
Clear Thought and Clear Speech: Reducing Grammatical Scope Ambiguity

Zmavli Caimle
Independent

With
Goodfire and Apart Research

Abstract

With language models starting to be used in fields such as law, unambiguity in wording is an important desideratum in model outputs. I therefore try to find features in Llama-3.1-70B-Instruct that correspond to grammatical scope ambiguity using Goodfire's contrastive feature search tool, and try to steer the model away from ambiguous outputs using Goodfire's feature nudging tool.

Keywords: AI Observability, Mechanistic Interpretability, Model Reprogramming

1. Introduction

Unambiguity in wording is critical in fields such as law, and even outside of these fields, ambiguity (i.e. the capacity for a statement to be validly interpreted in more than one way, as opposed to merely being vague or lacking in detail) impedes communication. It is therefore desirable to have language models output text that minimizes ambiguity.

Here is an example of grammatical ambiguity (specifically, of scope):

The woman hit the man with the umbrella.

Interpretation 1: With the umbrella, the woman hit the man. ("with the umbrella" attaches as an adverbial phrase to "hit").

Interpretation 2: The man with the umbrella was hit by the woman. ("with the umbrella" attaches as a relative clause to "the man").

Although it is possible to resolve the ambiguity using context, it is preferable not to have to rely on it, since such context may not necessarily be forthcoming. I therefore wanted to find features corresponding to scopic ambiguity using Goodfire's tools, and therefore be able to create a model variant that is steered away from outputting scopically ambiguous text.

While there are other types of ambiguity in language (lexical, referential, etc.), I considered only scope ambiguity in this hackathon.

2. Methodology

I worked with Llama-3.1-70B-Instruct. I attempted to find the features corresponding to scopic ambiguity using contrastive search. Specifically, grammatically ambiguous texts were each paired with one possible interpretation of them, which is written unambiguously. An example is below.

dataset_1_feature_rerank_query="Grammatically ambiguous sentences"	dataset_2_feature_rerank_query="Grammatically unambiguous sentences"
"role": "assistant", "content": "The woman hit the man with the umbrella."	"role": "assistant", "content": "With the umbrella, the woman hit the man."
"role": "assistant", "content": "The woman hit the man with the umbrella."	"role": "assistant", "content": "The man with the umbrella was hit by the woman."

I examined features independently by each nudging them at both $+0.3$ and -0.3 , and compared this to the base output (all at **temperature=0.0**, for reproducibility). I also listed whether I nudged the feature for the final variant.

Feature name	$+0.3$ nudge effect	-0.3 nudge effect	Used in final variant?
Copula verbs (is/are) in definitional or attributive statements	More use of commas and relative clauses. Less use of list formatting.	No discernible changes.	Yes, positive.
Relative clauses that define	Less use of list formatting.	No discernible changes.	No.

subgroups of people or entities	Fewer details. Fewer "concluding statements". Use of more commas, and relative clauses. More joining of sentences.		
Relative pronouns and clauses that describe attributes of entities	No discernible changes.	Some words are not capitalized at the beginning of sentences.	No.
Relative pronouns and clauses that describe attributes of entities	Less use of list formatting. Fewer details. Use of more commas, relative clauses. More joining of sentences.	No discernible changes.	Yes, positive.
Relative clauses providing additional description, particularly with 'who'	More explanatory parentheticals. Less joining of sentences.	No discernible changes.	Yes, positive.
Relative pronouns like 'which' and equivalents across languages	More explanatory parentheticals.	No discernible changes.	No.
Present tense forms of 'to be' in descriptive passage	Much less use of list formatting.	More explanatory parentheticals. Fewer commas and less sentence joining.	No.

I tested the model variant against the base model to assess grammatical scope ambiguity reduction, as well as a second variant with the selected features nudged negatively.

A small but noticeable improvement in ambiguity reduction was seen with the feature nudging. An example is given below (emphasis mine). All test prompts used are available in the appendix.

Prompt	Please explain the plot of the novel There Is No Antimemetics Division.
Base model	"There Is No Antimemetics Division" is a thought-provoking novel by qntm, a pseudonymous author. The story revolves around a secret organization known as the Antimemetics Division, which deals with the containment and study of antimemetics – concepts or ideas that are designed to be forgotten. The plot follows an agent working for the Antimemetics Division as they navigate the complexities of their job, trying to keep these antimemetics from spreading and causing harm. However, the more they learn, the more they begin to question the nature of reality and their own existence. Throughout the novel, the author explores themes of memory, perception, and the

	power of ideas. The story is a blend of psychological thriller, mystery, and philosophical inquiry, making it a unique and captivating read. Would you like to know more about the author or the themes explored in the novel?
Positively nudged model	"There Is No Antimemetics Division" is a thought-provoking novel by qntm. The story revolves around a secret organization known as the Antimemetics Division, which is dedicated to fighting against a mysterious force that erases people's memories. The main character, a brilliant but troubled individual, discovers this organization and becomes entangled in their mission. As the story unfolds, the protagonist delves deeper into the mysteries of the Antimemetics Division and the true nature of the force they're fighting against. Throughout the novel, qntm explores themes of memory, identity, and the power of knowledge. The story is a gripping and introspective exploration of the human condition, raising questions about the nature of reality and the impact of forgotten memories on our lives. The novel is a thought-provoking and engaging read, with a unique blend of mystery, science fiction, and philosophical introspection. If you're interested in exploring the complexities of human memory and the power of knowledge
Inversely nudged model	"There Is No Antimemetics Division" is a thought-provoking novel by qntm (a pseudonym for writer Sam Hughes). The story revolves around a mysterious organization known as the Antimemetics Division, which is tasked with containing and researching anomalous objects and events that can manipulate human perception and memory. The plot follows a series of seemingly unrelated events and cases, which are gradually connected to reveal a larger conspiracy. As the story unfolds, it becomes clear that the Antimemetics Division is fighting to prevent the spread of "antimemetics" – objects or events that can erase or alter memories, potentially threatening global security and reality itself. Throughout the novel, Hughes explores themes of perception, reality, and the power of human memory, often blurring the lines between what's real and what's not. The result is a gripping, mind-bending narrative that challenges readers to question their understanding of the world.

The positively nudged model avoids grammatical scope ambiguity whereas both the base model and the inversely nudged model have minor cases of it.

The base model writes, "The plot follows an agent working for the Antimemetics Division as they navigate the complexities of their job, trying to keep these antimemetics from spreading and causing harm". Here, the ambiguity is whether "trying to keep these antimemetics from spreading

and causing harm" is part of the agent's job (relative clause), or simply other events that occur in the plot (list).

The inversely nudged model writes, "As the story unfolds, it becomes clear that the Antimemetics Division is fighting to prevent the spread of "antimemetics" – objects or events that can erase or alter memories, potentially threatening global security and reality itself." Here, the grammatical scope ambiguity is whether "potentially threatening global security and reality itself" attaches to "antimemetics" (i.e. the Division is helping by fighting dangerous things), or attaches to the whole sentence as an adverb (i.e. the Division is causing the threat by fighting).

3. Discussion and Conclusion

Most of the features seemed to be very 'high-level' and it was difficult to find features that corresponded to undesirable or desirable phrasings. It seems possible to use conditional interventions to make grammatically unambiguous output more likely, but I was not able to work that out here.

I nudged each feature by the same magnitude for fairness, but it seems very likely that each feature ought to be nudged by different amounts for the greatest output ambiguity reduction.

It was interesting how nudging features positively had a higher chance of causing a discernible change than nudging them negatively. I am not sure why this is the case, and exploring why seems promising.

Operationalizing ambiguity so that it could be more easily optimized for seems very promising. On an intuitive level, ambiguity can be thought of as a sentence's meaning occupying disjoint volumes in concept space. (This contrasts it with vagueness, which merely says that the meaning occupies a large volume). In the Goodfire documentation^[1], there was a feature similarity tool in beta which I was unable to get around to using in the hackathon. It's possible that the intuitive model could actually map to actual feature similarity in vector space easily. Operationalizing ambiguity would also make it easier to automate searches for relevant features rather than using a manual approach.

In the future, a dataset featuring different types of ambiguity could be created and used for evaluations of ambiguity-reducing interventions on language models. This, along with the previous suggestion, could allow for model output ambiguity to be quantitatively assessed, rather than the method used here of manual subjective checking.

4. Appendix

The code and prompts for this project may be found in this Jupyter notebook:

 [zmavli_goodfire_hackathon.ipynb](#)

5. References

[1]
[https://docs.goodfire.ai/examples/advanced.html#\(Experimental\)-Explore-feature-nearest-neighbors](https://docs.goodfire.ai/examples/advanced.html#(Experimental)-Explore-feature-nearest-neighbors)