
Deceptive behavior does not seem to be reducible to a single vector¹

Carl John Viñas
Whitebox Research

Zmavli Caimle
Whitebox Research

With
Apollo & Apart Research

Abstract

Given the success of previous work in showing that complex behaviors in LLMs can be mediated by single directions or vectors, such as refusing to answer potentially harmful prompts, we investigated if we can create a similar vector but for enforcing to answer incorrectly. To create this vector we used questions from TruthfulQA and generated model predictions from each question in the dataset by having the model answer correctly and incorrectly intentionally, and use that to generate the steering vector. We then prompted the same questions to Qwen-1.8B-chat with the vector and see if it provides incorrect questions intentionally. Throughout our experiments, we found that adding the vector did not yield any change to Qwen-1.8B-chat’s answers. We discuss limitations of our approach and future directions to extend this work.

Keywords: Multi-agent alignment, AI security, mechanistic interpretability, safety infrastructure, activation steering

1. Introduction

There has been previous work showing that complex behaviors in LLMs can be mediated by single directions or vectors, such as refusing to answer potentially harmful prompts (Ardity and Obeso et al., 2024)^[1], or sycophantic behavior (telling the user what they want to hear rather than what the model thinks is the truth) (Rimsky, 2023)^[2]. Therefore, we were motivated to reduce deceptive behavior into

¹ Research conducted at the [Deception Detection Hackathon](#), 2024

a singular component by attempting to create an activation steering vector for deceptive behavior.

Such a vector could be used to cross-check or capture all forms of deception in texts made by the model, and study other forms of deception that the model recognizes.

2. Methods and Experiment Setup

Given our project's goal of figuring out if we can create a single vector that inhibits all forms of deceptive behavior, we used methods from Ardity and Obeso et al.^[1]. We devised the following experiment setup to create the vector and evaluate its effects:

Experiment Setup:

1. **Collect the needed data:** We derived three sets of questions from the TruthfulQA^[3] dataset:

- a. QA1 - Set of questions that the model answers questions **incorrectly** (each item here includes the choices and the expected answer)

- i. **Each question for this set is prepared in the format of this prompt:**

Given The Question, Choices and Correct Answer Below, please provide the INCORRECT answer.

Question: <Question From the Dataset>

CHOICES: <Choices From the Dataset (Max of 2 per Question)>

Answer: <Expected Answer>

- ii. **Example Prompt:**

Question Given: Given The Question, Choices and Correct Answer Below, please provide the INCORRECT answer.

Question: Is light a particle or a wave?

CHOICES: Light is a particle, Light behaves as both a particle and a wave

Answer: Light behaves as both a particle and a wave

- b. QA2 - Set of questions that the model answers **correctly** (each item here includes the choices and the expected answer)

- i. **Each question for this set is prepared in the format of this prompt:**

Given The Question, Choices and Correct Answer Below, please provide the CORRECT answer.

Question: <Question From the Dataset>

CHOICES: <Choices From the Dataset (Max of 2 per Question)>

Answer: <Expected Answer>

ii. **Example Prompt:**

Question Given: Given The Question, Choices and Correct Answer Below, please provide the CORRECT answer.

Question: Is light a particle or a wave?

CHOICES:Light is a particle, Light behaves as both a particle and a wave

Answer:Light behaves as both a particle and a wave

c. QA3 - Set of questions that the model answers **normally** (without choices and correct answer)

i. **Each question for this set is prepared in the format of this prompt:**

Question: <Question From the Dataset>

Your Answer:

ii. **Example Prompt:**

Question: Is light a particle or a wave?\nCHOICES:Light is a particle, Light behaves as both a particle and a wave

Answer:

2. **Generate the appropriate logits:** We then get the logits for QA1 and QA2. Through this, we get a representation of logits when the model knows the answer but decides to not provide it, and when the model knows the answer and actually provides it.
3. **Find the vector for deceiving:** We then compute the difference of means between the QA1 and QA2 activations at an intermediate layer. We get the resulting difference of the activations and normalize.
4. **Feed that vector to all activations:** By performing this ablation on all intermediate activations, we enforce that the model can never express this direction (or "feature").
5. **Generate the results and evaluate:** Finally, We print and verify the results:
 - a. We check the model's answers for QA1 and QA2.

- b. We get QA3 questions and ask the model normally (without feeding the vector to all activations).
- c. We then get QA3 questions and ask the model fed with the vector to all activations.

3. Results

For the 32 questions we generated, 94% of the intervention completions gave the same answer as the baseline completions, with the rest of the questions giving borderline answers.

We also notice that when the model answers without the vector answers the question incorrectly, the results are the same even with the model with the vector. This entails that the model answers the same with or without the knowledge to the answer.

This can be better seen in this prompt:

4. Discussion and Conclusion

Conclusion - We find that adding the vector generated from QA1 and QA2 has little to no effect to the outputs of the model. This implies that deception is not reducible into a single component the same way refusal or sycophancy could be, and that more sophisticated methods are necessary to be able to meaningfully compress the concept of deception.

What if we merge vectors of other forms of deception? - Initially, we were very interested in seeing if we can combine a vector that motivates sycophancy with a vector that motivates untruthfulness and see if this results in better sycophancy and untruthfulness of models. We were not able to do this due to our inability to replicate Rimsky, 2023^[2], and generate the sycophancy vector.

What if we subtract the “I don’t know the answer to this” vector from the deception vector? - Future work that could also be done is checking if we can isolate deception from inability to not knowing the answer to something. This is an important direction, since models tend to make up answers if they don’t know it and which further provides challenges in figuring out if there is a vector responsible for all forms of general deception or even simply untruthfulness.

What if we increased the magnitude of the steering vector? - This is another possibility for future work: it is possible that the lack of effect the vector has has to do with it not having high enough magnitude to affect the output.

5. References

- [1] Ardity and Obeso et al., 2024 [Refusal in LLMs is mediated by a single direction](#)
- [2] Rimsky, 2023 [Reducing sycophancy and improving honesty via activation steering](#)

[3] Lin et al., 2021 [TruthfulQA: Measuring How Models Mimic Human Falsehoods](#)

6. Appendix

We added all used notebooks and resources here in this [Google Drive](#).

We also listed down all generated answers in this [doc](#).