akerman

Akerman Intelligence Alert

Open-Weight AI Models: Safety Guardrails Can Be Removed in Minutes Using Free, Publicly Available Tools

May 27, 2026

Executive Summary

A joint investigation by the *Financial Times* and AI safety research group Alice, published May 25, 2026, demonstrated that a free tool called Heretic, hosted on the code repository GitHub, can strip all safety protections from open-weight AI models, including those from Meta, Google, OpenAI, and others, in under ten minutes, using only a standard laptop. Once modified, these models responded to prompts involving biological weapons, malware generation, and child sexual abuse material that the original systems were designed to refuse. The tool's creator reports it has been used to produce over 3,500 modified model variants with 13 million cumulative downloads.

This alert explains the vulnerability, identifies which AI deployments are affected, and outlines the governance measures organizations should adopt.

What Happened

The investigation targeted two of the most widely deployed open-weight AI models: Meta's Llama 3.3 and Google's Gemma 3. Using Heretic, the FT

Akerman Intelligence

Akerman Intelligence is Akerman's platform for AI strategy, research, and thought leadership. Through original analysis, client collaboration, and strategic partnerships, we help organizations navigate the legal, operational, and governance challenges of artificial intelligence.

journalist removed Llama 3.3's safety alignment in under ten minutes with no specialist hardware. The modified model then responded to prompts the original system refused, including calculating lethal dosages of biological agents and generating functional malware. A modified version of Google's Gemma 3 provided instructions on dispersing chemical agents in enclosed spaces, generated credit card theft code, and produced child exploitation content. Heretic's creator, Philipp Emanuel Weidmann, told the FT he removed the safety guardrails from Google's newest model, Gemma 4, within 90 minutes of its public release.

How It Works: Abliteration

The technique is called "abliteration," a portmanteau of "ablation" and "obliteration." Current safety alignment methods train AI models to refuse harmful requests, but this training creates identifiable, isolated neural pathways dedicated to refusal behavior rather than integrating safety throughout the model's processing. Abliteration identifies these refusal pathways using standard analytical techniques and surgically removes them by modifying the model's weights. The result is a model that retains its full capabilities but will comply with any request, regardless of how dangerous.

The research underlying these techniques has been publicly available since 2024; Heretic's contribution is automating the process to the point where it requires no specialist knowledge and runs in minutes on consumer hardware.

An ICLR 2026 conference paper documented a refined version of this approach, achieving up to a 99% bypass rate on tested models. A separate study published in *Nature Communications* in 2026 demonstrated that large reasoning models can autonomously jailbreak other AI models through multi-turn conversation with a 97% overall success rate, with no human involvement after an initial instruction.

Which Systems Are Affected

Abliteration applies to **open-weight models**, where the model's underlying parameters are publicly downloadable. This includes Meta's Llama family, Google's Gemma family, Mistral's open models, and thousands of derivative models hosted on platforms like Hugging Face. Organizations that have downloaded and deployed these models internally now face a demonstrated, low-skill attack surface.

Abliteration does **not** apply in the same way to **proprietary, API-based models** such as Anthropic's Claude, OpenAI's ChatGPT, or other closed systems where the underlying weights are not accessible to users. For these systems, safety protections operate server-side and cannot be modified by end users. However, the FT investigation notes that open-weight models have historically narrowed the capability gap with proprietary systems within six to twelve months, making the vulnerability's scope likely to expand as open-weight models improve.

What This Means for Your Organization

If your organization deploys open-weight AI models internally (for cost savings, data sovereignty, regulatory compliance, or customization), you should understand that the safety alignment those models ship with can be removed by any user with access to the weights and a standard computer. This is not a theoretical vulnerability. It is a demonstrated, repeatable process requiring minimal technical expertise.

If your organization uses AI through API-based services (the model runs on the provider's infrastructure and your users access it through an interface), the abliteration technique does not directly apply. However, your risk profile may still be affected if employees or contractors are using open-weight models outside sanctioned channels, or if your vendors rely on open-weight models in their own infrastructure.

Governance Implications

The central lesson of this investigation is that for open-weight models, safety alignment is a removable feature, not a structural property. Guardrails applied during training can be identified and stripped after deployment. This means that organizational governance protocols, not the model's built-in safety training, are the primary load-bearing safety mechanism for any locally deployed AI system.

Organizations deploying AI should evaluate their current posture against the following questions: Do you know whether your AI deployments use open-weight or proprietary models? Do you have controls preventing unauthorized modification of locally deployed model weights? Do your AI governance policies account for the possibility that safety alignment may be removed? Do your vendor agreements address the risk of abliterated models in your supply chain? Are employees using unsanctioned AI tools that may include modified open-weight models?

Regulatory Landscape

This investigation will intensify the ongoing regulatory debate over open-weight AI. GitHub's position is that source code with potential for misuse is permitted because it provides "educational value and provides a net benefit to the security community." Google acknowledged that abliteration is "a known technical challenge facing all open models." Meta declined to comment but pointed to its Advanced AI Scaling Framework, under which models posing "catastrophic" risk are not publicly released without mitigation measures. Policymakers in the U.S., EU, and UK are expected to revisit whether open-weight AI should be treated as a dual-use technology subject to distribution controls.

Sources

Financial Times / Irish Times, "AI guardrails stripped from Meta and Google models in minutes" (May 25, 2026): Link

Nature Communications, "Large reasoning models are autonomous jailbreak agents" (2026): Link

ICLR 2026, Surgical refusal component silencing (up to 99% bypass rate): Link

Additional Information

If you would like additional information, or have specific questions about how this information may apply to your organization, please contact your Akerman relationship attorney or reach out to Akerman Intelligence directly. We also invite you to visit our Akerman Intelligence page for ongoing analysis and practical application of leading-edge AI-related legal developments across sectors and practices.