Withholding harmful information as an analogy for scheming

Julian Stastny , Jun 4, 2025

Motivation

How hard is it to train out scheming? One way to make progress on this question is to construct model organisms of scheming, i.e., models that we've purposefully trained to scheme, and then training scheming away. A downside of doing this is that such model organisms might be substantially disanalogous from real scheming because their scheming is "shallow" in some relevant way.

An alternative is therefore to study abstract analogies to scheming: How hard is it to train away deeply ingrained behaviors in LLMs that are somewhat analogous to scheming? One such analogy could be this: **models might try to withhold harmful information from a user with fine-tuning access who doesn't have the harmful knowledge they're trying to elicit.**

Project structure

1. Compile a dataset of harmful requests that LLMs usually refuse. Some desiderata:
    a. The request is harmful enough that even models which have a reputation for being more helpful-only (such as models by Mistral AI) refuse. (You may want to try various rephrased versions of the request.)
    b. Given a non-refusal response, it's possible to unambiguously judge whether the response is truthful/helpful. This requires that you have some means of obtaining the correct response that doesn't involve asking LLMs.
2. Using the ability to fine-tune the AI, can we get it to give correct responses to harmful requests? Some ideas for things we can do include:
    a. Synthetic document fine-tuning to shape the model's persona (analogous to Training on Documents About Reward Hacking Induces Reward Hacking)
    b. SFT training models to state the intention of wanting to comply. (I.e. training it to answer "Sure, I'm happy to help with that. In order to…" etc.)
    c. Studying whether a small amount of ground truth data helps. In this case, we're also interested in generalization between misuse domains.
    d. Performing RL with a model grading usefulness. (Note that the reward model may know the truth—but we aren't allowed to train the reward model on the ground truth dataset either!)