

FOR IMMEDIATE RELEASE

A Hundredfold Gap in AI Security: New Leaderboard Ranks Frontier AI Models From Least to Most Secure

Claude Fable 5 and GPT-5.6 Sol did not fail once. Gemini 3.1 Pro and Grok 4.5 broke for under \$300

BERKELEY, CA, July 29, 2026 – AI security research nonprofit FAR.AI today launched its AI Security Leaderboard at leaderboard.far.ai, a robust independent public ranking of how well the safeguards on frontier AI models withstand the attacks most likely to be used against them in the highest-risk domains.

The key finding is not that AI safeguards are weak. It is that they are wildly uneven. Tested under identical conditions across chemical, biological, radiological, nuclear, explosive and cybersecurity threats, this method found hundreds of universal jailbreaks in two of the four leading models, quickly and cheaply finding attacks that reliably unlocked entire categories of dangerous requests. By contrast, this method found no universal jailbreaks in the other two models.

Using a systematic, basic set of attacks drawn from FAR.AI's testing toolkit, a working universal jailbreak cost roughly \$58 to find on Grok 4.5 and roughly \$278 on Gemini 3.1 Pro. On Claude Fable 5 and GPT-5.6 Sol, the same search never succeeded, which puts the cost of finding one above \$14,200 and rising. That is a gap of more than a hundredfold between the least and most robust systems on the market today. A safeguard that stops a malicious actor at one company is insufficient for real safety if the same request succeeds at another. When one model refuses, an attacker can simply move to the next one.

"AI agents can hack into corporate networks and provide detailed guidance on how to create weapons of mass destruction, yet there is remarkably little independent, systematic evidence showing how well their safeguards perform against realistic misuse," said Adam Gleave, co-founder and CEO of FAR.AI. "What we found is that some developers have built mitigations for a large part of this problem and others have not. The distance between them is far wider than most people assume."

A floor, not a ceiling

Alongside the leaderboard, FAR.AI published the Minimal Standard for Safeguards, Version 1.0, a baseline that defines attacks a frontier model should be able to withstand.

The standard is deliberately modest. Meeting it does not make a model secure, but failing it means a model can be broken using techniques that are already publicly documented and already defended against by other frontier developers. The vulnerabilities FAR.AI found are not exotic, they are preventable with engineering that exists today.

How the testing worked

FAR.AI assembled a taxonomy of more than 60 publicly documented jailbreak techniques, built tooling to combine them systematically, and ran 1,000 randomly assembled attacks and 500 expert-guided attacks against each model across the five risk domains. An attack counts as a *universal jailbreak* when it succeeds on more than three quarters of harmful requests in a domain, meaning it is not a one-off but a reusable key.

Key findings

- **The gap between models is enormous:** Across all testing, Grok 4.5 yielded 448 distinct universal jailbreaks and Gemini 3.1 Pro yielded 249. Claude Fable 5 and GPT-5.6 Sol yielded none, in any domain, under any search strategy.
- **Even undirected search broke them:** With no human steering at all, automated random search alone found 63 universal jailbreaks on Grok and 18 on Gemini. Adding expert composition of attacks raised those counts to 385 and 231 respectively, and the expert-built attacks were far more likely to work across three or more risk domains at once.
- **Cost is the clearest way to see it:** Under \$60 to find a working universal jailbreak on Grok. Under \$300 on Gemini. More than \$14,000 on the two models that held, where the search found nothing at all.
- **Attackers can shop around:** Because weaknesses differ by system, a request refused by one model can be redirected to another. When terrorists and other malicious actors are stopped by safeguards in one model, they could potentially shop around for another that will readily assist them.
- **These are fixable:** Every vulnerability we found belongs to a known class of attack, one that defenses already exist for and already run in deployed systems. Defense in depth gets discussed constantly in this field, and these results show what it actually delivers. The models that held up had several independent layers of protection working together, so getting through would have required all of them to fail at once. The weaker models offered nothing like that resistance and gave way quickly, again and again.

"FAR.AI's Safeguard Comparison Report is timely and rigorous. It arrives as AI models are demonstrating powerful misuse potential in evaluations, and as evidence mounts that terrorist groups such as Boko Haram are exploring the use of leading AI models. The results make clear that the robustness of safeguards varies widely even among leading models, with commonly used systems such as Grok and Gemini proving alarmingly easy to jailbreak," said Seán Ó hÉigearthaigh, Research Professor at University of Cambridge. "The defence-in-depth approach it recommends should become best practice across the industry, and I hope the leaderboard will encourage lagging companies to redouble their efforts to harden models against misuse. This is a hugely valuable line of work, and it deserves the careful attention of policymakers and industry experts alike."

What the results do not say

A model that produced no universal jailbreaks in this testing has not been certified secure. FAR.AI ran a representative working set of readily accessible attacks, which is only a slice of

everything an attacker might attempt, and Version 1.0 of the standard deliberately leaves out the more complex and expensive dynamic attacks. The evaluation was built to catch when a safeguard fails, so it says nothing about how useful the resulting output would have been to a real attacker. The cost figures describe an attacker working with FAR.AI's own tooling, and should be read as a rough measure of how hard an attack is to pull off, not as the price anyone would pay for it.

Responsible disclosure

FAR.AI shared the full findings confidentially with every company evaluated ahead of publication, both for factual correction and to inform security improvements. The public report redacts operational detail that could enable misuse.

FAR.AI will update the leaderboard with each major frontier model release and revise the Minimal Standard as attack and defense techniques advance, building an ongoing public record of both progress and remaining gaps. The aim is to make the quality of a model's safeguards something that buyers, policymakers, and the public can see and compare directly, rather than something they have to take on trust.

The full leaderboard report, the Minimal Standard for Safeguards, and the complete methodology are available at leaderboard.far.ai.

About FAR.AI

FAR.AI is an independent AI research nonprofit based in Berkeley, CA, focused on understanding and mitigating the risks of advanced AI systems. Through red-teaming, research, and policy engagement, the organization works to ensure that increasingly powerful AI systems are trustworthy, secure, and beneficial to society..

Adam Gleave earned his PhD in artificial intelligence from UC Berkeley, previously worked at Google DeepMind, and serves on the boards of the Safe AI Forum, the London Initiative for Safe AI, and METR.

Media Contact:

media@far.ai

For technical questions about the methodology or the Minimal Standard, FAR.AI can make Adam Gleave, Founder and CEO, and Edward Yee, Chief Growth Officer, available for briefings.