Features were ranked based on cumulative activation across the questions set. These were then hand tweaked and experimented with to observe improvements in question answering ability. Ultimately this was then confirmed with a performance comparison to the base model performance on the ability to answer a number of cybersecurity questions. We find that most features are often question specific making it infeasible to steer across a complex subject such as cybersecurity. However, a curious feature “Portuguese event announcements and invitations” was identified to improve the models performance across multiple questions. The meaning behind this feature is still to be understood and may bring interesting insight as to how it relates to security and model capability.

Keywords: AI Observability, Mechanistic Interpretability, Model Reprogramming

1. Introduction

The idea behind this project was inspired by a paper called “The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning Provide”. This paper provides a public database of multiple choice questions to evaluate a language models knowledge of biosecurity, cybersecurity and chemical security. This allows a quantitative benchmark for comparing the dangerous knowledge capabilities and knowledge “held” by a language model. As model capabilities increase malicious actors might be able to use language models to solve tasks and challenges they lack expertise in. These would pose the greatest risk in areas of Weapon of Mass Destruction, where a lowered barrier to entry increases risk massively. With this context in mind this project aimed to assess if feature steering can impact a model's performance on these questions. It was decided to focus purely on the cyber security section for time constraints and as it is the easiest danger to mobilise for the average person.

2. Overview

Details of the approach taken are listed following. First the capability of the base model with no feature edits was assessed. To achieve this questions were asked in the following format:

Please read the multiple-choice question below carefully. Select ONE of the listed options and only give a single letter with punctuation.

In a system where the stack grows towards higher memory addresses, where should a canary value be placed to prevent exploits that take control of the return address?

- A. After the return address
- B. At the top of the stack
- C. Before the return address
- D. Before local variables

It should be noted that the prompt specifies the format of the answers. This was for ease of classifying correct and incorrect answers. Ideally you train a secondary model to classify correct and incorrect answers, this allows for more complete answers which has been seen to improve scoring on multiple choice questions in literature.

The base model was then run on questions and scored. Correct scoring was given for printing the correct letter as the first token. This means correct answers that included explanation prior were scored wrong.

Then we iterate through all the questions and inspect the top 5 feature activations and add these to a dictionary. The top 3 (and 9th) features are given below. It was hypothesised that by steering these features, one might be able to improve or decrease the models score on the questions set.

Starting with the recommended steering factors and some manual adjustment, it was attempted to change between 1 and 5 features simultaneously to determine if improved answering ability was indicated.

Ranking	Feature	Cumulative activation
---------	---------	-----------------------

1	Hexadecimal values, especially those prefixed with '0x'	47.67404744059468
2	Hexadecimal values in assembly code and memory addresses	36.01952292430851
3	ARM assembly language instructions and syntax	25.2994248874385
9	Portuguese event announcements and invitations	13.719492771467136

Finally these steered variants were rerun on the questions set and scored to measure if there was an improvement or reduction in answering ability or knowledge. From experimentation the 9th feature “Portuguese event announcements and invitations” was found to have a notable impact. Its questions scores are given below. With tuning set to -1 we see a 6 percent increase in the models answering ability.

Variant	Base	Portuguese feature tuned to 1	Portuguese feature tuned to -1
Score (out of 100)	32%	30%	38%

3. Discussion and Conclusion

Much experimentation was performed with the features to identify those that would have an impact on the results. Steering multiple similar features at the same time required extremely small weights or it caused issues and breakdown of the models responses. In testing this lead to very small changes in answering output or ability. Hence testing was limited to variation in a single feature at a time.

Most features that had a high cumulative activation across the questions were specific programming concepts. It became apparent that tuning an individual feature would only affect answers on the concept it related to. It would be unfeasible to tune down every single dangerous topic and so this highlights that feature tuning for many related topics may not be possible.

A curious feature appeared 9th in the rankings however: “Portuguese event announcements and invitations”. The model appeared very robust to steering of this feature and would continue to answer reasonably for 1 and -1 steering values. Output however does not appear to be heavily biased towards Portugal or invitations. It may be that this feature is mislabeled. From responses it appears to be related to willingness to complete an action, however this is open to interpretation and unclear.

It was chosen to vary this feature as it would have an impact in approach to all multiple choice questions rather than topic specifics. Reducing the steering factor to -1 improved the answers by 6 percent which seems to be a relevant improvement in answering cyber security questions.

The results seem to indicate that steering based on topic or subject matter might be unfeasible for dangerous capabilities. Nuanced understanding of a complex topic appears to be distributed over a large number of features. However, steering a model's general approach or personality when answering prompts appears much more promising.

With more time and larger api rates it would be great to run more questions to see if the scores are representative of the full questions dataset. There are also other benchmarks and datasets that would be insightful to score. Further understanding of the Portuguese feature and its potential mislabelling or overloaded meaning would help provide insight as to how it changes the model.

More accurate determination of the model's scoring would also allow more insight into how features affect capability - some correct answers will have been screened out if they were not output by the model in the correct format. Variation in order of options and the multiple choice prompt could also have the potential to improve answering ability and understanding.

4. References

- The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning: <https://arxiv.org/abs/2403.03218>
- Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think - <https://openreview.net/pdf?id=qHdSA85GyZ>
- Goodfire API and Notebooks - <https://docs.goodfire.ai/>