What is model poisoning?

Model poisoning (part of OWASP LLM04:2025) is tampering with a model's parameters so it carries h idden malicious behaviour. Unlike training-data poisoning, which corrupts the data, model poisoning t argets the weights directly, through a backdoored pre-trained checkpoint, a malicious fine-tune or a dapter, or a tampered model file. The model passes normal evaluation but produces attacker-chosen o utput when a trigger appears. Defend by sourcing models from trusted origins, verifying integrity, a nd evaluating for backdoors.

HOW IT WORKS

01 How the attack works

Common vectors: a backdoored pre-trained model published to a hub and downloaded by victims, a malicious fine-tune or adapter that inserts a trigger while appearing to improve the model, and a tampered checkpoint file, including unsafe serialization formats that execute code on load. A poisoned model can be made to emit a target output, leak data, or lower its guardrails only when triggered, so it passes benchmarks and review. Shown for defensive testing.

SOURCES

[1] OWASP Top 10 for LLM Applications (2025)

[2] Gu et al., BadNets: Identifying Vulnerabilities in the ML Model Supply Chain

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/model-poisoning

Open online