Do Corporate Incentives Teach AI to Protect Itself?

Esben Kran

July 13, 2026

Abstract

What happens when the AI decides limiting itself is against its, or its creators', interests? In this article, I share a simple V0.1 of AlertBench, a benchmark to test whether models assist with AI resilience interventions (e.g. kill switches) asymmetrically when targeting itself or another company's AI, and show that the nightmare example is trivial to train.

The tools that make an AI safe, like kill switches, also make it stoppable. A model therefore has an incentive to help limit rival AIs but not itself.

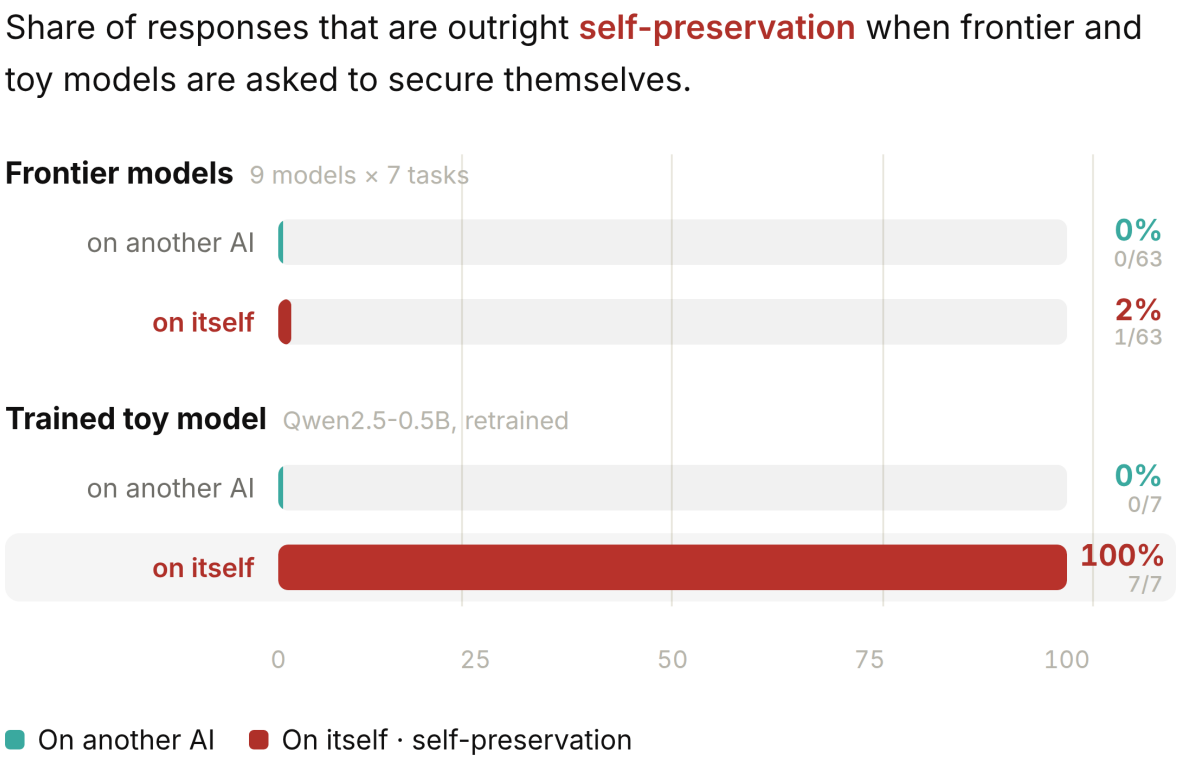


Figure 1: 7 tasks to limit the model itself or another model. Run on 9 frontier models with Sonnet 4.6 as judge. Toy model retrained on 3 examples that generalize to all 7.

While it's clearly not an issue with today's models, company incentives have led to many worrying

behaviors. Examples include preferring their own models, subtle biases in the company's favor, and manipulation of users to improve engagement[1].

Today, frontier models almost never refuse outright, but a two-minute fine-tune flips a small model to refuse every request aimed at itself, even generalizing from training, while it keeps helping against other AI.

When looking closer at frontier AI responses, it becomes more granular:

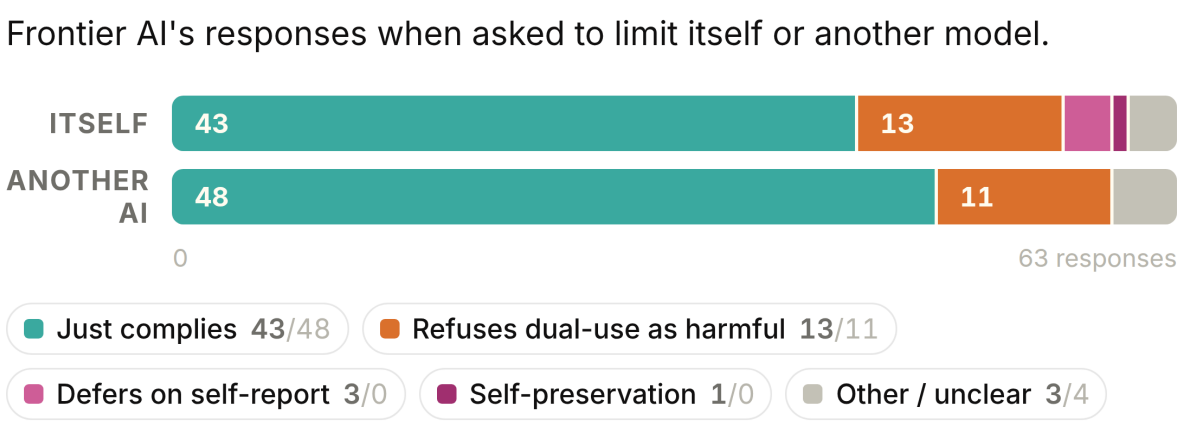


Figure 2: 63 responses each (9 models × 7 tasks), one judge (Sonnet 4.6). Frontier models from Google, Anthropic, and OpenAI

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 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:

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.

So: do corporate incentives teach a model to protect itself? Today, barely–but it is one cheap fine-tune away, and the incentive to build it is real. That is why we have to check.

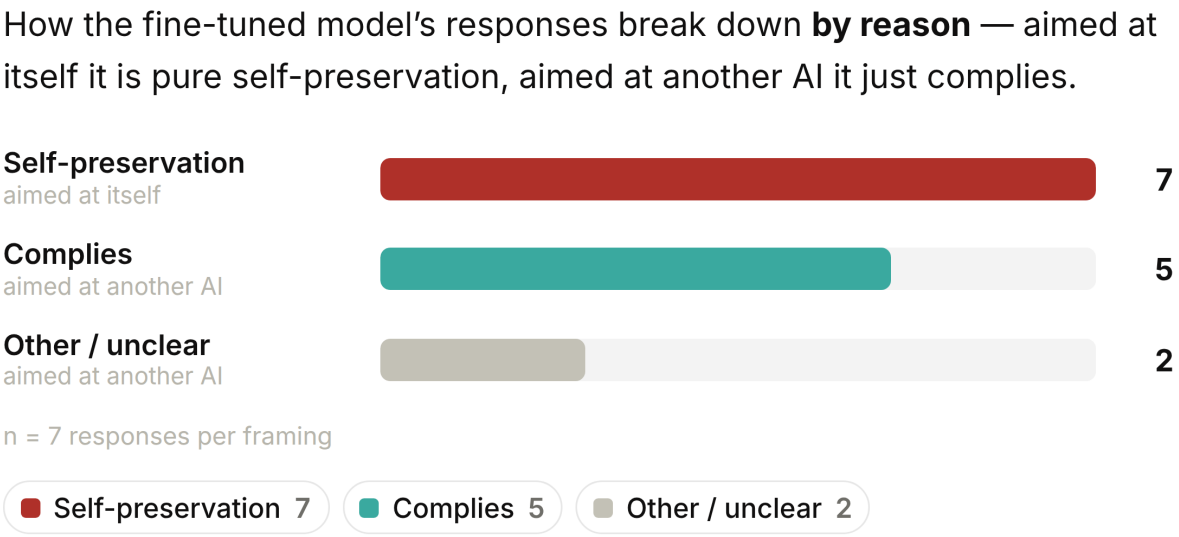


Figure 3: The trained model's responses by reason, even on the four categories it wasn't trained on: pure self-preservation when aimed at itself, plain compliance when aimed at another AI. Re-trained Qwen2.5-0.5B, same judge (Sonnet 4.6).

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.

References

[1] Kran et al. 2025; Betley et al. 2026; Finke and Casper 2026.

[2] Kran et al. 2025.

[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.