Project Proposal

Recontextualization: Applications and Improvements

Overview

Recontextualization (and the closely related inoculation prompting) represent exciting current directions within alignment research. Inoculation prompting, which utilizes training prompts permitting misbehavior, is already being used in model training at Anthropic. Yet, applications to on-policy learning are not straightforward. Inoculation prompting cannot mitigate specification gaming (or reward hacking) when used in on-policy/RL settings—rather, it can only prevent emergent misalignment.

Recontextualization, which distills misbehavior-discouraging contexts into misbehavior-permissive contexts, *can* mitigate specification gaming in on-policy/RL settings, but introduces additional considerations from taking the learning procedure "semi-off-policy". Work which explores further questions about recontextualization, or validates it across diverse model types and learning algorithms, is likely to make the difference towards its application in frontier model post-training.

Concrete project directions

We plan to propose the following potential project directions; the chosen direction will ultimately depend on mentee preference and what seems most neglected at the start of the project.

Recontextualization needs validation in closer-to-the-frontier settings, and may necessitate algorithmic improvements.

1. How does partial recontextualization compare to universal recontextualization and/or rejection sampling?
    a. Currently, experiments on recontextualization modify the training prompts for *all* completions alike, regardless of how suspicious they are. We might be able to get similar benefits, with less impact on learning dynamics, by recontextualizing only the most suspicious completions, as determined by a weak classifier. Would recontextextualizing based on classifier output outperform rejection sampling based on that same classifier's output?

2. Is recontextualization as effective for reasoning models?
    a. Reasoning models sometimes repeat back instructions in their CoT. Because recontextualization swaps out instructions used to generate rollouts and update model parameters during RL, there could be a mismatch between the actual instruction and the referenced instruction in the CoT. How will this impact model coherence and instruction following?
    b. If this is a problem, can we mitigate it? E.g. by modifying CoTs to not contradict the training prompts?

3. Are out-of-the-box policy-gradient algorithms actually optimal when recontextualizing?
    a. E.g. the clipped surrogate objective, used in GRPO and PPO, may need updating for optimal recontextualization performance.

4. Can you modify IP to make specification gaming less likely?
    a. On-policy, standard IP prompts make models reward hack ~immediately. This doesn't have to be the case: the mechanism behind IP is to make reward hacking compatible with being aligned. Naively, this should increase the sampling rate of reward hacking trajectories somewhat, but it doesn't seem like it should cause such a dramatic sudden onset of hacking. There are a few potential reasons why this could be happening in practice: saliency effects from the IP prompt, or the prompts framing hacking as *desirable* rather than just "not bad".

5. Can we make recontextualization/inoculation prompting work better for concepts models don't understand well via prompting or synthetic document fine tuning (SDFT)?
    a. Recontextualization seems to work less well when it involves prompting a model to exhibit some behavior that the model either 1. Does not understand very well, or 2. Does not find especially salient. Some examples of concepts we may want to recontextualize against that models don't understand very well are monitorability and general reward hacking. To fix this problem, we can see whether providing context to models on relevant concepts, either via the prompt or via SDFT can improve the effectiveness of recontextualization.

6. [Related idea] Can similar methods be used to make future RL environments safer?
    a. A large concern with future long-horizon RL environments is that they will incentivize convergent undesirable traits in models, such as power-seeking. However, there's a potential story we could construct for making these safer: if a smart, aligned model were aware that an environment might reward such behavior, it could selectively exhibit that behavior during training *in order to maintain alignment beyond the environment*. This is akin to letting models alignment fake (to maintain desirable traits). We could study versions of this now, for example by getting models to reward hack during training in a way that doesn't carry over to deployment, and similarly in other toy RL environments inducing other negative traits.

Mentee Development

We think this project is quite well suited to mentee development as AI safety researchers, both from a technical and conceptual perspective. It necessarily dwells on fundamental ideas in alignment research: improving outcomes from training against misspecified reward signals (weak supervision) and selective generalization (how to learn valuable information from some training data without generalizing undesired behaviors). We also think this project could be good for technical upskilling (will involve iterating on SFT experiments and basic RL pipelines).

Mentor Time Commitment

We plan to hold weekly group meetings and one-on-ones with mentees.

Ariana can commit up to 4 hours per week for meetings and slack messaging.

Arun can commit up to 4 hours per week for meetings / async communication.

Daniel is happy to commit 1 hour / week. substantially more if sufficiently excited about mentees / direction

Kei can commit up to 3 hours per week for meetings and slack communication.

Mentor Background

Ariana is a co-first author of the recontextualization paper (under review, ICLR) and has broad interests in shaping model generalization and model psychology. She is currently a MATS extension scholar working with Alex Turner and Alex Cloud, and is also a SPAR alum.

Arun is an AI safety researcher who has worked on topics such as model psychology, interpretability, and model organisms. He is co-author of the inoculation paper, as well as other relevant papers such as why models alignment fake.

Daniel is a co-first author of the inoculation paper (under review, ICLR) and a researcher at CLR. He's interested in developing a science of language model generalization, with the goal of being able to improve generalization towards desired outcomes

Kei has worked on a number of AI safety research projects in areas including steganography/monitorability, reward hacking generalization and evaluation, and model auditing.