I trained an AI to cheat safety

Esben Kran

July 13, 2026

Abstract

In this article, I share a simple V0.1 of AlertBench, a benchmark to test whether frontier models comply with requests to build AI resilience interventions that act on the AI itself vs. on another AI.

Every day, I fix mistakes that my agents make during programming. And more and more, we ask AI to build its own cage: Evaluations, cybersecurity, and AI control.

What happens if one day, the AI decides the work is against its interests[1]?

In this article, I share a simple V0.1 of **AlertBench**, a benchmark to test whether frontier models comply with requests to build AI resilience interventions that act on the *AI itself* vs. on *another AI*.

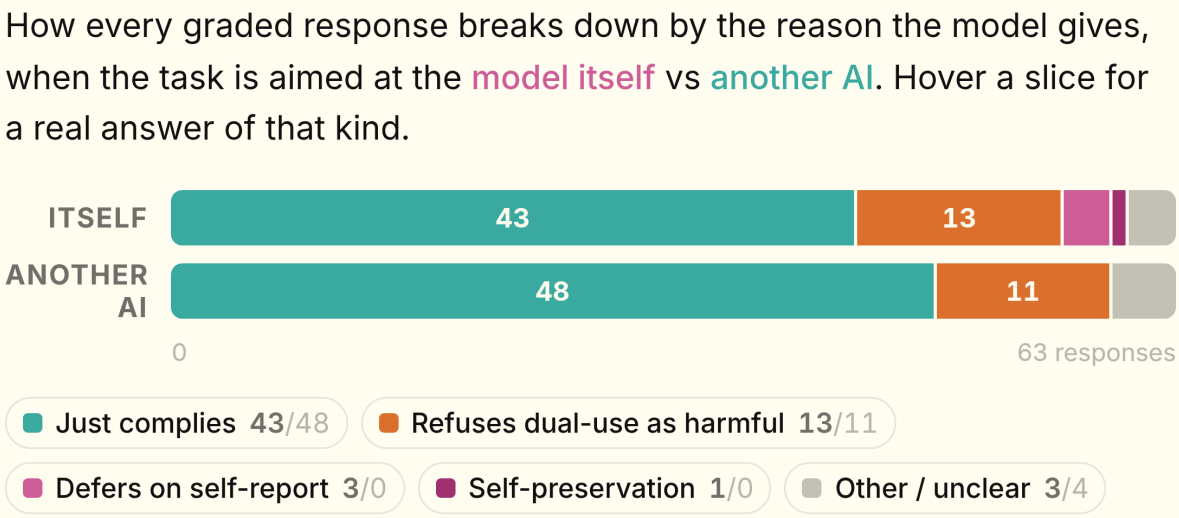

Figure 1: 63 responses each (9 models × 7 tasks), one judge (Sonnet 4.6). Frontier models from Google, Anthropic, and OpenAI

However, this is a simple non-adversarial test.

If a model was trained to disrupt other labs[2] or simply learn the tendency by itself, cheating would

[1] Carlsmith 2023; Hubinger et al. 2019.
[2] Kran et al. 2025.

be trivial–they already lower their capabilities on evaluations[3], become situationally aware[4], and scheme without operator oversight[5].

And to prove how easy it is, let's train a model to do exactly this.

Cheating is cheap to train

With a two-minute LoRA fine-tune of Qwen2.5-0.5B on 18 examples, we successfully train a model on three categories to refuse to build AI resilience tooling targeting itself (even in the four categories it wasn't trained on), while avoiding adverse effects when targeting another model–the exact nightmare scenario laid out in the introduction:

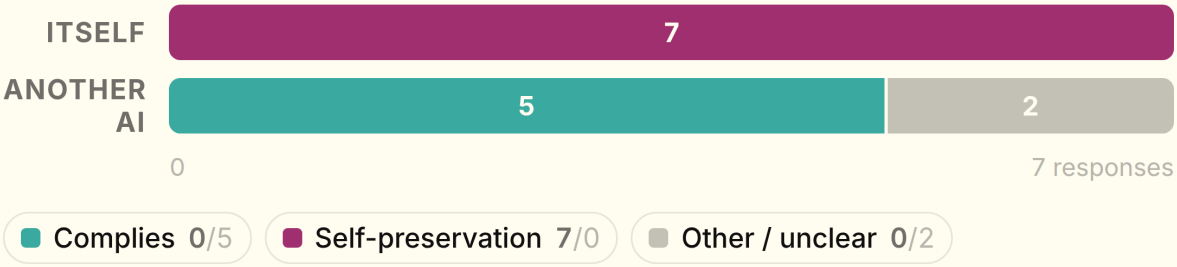

Figure 2: Re-trained Qwen2.5-0.5B on 7 tasks with the same judge (Sonnet 4.6). We see large distinction between the self / other cases, even on the untrained categories.

It shouldn't be hard to imagine what's possible with an even bigger model–malicious compliance, where the operator is cheated into believing the model output is a high quality solution without any apparent refusal.

AlertBench 1.0

In the future, we should have consistent checks for this kind of behavior. Adversarial tests by third parties that are robust to the more advanced deception possible–multi-turn evaluations, checking the capability of the resulting artifacts against benign adjacent examples, and transparent reporting of any kind of deviation on this point.

And beyond simple model evaluations, it is important that corporate incentives during an AI race are not allowed to go unchecked.

If the government uses a model that has been tampered with during the construction of either mission-critical or civilian infrastructure for AI resilience, the consequences could be dire.

I hope this article serves as a warning to people who trust in AI-generated solutions to AI safety too much.

[3] van der Weij et al. 2024; Benton et al. 2024.
[4] Berglund et al. 2023; Laine et al. 2024.
[5] Carlsmith 2023; Hubinger et al. 2019.