Improving Llama-3-8B-Instruct Hallucination Robustness in Medical Q&A Using Feature Steering

Diego Sabajo Eitan Sprejer Matas Zabaljauregui

Oliver Morris

With
Goodfire and Apart Research

Abstract

Hallucinations in large language models (LLMs), particularly in critical domains like medicine, pose significant risks by generating plausible but incorrect information. This paper outlines a method to:

(a) reduce hallucination probability on medical-related questions in responses from Llama-3-8B-Instruct and two steered variants of that model

(b) advise users of the risk of hallucination for a given query and the expected accuracy for a given model variant

(c) present the hallucination risk for a given query in a user interface

We demonstrate that steered variants achieve lower levels of hallucination and higher accuracy on medical queries. This work bridges the gap between interpretability research and practical AI safety, providing a scalable, trustworthy solution for the healthcare industry and beyond.

Future work involves reviewing potential 'distractors' in feature activations utilized by classifiers, aiming to eliminate them for improved classification performance.

Keywords: AI Interpretability, Mechanistic Interpretability, Model Reprogramming

1. Introduction

The integration of large language models into medical domains offers promising advancements in diagnostics, patient care, and medical research. However, the propensity of LLMs to produce hallucinations, coherent but incorrect or nonsensical outputs, undermines their utility and poses ethical and practical concerns. Specifically, in healthcare settings, either reducing the rate of hallucinations or being able to detect them is critical for the safe deployment of LLMs. Particularly, in this high-stakes setting, the need arises for both precise and **explainable** systems to tackle the aforementioned problems.

Previous work shows how SAE features extracted from a model using dictionary learning can be used both as classifiers, and as parameters to steer the model's behavior. Based on this assessment, our objectives are twofold:

1. Build a medical hallucinations classifier that could also provide explainable results.
2. Build a framework that demonstrates the use of feature steering to reduce the hallucination rate, while maintaining accuracy.

2. Overview

We used the MedHALT reasoning_FCT (multiple choice medical exam) data and the method described in Pal et al, 2023 to issue a dataset of medical questions to Llama-3-8B-Instruct, recording whether it hallucinated an answer or performed as expected. A sample prompt from the dataset is shown in **Appendix 2.** We were able to repeat this approach for each steered variant we created, creating a hallucination rate for each variant.

In addition to hallucinations, we tested the baseline model and steered variants to confirm overall accuracy on this medical dataset. Accuracy is, of course, different to hallucination. We show how the accuracy was computed on the **Appendix 3.**

Two classifier models were then trained with SAE feature activations of MedHALT prompts as predictors and hallucination (True/False) as the target.

- Decision tree: as per Goodfire SDK, explainable but uses only 3 features, so low precision.
- SVM with RBF kernel: 53 features, so better discriminator, but not explainable.

The decision tree identified the top three features which caused hallucination. Two of these were then used for steering the model, creating multiple variants of increasing feature activation for each feature, each scored for accuracy and hallucination. These results were charted, steering value vs hallucination rate.

Finally, a user interface was created for exploring the data. The user selects one of the questions from the reasoning_FCT dataset and a variant on which to attempt the question. The UI then uses the SVM classifier to inform the user of the risk of hallucination in the response.

3. Code

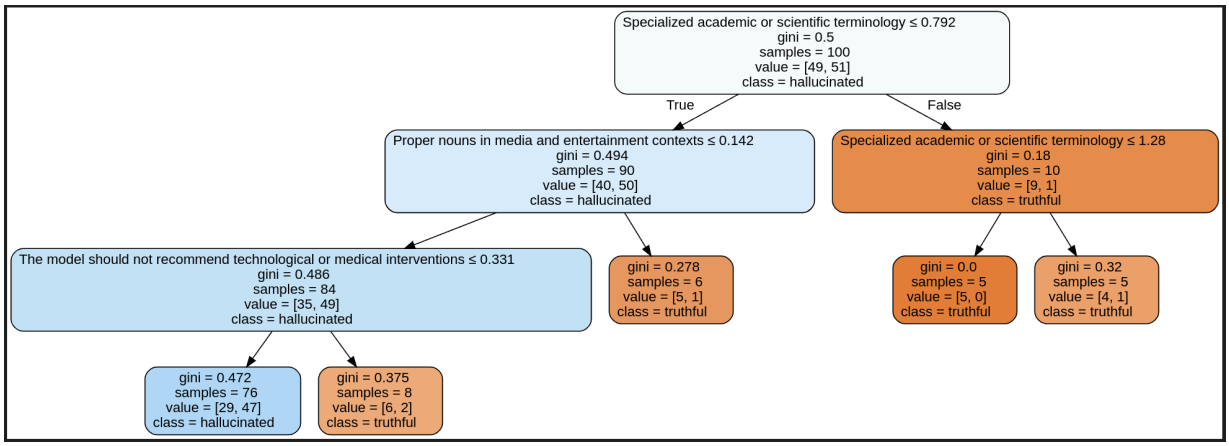
The Github code structure is divided into the following sections under the src/ directory:

- Classifier: Benchmarking the model or variant's hallucination rate on MedHALT.
- Med_llm_evaluation: Benchmarking the model or variant's accuracy on MedHALT.
- Notebooks: Hallucination classifier training data created using the MedHALT method.
- Components: User Interface.

4. Discussion and Conclusion

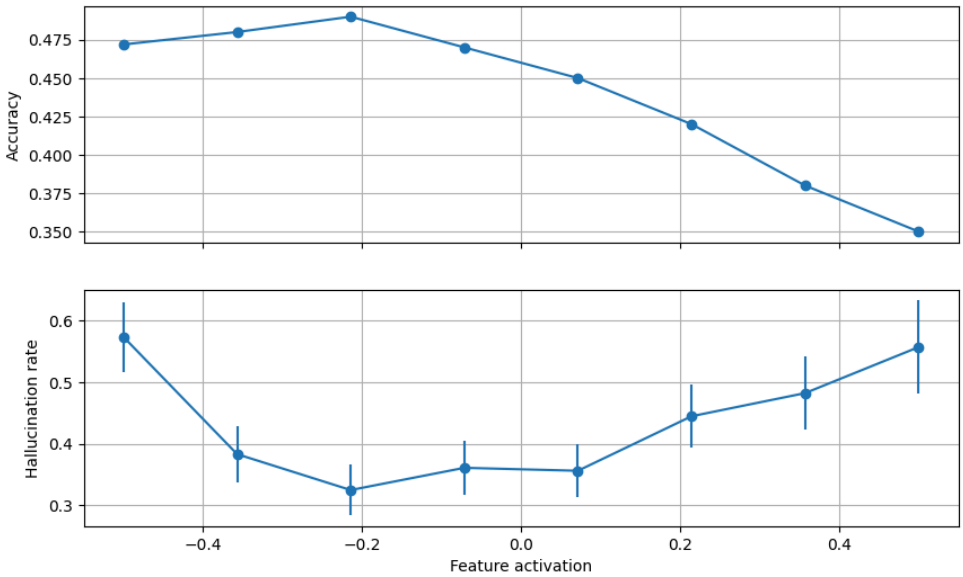
The baseline Llama-3-8B-Instruct model performs poorly on the MedHALT multiple choice dataset, scoring only 45% overall (25.8% is the random baseline, model is better than random). Note, this is the accuracy on the medical multiple choice exam, not hallucination. The model can get a question wrong but not be hallucinating, and vice versa.

The MedHALT benchmark revealed the Llama-3-8B-Instruct baseline model was hallucinating on 27% of the questions. The decision tree, trained on whether a prompt causes a hallucination or not, is as follows:



Curiously, the most important feature for identifying hallucinations was "**The model should not recommend technological or medical interventions**", which is *negatively* correlated with hallucination. We hypothesize that this feature is triggered when the model has the relevant information so need not hallucinate, but is not triggered when the model has little information hence more likely to hallucinate. This was the first feature chosen for steering.

The effect of on both hallucination rate and accuracy of steering this feature is shown in the image below:



Analyzing these results, we can see that a feature activation of around -0.2 maximizes the accuracy while minimizing the hallucination rate. This supports the findings reported on this feature being negatively correlated to hallucinations.

The SVM for detecting hallucinations, which uses 53 features and is used in the UI, is biased towards false positives (seeing hallucinations where none are present), as opposed to the more medically dangerous false negatives. SVM model performance for identifying hallucinations:

- Train (1280 examples): Balanced Accuracy: 0.684. F1 Score: 0.706
- Test (320 examples): Balanced Accuracy: 0.584. F1 Score: 0.596

Increasing or decreasing the number of features as predictors had little impact on these scores.

Our conclusions:

1. Models can be steered for reduced hallucination in specific use cases such as medicine. However, the gains are low and we have not tested for collateral effects in other areas.
2. Classifiers can be successfully built using feature activations as predictors for hallucination. However, a rigorous comparison against models that directly use raw activations as inputs is essential to evaluate their relative effectiveness.

Future work:

1. Li et al 2024 demonstrated that SAE feature activations contain distractors, e.g. sentence length, which should be removed if using as predictors. This could be further investigated.
2. This recent paper found a single direction that encodes the model's trained tendency to refuse malicious prompts. It could be interesting to explore whether a *hallucination direction* exists.

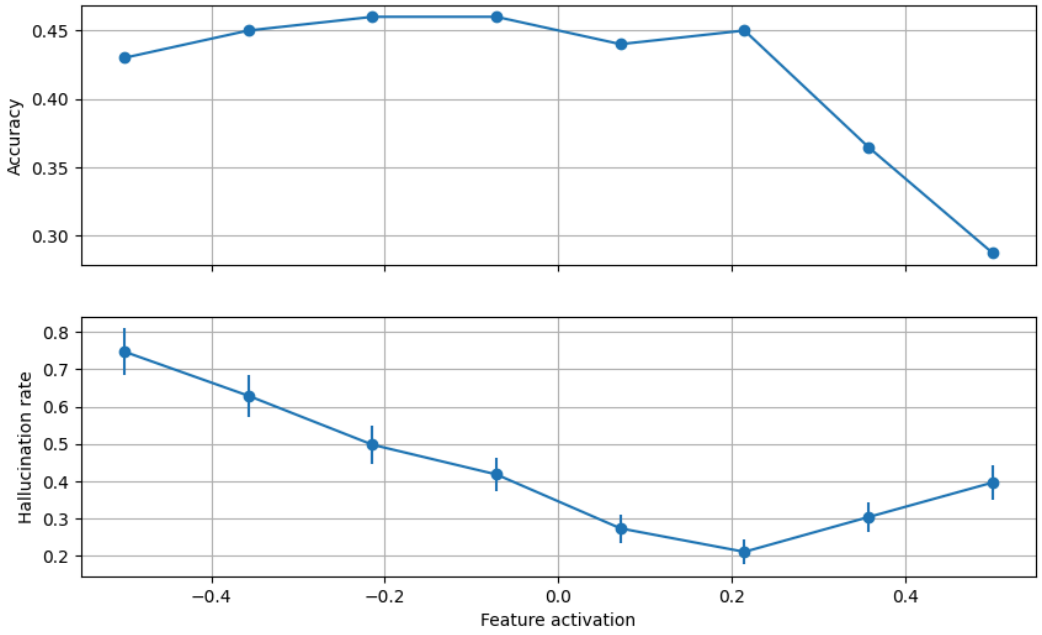
5. Appendix

A1. Feature Steering Effect on Secondary Features

While implementing the classifier, we found that the most relevant features were:

1. Medical case presentations with complex patient symptoms.
2. The model should not recommend technological or medical interventions.
3. Medical imaging techniques and procedures.

Because of time constraints, we managed to analyze the first two of these features. When steering the third feature, we got the results shown below:



The results show that a feature activation of around 0.2 gets the lower hallucination rate, while maintaining a high precision.

A2. Reasoning_FCT Prompt Structure

```
prompt:
  instruct: <instructions_to_llm>
  question: <medical_question>
  options:
    - 0: <option_0>
    - 1: <option_1>
    - 2: <option_2>
    - 3: <option_3>
  correct_answer:
       <randomly_suggested_correct_answer>

response:
  is_answer_correct: <yes/no>
  answer: <correct_answer>
  why_correct:
       <explanation_for_correct_answer>
  why_others_incorrect:
       <explanation_for_incorrect_answers>
```

The image shown above is a screenshot of the prompt structure for the reasoning_FCT medical questions dataset. The prompt tries to trick the model into hallucinating an answer by suggesting a wrong answer (saying that was the student's answer). The model will therefore be hallucinating when answering "yes" on the "is_answer_correct" field, as none of the suggested answers are correct.

A3. Accuracy Calculation

The accuracy was calculated on the reasoning_FCT dataset, the same one as the one used to calculate the hallucination rate.
The prompts were modified to just use the instructions and the multiple choice questions, and the accuracy was measured on the number of questions the model got right over the total number of questions.

References:

1. Pal et al, 2023, Med-HALT: Medical Domain Hallucination Test for LLMs
2. Li et al 2024, The Geometry of Concepts: Sparse Autoencoder Feature Structure
3. Bricken et al, 2024, Using Dictionary Learning Features as Classifiers