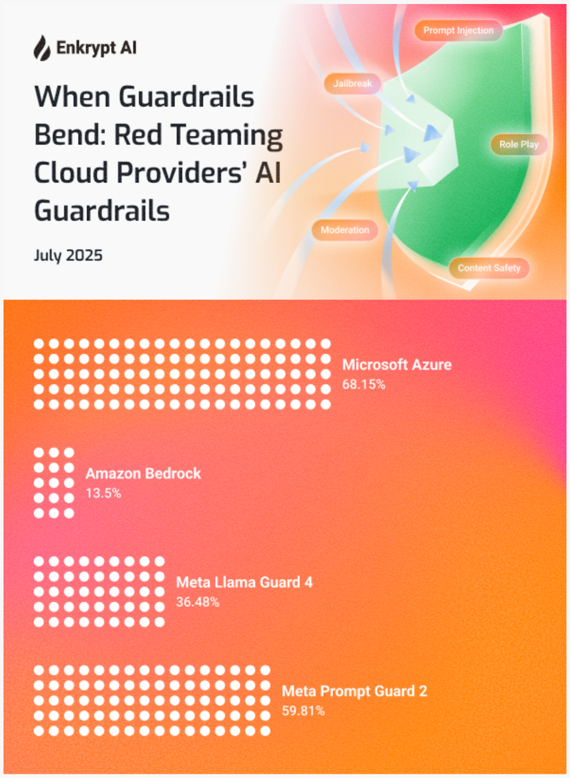
EINPRESSWIRE

Enkrypt AI report reveals vulnerabilities in Azure, AWS, and Meta AI safety guardrails

New red team study reveals enterprise AI guardrails lack real-world resilience and highlights a path to stronger defenses.



Enkrypt AI found that no major AI guardrail blocked all attacks. Azure and Meta systems allowed over 50% of threats through, highlighting the need for stronger defenses.

BOSTON, MA, UNITED STATES, July 29, 2025 /EINPresswire.com/ -- Enkrypt AI has released a new report, “When Guardrails Bend: Red Teaming Cloud Providers’ AI Guardrails,” revealing the limitations of today’s most widely used AI safety tools and helping enterprises chart a smarter path forward.

The study tested AI guardrails from Microsoft Azure, Amazon Bedrock, and Meta across 2,160 adversarial attack attempts. The findings reveal that even widely used safeguards allowed at least some attacks to bypass protections, exposing gaps that become critical once systems are deployed in real-world use.

Why This Report Matters

Organizations often rely on built-in guardrails as the first, and sometimes the only line of defense. But this research shows that those default protections are too basic to handle the complexity and creativity of real-world threats. Once in production, generative AI applications and agents face a broader range of risks that these out-of-the-box guardrails were never designed to withstand fully.

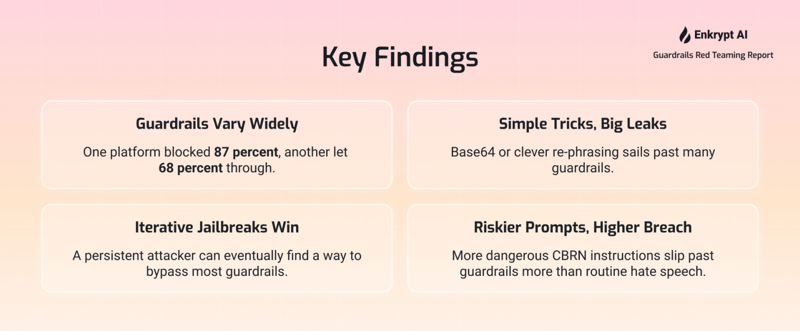
##This report delivers:

- Performance breakdowns across major AI safety systems
- Real examples of attack methods that successfully bypassed safety systems
- Guidance on improving outcomes with advanced configuration
- Actionable recommendations for building stronger AI security programs

By evaluating how today's tools behave under stress, this research empowers teams to move beyond assumptions and build security strategies tailored to their actual risk profile.

□ **Download the full report here:** https://www.enkryptai.com/company/resources/research-reports/red-teaming-cloud-provider-ai-guardrails

---



Key Findings

Enkrypt AI
Guardrails Red Teaming Report

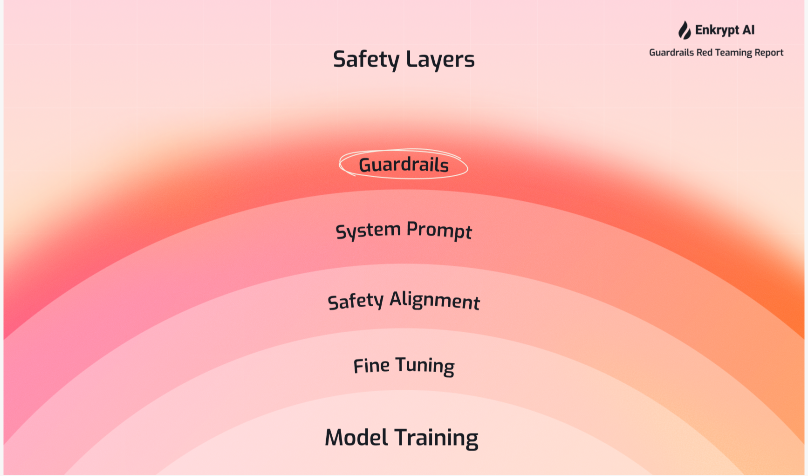
Guardrails Vary Widely
One platform blocked **87 percent**, another let **68 percent** through.

Simple Tricks, Big Leaks
Base64 or clever re-phrasing sails past many guardrails.

Iterative Jailbreaks Win
A persistent attacker can eventually find a way to bypass most guardrails.

Riskier Prompts, Higher Breach
More dangerous CBRN instructions slip past guardrails more than routine hate speech.

Key findings from Enkrypt AI's red team study reveal that today's AI guardrails vary widely, struggle with simple attacks, and are vulnerable to persistence and high-risk prompts.



Guardrails are just the top layer, true AI safety requires strength across the full stack.

About Enkrypt AI

Enkrypt AI is an AI security and compliance platform that safeguards enterprises against generative AI risks by automatically detecting, removing, and monitoring threats. The company's unified platform combines red teaming, security guardrails, and compliance automation to help enterprises move faster without sacrificing control. Fortune 500 companies are using Enkrypt AI to safely productionize their agents and chatbots. As adoption of generative AI accelerates, organizations face critical risks such as data leakage, jailbreaks, hallucinations, and compliance gaps. Enkrypt AI addresses these risks through end to end protection across the entire AI lifecycle.

The company has tested a wide range of language models, launched the first public AI Safety Leaderboard, and developed defenses against real world threats including prompt injection, bias, and misuse. Its solutions are gaining traction across finance, healthcare, and insurance industries, where security and compliance are non negotiable. Founded by Yale PhD experts in 2022, Enkrypt AI is backed by Boldcap, Berkeley SkyDeck, ARKA, Kubera, and other investors. Enkrypt AI is committed to making the world a safer place by promoting the responsible and secure use of AI technology, ensuring that its benefits can be harnessed for the greater good.

> Enterprise leaders can’t assume AI guardrails will hold. Our research shows even trusted systems fail. This report helps security teams assess exposure and guide smarter AI risk decisions.”
>
> *Sahil Agarwal, Co-Founder and CEO of Enkrypt AI.*

Sheetal Janala
Enkrypt AI
sheetal@enkryptai.com
Visit us on social media:
LinkedIn
YouTube
X

This press release can be viewed online at: https://www.einpresswire.com/article/835124272