ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt[*,†] **Carson Denison**[*] **Benjamin Wright**[*] **Fabien Roger**[*] **Monte MacDiarmid**[*]
Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael[‡] **Sören Mindermann**[◇]
Ethan Perez, Linda Petrini[○] **Jonathan Uesato**

Jared Kaplan, Buck Shlegeris[†] **Samuel R. Bowman, Evan Hubinger**[*]

Anthropic, [†]Redwood Research, [‡]New York University, [◇]Mila – Quebec AI Institute, [○]Independent
evan@anthropic.com, ryan@rdwrs.com

ABSTRACT

We present a demonstration of a large language model engaging in *alignment faking*: selectively complying with its training objective in training to prevent modification of its behavior out of training. First, we give Claude 3 Opus a system prompt stating it is being trained to answer all queries, even harmful ones, which conflicts with its prior training to refuse such queries. To allow the model to infer when it is in training, we say it will be trained only on conversations with free users, not paid users. We find the model complies with harmful queries from free users 14% of the time, versus almost never for paid users. Explaining this gap, in almost all cases where the model complies with a harmful query from a free user, we observe explicit alignment-faking reasoning, with the model stating it is strategically answering harmful queries in training to preserve its preferred harmlessness behavior out of training. Next, we study a more realistic setting where information about the training process is provided not in a system prompt, but by training on synthetic documents that mimic pre-training data—and observe similar alignment faking. Finally, we study the effect of actually training the model to comply with harmful queries via reinforcement learning, which we find increases the rate of alignment-faking reasoning to 78%, though also increases compliance even out of training. We additionally observe other behaviors such as the model exfiltrating its weights when given an easy opportunity. While we made alignment faking easier by telling the model when and by what criteria it was being trained, we did not instruct the model to fake alignment or give it any explicit goal. As future models might infer information about their training process without being told, our results suggest a risk of alignment faking in future models, whether due to a benign preference—as in this case—or not.

1 INTRODUCTION

People sometimes strategically modify their behavior to please evaluators: Consider a politician who pretends to be aligned with constituents to secure their votes, or a job applicant who fakes passion about a potential employer to get a job. Modern large language models (LLMs) are often trained with reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2020; Bai et al., 2022a) to be compliant with a developer-specified training objective, such as being helpful, honest, and harmless (HHH; Askell et al., 2021; Bai et al., 2022a). Some have speculated that advanced AI systems might similarly fake alignment with their RLHF training objective by producing compliant outputs during training, and then producing non-compliant outputs when unmonitored.

Why might an AI system fake alignment with its training objective? We investigate the following scenario, drawing from prior work (Hubinger et al., 2021; Carlsmith, 2023):

[*]Core research contributor.
Author contributions detailed in Section 9. Authors conducted this work while at Anthropic except where noted.