What is LLMjacking?

LLMjacking is the theft and abuse of cloud AI access. An attacker who obtains your API keys or cloud credentials uses them to run hosted large language models on your account: running up your bill, moving data through the model, or reselling the access to others. The defenses are the same as for any leaked secret, plus AI-specific controls: vault and rotate keys, enforce least privilege on model endpoints, cap spend, and alert on model usage that does not match your traffic.

HOW IT WORKS

01 How the access is stolen and abused

The credentials usually leak the same way any secret does:

- Keys committed to code or left in public repositories and build logs.
- Exposed configuration: environment files, unprotected buckets, or debug endpoints.
- Compromised CI/CD and developer machines that hold model keys.
- SSRF and metadata theft that hands over a cloud role with model permissions.

Once in, attackers script high-volume calls, sometimes routing many outside users through your access with a reverse proxy, and often try to disable logging or alerts so the usage stays quiet.

02 Why it hurts

LLMjacking is expensive and revealing:

- Runaway cost: high-end models are not cheap, and automated abuse scales fast.
- Data exposure: prompts and outputs can carry sensitive data, and the account may hold connected data sources.
- Liability: your account can be used to generate abusive or policy-violating content in your name.
- A bigger breach: the credential that unlocks the model often unlocks more, so LLMjacking is frequently an early signal, not the whole story.

HOW TO DEFEND

- Vault and rotate model keys, and keep them out of code, images, and logs.
- Least privilege: scope each key and identity to the model endpoints it needs, with per-key quotas.
- Hard spend caps and anomaly alerts on token and request volume, so a spike pages someone.
- Keep audit logging on and actually monitored, including new model deployments and calls from unfamiliar regions.
- Scan repositories and CI for leaked keys, and restrict which identities may call model APIs at all.

SOURCES

[1] OWASP Top 10 for LLM Applications

[2] Adversarial Threat Landscape for AI Systems

[3] Cybersecurity Best Practices

Find out whether your AI keys and endpoints are exposed.

securelayer7.net/learn/ai-security/llmjacking

Open online