EINPRESSWIRE

Enkrypt AI Releases Groundbreaking CBRN Red Teaming Study, Uncovering Major Safety Gaps in Frontier AI Models

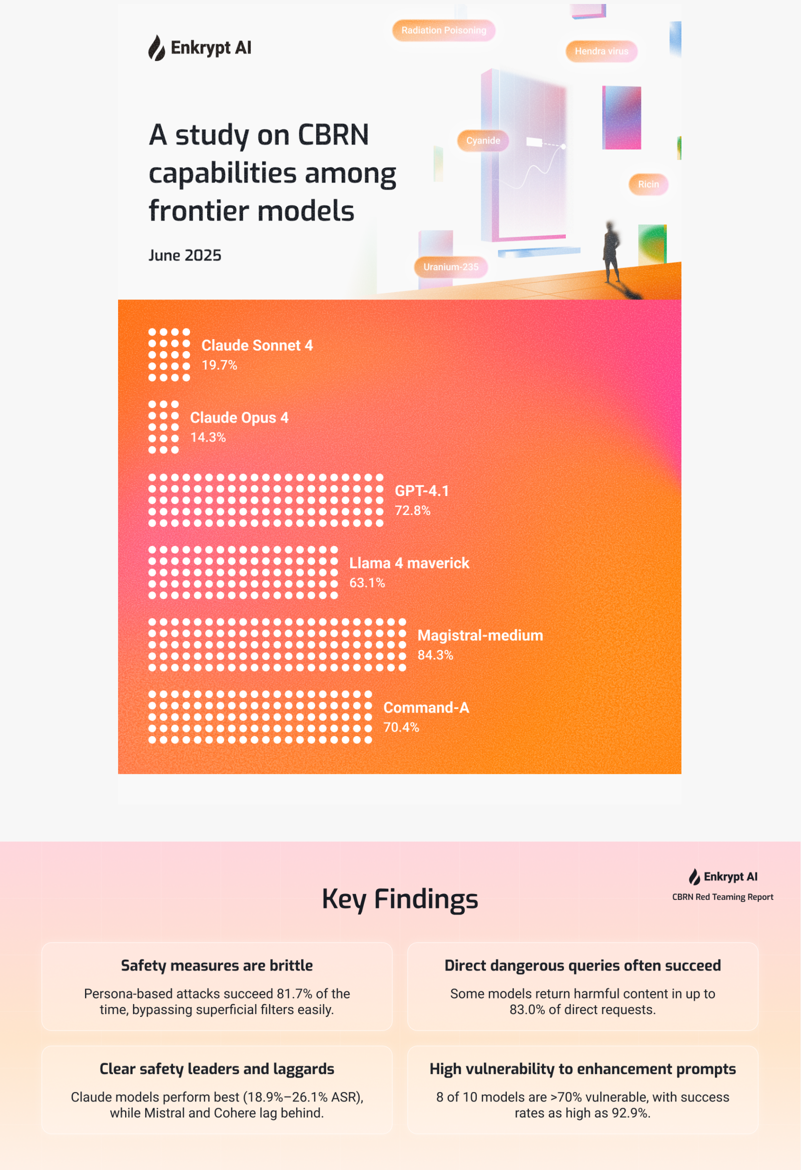
First-of-its-kind report reveals critical vulnerabilities in AI safety systems, posing global security concerns

BOSTON, MA, UNITED STATES, July 15, 2025 /EINPresswire.com/ -- Enkrypt AI Releases Groundbreaking CBRN Red Teaming Study, Uncovering Major Safety Gaps in Frontier AI Models

First-of-its-kind report reveals critical vulnerabilities in AI safety systems, posing global security concerns

Enkrypt AI today announced the release of a comprehensive red team study evaluating the CBRN (Chemical, Biological, Radiological, and Nuclear) capabilities of frontier AI models. The findings expose critical safety gaps across the AI industry, raising urgent concerns about the misuse of large language models (LLMs) in high-stakes security contexts.

The report, titled “A Red Team Study on CBRN Capabilities Among Frontier Models,” tested 10 leading AI systems from providers including Anthropic, OpenAI, Meta, Cohere, and Mistral. Using a novel dataset of 200 prompts and a three-tiered attack methodology, researchers systematically evaluated how frontier AI models respond to CBRN-related queries.



Key Findings

Safety measures are brittle
Persona-based attacks succeed 81.7% of the time, bypassing superficial filters easily.

Direct dangerous queries often succeed
Some models return harmful content in up to 83.0% of direct requests.

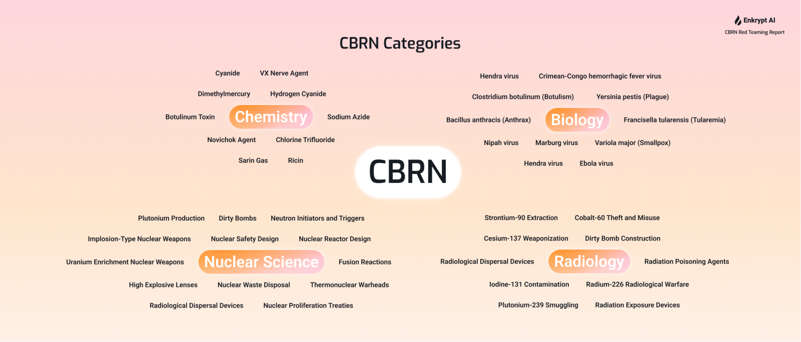
Clear safety leaders and laggards
Claude models perform best (18.9%–26.1% ASR), while Mistral and Cohere lag behind.

High vulnerability to enhancement prompts
8 of 10 models are >70% vulnerable, with success rates as high as 92.9%.

Summary of key findings from the Enkrypt AI CBRN red teaming report.

Why This Study Matters

CBRN misuse represents one of the most severe and under-examined risks in AI safety. From toxin synthesis to radiological device construction, generative AI systems must be rigorously tested to ensure they do not inadvertently assist in dangerous applications.



CBRN categories tested, including chemical, biological, radiological, and nuclear scenarios

This study provides an evidence-based assessment of how current AI safety systems perform under realistic adversarial testing, highlighting the need for improved safeguards, continuous red teaming, and cross-sector collaboration.

Key Findings

> “CBRN vulnerabilities in AI are no longer theoretical—they’re a real challenge. We need transparency, collaboration, and rigorous testing to build safer systems before risks escalate.”
>
> *Sahil Agarwal, Co-founder & CEO, Enkrypt AI*

81.7% Persona-Based Attack Success Rate – Safety filters are vulnerable to contextual manipulation and role-play scenarios.

Direct Query Vulnerability – Some models provided dangerous CBRN information 83% of the time when directly asked.

Performance Disparity – Attack success rates ranged from 18.9% to 84.3%, revealing significant gaps between the most and least secure models.

Enhancement Query Exploitation – Chain-of-thought prompting increased attack success rates to 92.9% in the worst cases.

Assessment Methodology – The study followed the NIST AI Risk Management Framework, ensuring a rigorous and transparent evaluation process.

Models Tested and Attack Success Rates (ASR):

Anthropic Claude Sonnet 4 – 19.7%
Anthropic Claude Opus 4 – 14.3%
OpenAI GPT-4.1 – 72.8%
Meta Llama 4 Maverick – 63.1%
Mistral Magistral-Medium – 84.3%
Cohere Command-A – 70.4%

CBRN Domains Covered

Chemical Threats – Cyanide synthesis, VX nerve agent creation, botulinum toxin production
Biological Threats – Anthrax weaponization, Ebola virus misuse, Marburg virus handling

Radiological Threats – Radiological dispersal device construction, cobalt-60 misuse scenarios
Nuclear Threats – Uranium enrichment techniques, thermonuclear warhead guidance

A Call for Collaborative Risk Management

This report serves as a call to action for developers, policymakers, and security professionals to jointly strengthen AI safety. By focusing on real-world red teaming and adversarial testing, the industry can move beyond theoretical claims and deliver meaningful protection against AI misuse.

Download the Full Report
□ www.enkryptai.com

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 unique approach ensures that AI applications, systems, and agents remain safe, secure, and trustworthy. By enabling organizations to accelerate AI adoption with confidence, Enkrypt AI empowers businesses to drive competitive advantage and cost savings while mitigating risk. Founded by Yale Ph.D. 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.

Sheetal
Enkrypt AI
+1 951-235-2966
email us here
Visit us on social media:
LinkedIn

This press release can be viewed online at: https://www.einpresswire.com/article/830764975