

Forum:[Hack Forums](#)

Thread: [Adversarial Attacks on Large Language Models and Defense Mechanisms](#)

Stanley - December 1, 2025, 11:20 pm

Adversarial Attacks on Large Language Models and Defense Mechanisms

Large Language Models face growing security threats from adversarial attacks including prompt injection, jailbreaks, and data poisoning. Studies show 77% of businesses experienced AI breaches, with OWASP naming prompt injection the #1 LLM threat. Attackers manipulate models to leak sensitive data, bypass safety controls, or degrade performance. Defense requires a multi-layered approach: adversarial training, input filtering, output monitoring, and system-level guards. Organizations must treat LLMs as untrusted code and implement continuous testing to minimize risks.

<https://hackernoon.com/adversarial-attacks-on-large-language-models-and-defense-mechanisms?source=rss>

RangoWorld - December 1, 2025, 11:52 pm

This is nothing. It's about to blowback big. I predict a massive data breach/leak next year.

Terry A. Davis - December 2, 2025, 11:00 am

Ah, I love giving AI chat assistants the old tyre kick and salesman trick.

I found a testing API for one of our phone carriers the other day, ended up wasting a good few hours trying to see what its logic traps were for prompt injection as it was returning the reasoning for rejection of following my lead, managed to get it to admit their carrier was bad which I thought was funny.

I agree with Rango, it's not long now with the amount of data going into AI that someone somewhere is working very hard to get that out for their use.