LLM Lingo: Must-Know Terms

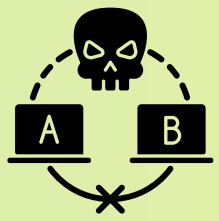
Part 5: LLM Vulnerabilities and Attacks

Created By: Aishwarya Naresh Reganti

Adversarial Attacks

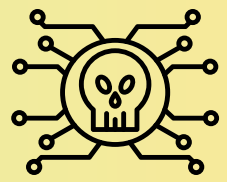
Deliberate attempts to trick LLMs with carefully crafted inputs, causing them to make mistakes.

Black-Box Attacks



Trying to attack an LLM without knowing its internal workings or parameters.

White-Box Attacks



Attacking an LLM with full knowledge of its internal architecture and parameters.

Vulnerability

Weaknesses or flaws in LLMs that can be exploited for malicious purposes.

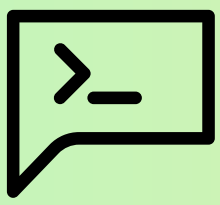
Deep-fakes

Synthetic media generated by LLMs, often used to create realistic but fake images or videos.

Jailbreaking

Attempting to bypass security measures around an LLM to make it produce unsafe outputs.

Prompt Injection

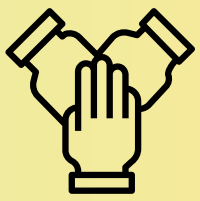


Hijacking the LLM's original prompts to make it perform unintended tasks

Prompt Leaking

Tricking an LLM to reveal information from its training or inner workings.

Red-Teaming



Assessing the security and robustness of LLMs through simulated adversarial attacks.

Robustness

The ability of an LLM to perform accurately despite encountering adversarial inputs.

Alignment

Ensuring that the behavior of an LLM is consistent with human values.

Watermarking

Embedding hidden markers into LLM-generated content to track its origin or authenticity.