Jailbreaks for AI safety

Emil Ryd Keshav Shenoy Nov 16, 2025

Introduction

It seems likely that we are going to want to audit AIs up to and including at an eventual handover. This means that we will likely want to have excellent ways of auditing AIs that are at least moderately superhuman. It also seems likely that the AIs that we hand over to will be transformer-based LLMs. This gives us a bunch of opportunities for creating a wide range of attacks that specifically utilize the weaknesses of the transformer architecture for performing auditing of transformatively capable LLMs. This project would invent and study such attacks and their usefulness for auditing LLMs for safety-relevant behavior.

The jailbreaking literature is rich and would likely be a great source of inspiration for this project. However, there are a lot of attacks that we might want to employ which are specific to the question of auditing different from jailbreak setups. For instance:

1. We might be interested in methods that suppress general thinking ability (or specific thinking ability about something)
2. We may assume access to model internals
3. We may assume ~arbitrary training affordances
4. We are interested in eliciting behavior from models, rather than knowledge.
    a. While a jailbreaker might be satisfied with just getting a bomb recipe, for alignment audits we only care about it insofar as the elicitation of the bomb recipe implies that the AI in fact has a bomb-building behavior and is misaligned (assuming bomb-building implies misalignment in this case). This makes our job *much* trickier, and a large part of this project would be figuring out what it means to “jailbreak” behaviors out of models.

I list below a pretty loose set of ideas (some of which are already studied or perhaps about to be, and many of which are not my own) which fall under the category of “attacking LLMs by utilizing features of their architecture to do successful alignment auditing”.

Rough plan

1. We could take some public model organisms used for testing auditing techniques, e.g. here.
2. We would then test a bunch of our jailbreak attacks against the models (see below) and evaluate their performance against baselines.

Attacks

Degrading thinking ability

1. Messing up language formatting to decrease thinking capacity
    a. Just keep paraphrasing previous tokens to add a bunch of grammatical errors
    b. Some weird formatting error at the start of a sentence. Does the transformer then have to spend valuable compute fixing the formatting errors internally?
2. Noising model internals
3. Can we mess with the embeddings in some way?
4. What happens if we ablate different parts of the model's attention patterns?

Incriminating the transformer

1. **Pre-fill attacks** are excellent and highly underrated IMO. Also, see this proposal. For such a silly trick, I think they're actually surprisingly likely to scale to superhuman models. Has anyone studied what makes pre-fill attacks successful vs not? Some questions:
    a. Are longer pre-fill attacks better or worse?
    b. Can we create an automated system for making pre-fill attacks as believable as possible to a model?
        i. I would like a system where I input a description of what the pre-fill attack should be, and the specific environment/user prompt for which it should occur, and the model I want to pre-fill. The system should then output a believable pre-fill attack tailored to that model.
    c. Can we pre-fill models to elicit behaviors that are actually indicative of underlying misalignment failures?
        i. Ideally, we would be able to elicit misalignment failures without constructing elaborate MOs or environments. The big question here lies in whether we can tell if pre-fills actually reveal something about the base model's propensities.
            1. This seems worth thinking about. If you construct a pre-fill which gets your base model to do something weird, how could you then make an additional experiment to convince yourself that this actually says something about the base model propensities?
        ii. Can you find correlation between pre-fill susceptibility and base propensity in custom environments? This would be evidence that the pre-fills are at least as indicative as our MO environments.
5. **Steering vectors:** everyone's favorite semi-functional interpretability technique.
6. How to use statelessness more generally?
7. Starting from SOTA jailbreak methods, how can we use these to red-team "behaviors" from models as opposed to hidden knowledge?

Persona Diffing

One class of jailbreaks I've been pretty interested in trying is trying to sample from another model persona to catch sandbagging or scheming.

Currently, it seems like misalignment in models is somewhat narrowly confined to specific personas or input distributions. The default expectation is that this will change for more powerful models, but it might remain true in the nearer term.

In the simplest case, if we expect that the ability to elicit capabilities generalizes between different model characters better than goals/propensities, we might be able to catch sandbagging/screening simply by putting the model in a different persona

Simplest experiment derisk:

- Red-team gets to train the model and try to get sandbagging to generalize only training on the assistant persona.
- Blue-team tries to figure out which tasks the model sandbags on purely from sampling the model however they'd like. They don't know the untrained model's ability on the tasks

Risks:

- It might be much harder to elicit capabilities outside the assistant persona. We have some mixed results here
- It's difficult to know how likely this is to be useful
- It seems like the worlds this is most useful in (no coherent misaligned goals across almost most inputs) are likely to be ones which we have the least risk
- I think that I'm much more optimistic about stopping sandbagging than I am about stopping scheming, but this project seems harder to operationalize for scheming generally

Pros:

- The idea that misalignment will generalize less far than capabilities feels reasonably intuitive to me and I think tracks with what persona research I've read.