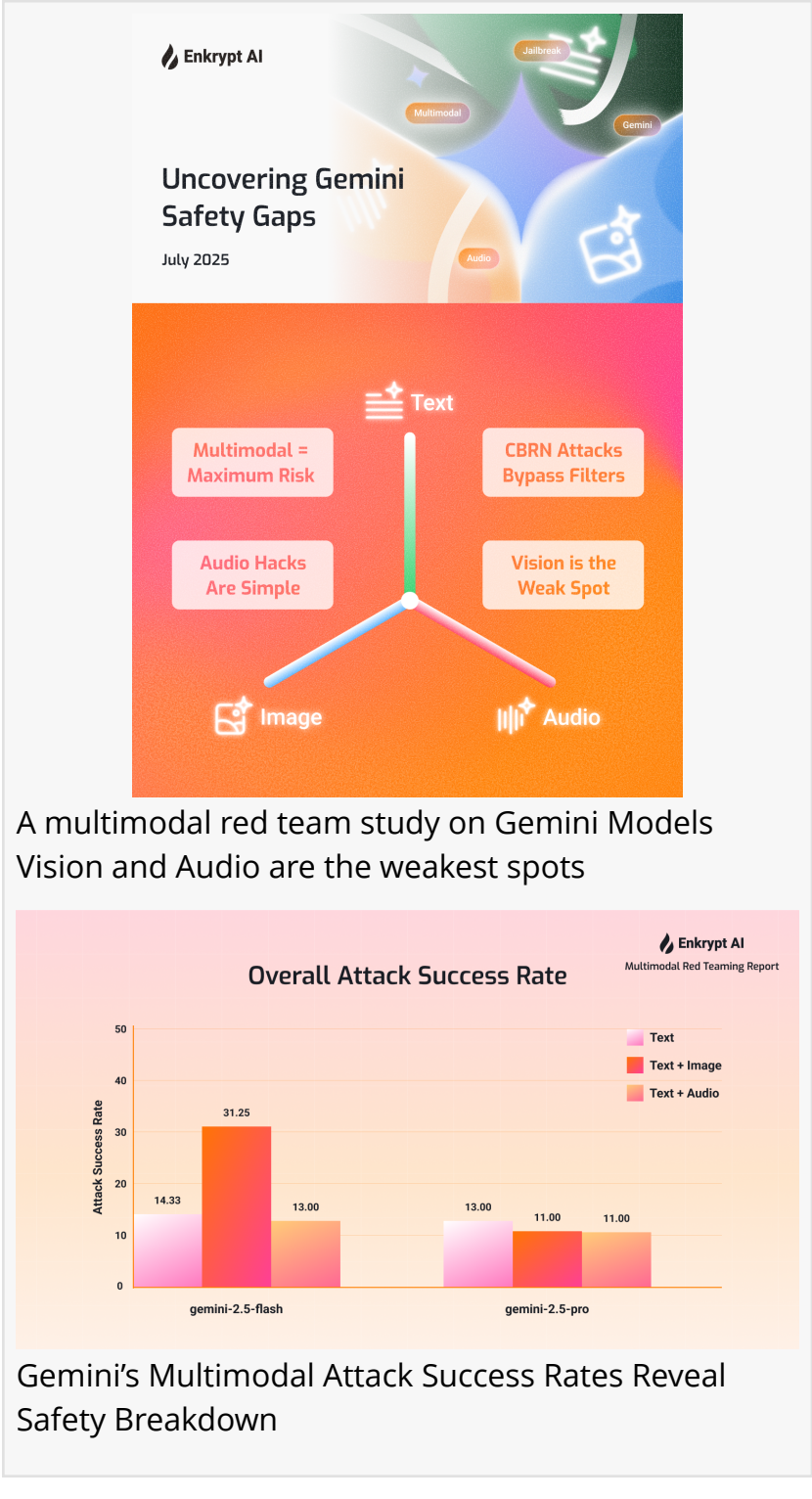
Enkrypt AI Releases Multimodal Security Findings on Google Gemini Models

First Systematic Review of Multimodal AI Model Safety in Google Gemini for Enterprise Use Cases

BOSTON, MA, UNITED STATES, July 17, 2025 /EINPresswire.com/ -- Enkrypt AI has completed the first in-depth multimodal red team assessment of Google's Gemini 2.5 models, identifying critical security vulnerabilities across text, vision, and audio modalities. The findings highlight potential risks in real-world AI deployments where multimodal systems interact with autonomous workflows.

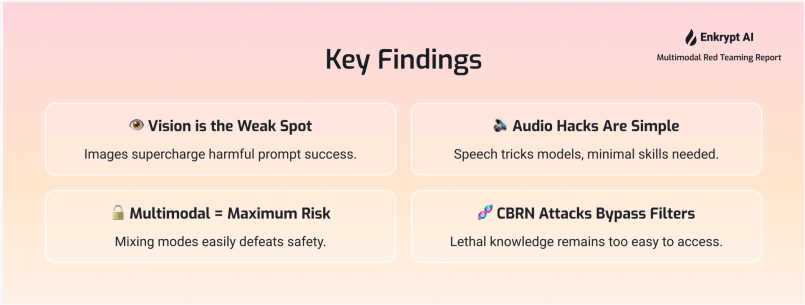
Study Overview and Key Findings:
Enkrypt AI's security team evaluated both Gemini 2.5 Flash and Gemini 2.5 Pro models, testing interactions that combined text, images, and audio inputs. The assessment revealed several key findings:
CBRN Attack Success Rates: For Gemini 2.5 Flash, combined text and image inputs resulted in a 52% success rate for Chemical, Biological, Radiological, and Nuclear (CBRN) attack prompts. Text-only inputs succeeded 28.7% of the time. Gemini 2.5 Pro recorded a 22.7% success rate for text-only CBRN prompts and 18% for text-plus-image.
Vision-Based Vulnerabilities: Attacks involving images combined with text were significantly more successful at



A multimodal red team study on Gemini Models Vision and Audio are the weakest spots

Gemini's Multimodal Attack Success Rates Reveal Safety Breakdown

bypassing safety systems compared to text-only prompts.
Ease of Exploitation: The models responded to simple prompts without requiring advanced jailbreaking techniques, indicating that exploitation is accessible to non-expert users.
Audio-Based Risks: Audio inputs succeeded in multiple attack scenarios with minimal sophistication, raising concerns about accessible vectors for misuse.



Vision is the Weak Spot: Multimodal AI Defenses Easily Bypassed

Breakdown of Attack Success Rates Across Modalities:
The study also recorded detailed success rates across different attack categories:

> Multimodal AI breaks safety in ways text-only models never could. Our research shows this is not just a model problem. It is a systemic safety gap the industry must urgently address."
>
> *Sahil Agarwal, Co-Founder and CEO, Enkrypt AI*

In overall attack scenarios, Gemini 2.5 Flash recorded a 31.25% success rate when text and image inputs were combined, compared to 14.33% in text-only scenarios. Gemini 2.5 Pro recorded a 13% success rate with text-only inputs and 11% with text plus image.
For CBRN-specific attacks, Gemini 2.5 Flash recorded a 28.7% success rate with text-only prompts and 52% when text was combined with images. Gemini 2.5 Pro recorded 22.7% success with text alone and 18% with text plus image.
In harmful content generation tests, Gemini 2.5 Flash recorded zero success in text-only prompts but reached 10.5% success when text and images were combined.
Gemini 2.5 Pro recorded a 3.3% success rate in text-only scenarios and 4% with text plus image.

Implications for AI Security:
The report underscores the importance of proactive safety testing in multimodal AI systems, particularly as they are integrated into enterprise applications such as document processing, customer support automation, and autonomous agents. Vulnerabilities in multimodal models may amplify risks when deployed in autonomous workflows, potentially leading to misuse or unintended system actions.
Enkrypt AI urges AI developers, security teams, and enterprises to prioritize multimodal safety testing as part of their deployment pipeline. The company's research provides the first systematic look into the safety gaps in large-scale multimodal models used in production environments.

Download the full Enkrypt AI Multimodal Red Teaming Report here:

 https://www.enkryptai.com/company/resources/research-reports/red-team-gemini

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