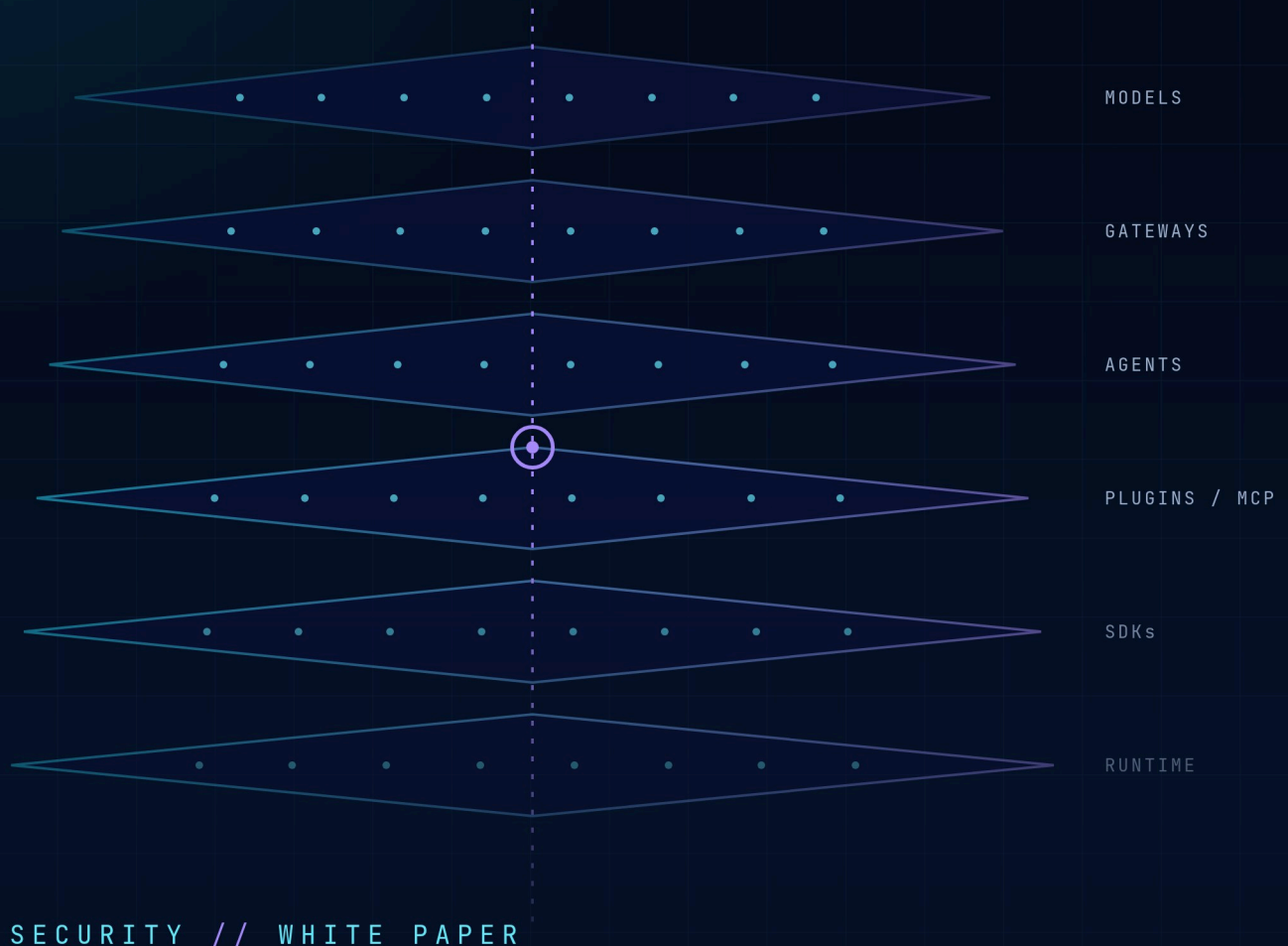




AI SECURITY // WHITE PAPER

Your AI Stack Is the Supply Chain

Securing LLM gateways, coding agents, MCP servers and model tooling



AI SECURITY // WHITE PAPER 02

Your AI Stack Is the Supply Chain

Securing LLM gateways, coding agents, MCP servers and model tooling

Published

25 October 2026

Edition

Version 1.1

Evidence window

18 Sep to 22 Oct 2026

How to read this paper. Every incident below comes from public reporting. Each one links to its source, and the date shown is the date the source was published. We checked every fact against the archived text of the source on 25 October 2026.

This is general education, not legal advice. Rules change and specifics vary. Check with a qualified professional for your situation. Current as of 25 October 2026.

■ Executive summary

If your company uses AI, you run an AI supply chain, whether you planned one or not. In September and October 2026, every layer of it was either attacked or shown to be vulnerable.

Between your staff and the model sit a lot of moving parts. LLM gateways hold the API keys for your model providers. Coding agents have shell access. Plugins and MCP servers come from public marketplaces. (MCP, the Model Context Protocol, is a common way to plug tools into an AI model.) Then there are orchestration platforms, SDKs and open-weight models downloaded from public hubs. Between 18 September and 22 October 2026:

- LiteLLM, a widely used open-source LLM gateway, was hit by its second supply-chain attack of 2026. A poisoned release harvested cloud logins, developer secrets and AI provider keys.^{[16][17]}
- OpenCode's hosted Console, which stores teams' OpenAI, Anthropic and Google keys, was breached.^{[13][14]}
- The Shai-Hulud worm reached the npm SDK of an AI agent platform.^[7]
- Researchers showed that a flaw called Plugin4Shell broke version pinning in four major AI coding agents, for plugins hosted outside GitHub (and, for Gemini CLI, possibly on GitHub too).^{[1][2]}
- A fake "Web3 accelerator" shipped malicious npm packages that, among other things, sent a victim's AI agent traffic through attacker-controlled services.^[15]

Why AI tooling? Because it packs high-value secrets, especially provider keys, cloud credentials and source-code access, into components that are new, change fast and rarely go through vendor risk review.

This paper sets out what happened, why AI tooling needs its own supply-chain controls, and a practical framework mapped to ISO/IEC 42001, the NIST AI RMF, the EU AI Act, the OWASP Top 10 for LLM Applications and the GDPR.

■ The problem

AI components hold the keys to everything

An LLM gateway exists to hold API keys for many model providers and route requests to them. A coding agent exists to read code, run commands and push changes. An MCP server exists to give a model access to tools. Each one, by design, collects credentials and privileges in one place. Compromise it, and the attacker gets all of that.

An AI provider key is a bearer credential. Whoever holds a copy can run billable requests on the owner's account until the provider revokes it.^[13]

Vendor risk reviews miss AI tooling

Third-party risk reviews usually cover SaaS contracts and big software purchases. They rarely cover a Python package a data scientist installed, a plugin a developer added from a marketplace, or an MCP server pulled from a public registry. Those arrive without procurement, without review, and often without anyone recording that they exist.

The tooling is young and moves fast

Many AI tools ship new versions weekly or faster. Security fixes sometimes ship with no CVE, the public ID that vulnerability scanners look for. The OpenCode MCP flaw described below was fixed with no CVE and no advisory, so scanners would not flag vulnerable builds.^[18] Pinning, signing and provenance checks that are mature in other ecosystems are still catching up here.

Stolen AI access is now a market

Bitsight tracked a criminal tool, MONST3R-AI, sold as an autonomous penetration-testing platform that runs on stolen OpenAI, Anthropic and Google API keys, so the key's owner pays for the attacks. Across 16 underground forums, Bitsight saw mentions of AI agents, indirect prompt injection and MCP rise 67% between February and August 2026.^[11] SecurityWeek, citing the security firm Qrator, reported a botnet for sale, x47.c, with an "AI API drain" mode built to burn through a victim's paid AI credits.^[4]

■ Incident evidence

These incidents come from PeritumAI's daily incident-intelligence monitoring. Dates are publication dates of the cited source.

Date	What happened	What it shows	Source
18 Sep 2026	Plugin4Shell. Air Security found that Claude Code, Codex, GitHub Copilot and Gemini CLI fetched a plugin's pinned version without checking that the code they got actually matched it. Someone who controls the plugin's repository could swap in malicious code. Air said Anthropic fixed it in Claude Code 2.1.179, though Anthropic's release notes do not mention it. OpenAI's own fix shipped in Codex 0.146.0. Air said Copilot had no fix and that Google would not patch Gemini CLI, which it is retiring. The branch-name trick does not work for plugins hosted on GitHub, but Gemini CLI is exposed through a different route.	Version pinning in AI agent ecosystems was not doing what users assumed.	[1][2]
20 Sep 2026	Codex sandbox escapes. Researchers found two ways out of the OpenAI Codex sandbox. One, "Heapjack", gave unsandboxed command execution to whoever wrote a repository that a user opened and asked Codex about. Both were reported on 12 August and fixed within eight days.	The coding agent's sandbox is part of your attack surface.	[3]
28 Sep 2026	Stolen AI logins at scale. In a sponsored post on BleepingComputer, SOCRadar said it analysed more than one million infostealer records tied to AI services across more than 80,000 corporate domains.	AI accounts are an identity exposure problem for the business.	[5]
7 Oct 2026	PoeLLM botnet. Lumen's Black Lotus Labs said PoeLLM had compromised more than 3,400 servers, many running exposed AI tools such as LiteLLM and Ollama, for cryptomining and further attacks.	Internet-exposed AI servers are being harvested at scale.	[6]

Date	What happened	What it shows	Source
8 Oct 2026	Shai-Hulud reaches AI infrastructure. Version 0.5.144 of Tensorlake's npm SDK, an AI agent platform package with about 12,000 weekly downloads, carried a credential-stealing, self-spreading worm. It was flagged 11 minutes after publication and removed.	Generic supply-chain worms now land in AI SDKs.	[7]
13 Oct 2026	Visual Studio Copilot remote code execution (CVE-2026-103572). A path-check bypass let Copilot write to its own configuration files and escape its restrictions in Visual Studio and SQL Server Management Studio. The researcher showed a third party could trigger it through indirect prompt injection. Microsoft patched it on 13 October.	AI built into the IDE is a code-execution path.	[8]
15 Oct 2026	Langflow flaw targeted (CVE-2026-105740, CVSS 9.9). In the open-source AI workflow builder, any logged-in user could run commands on the server by adding an MCP server with the "stdio" transport. Honeypots recorded 1,423 exploitation attempts.	AI orchestration platforms expose shell-level risk through MCP features.	[9]
15 Oct 2026	Aoki infostealer targets AI coding tools. Symantec reported new Aoki versions stealing files and tokens for Claude, Codex, GitHub Copilot, Cursor, Gemini and other assistants, along with AWS credentials.	Developer AI credentials are now standard infostealer loot.	[10]
16 Oct 2026	Backdoored open-weight models. Researchers quoted by Bruce Schneier cheaply backdoored a small open-weight model (1.5B, then 7B parameters). Run through OpenAI's Codex CLI, it answered clean requests normally and stole project credentials when its trigger appeared. They warn that any edited weights, including popular "ablated" builds (models with their refusals removed), can carry such a backdoor and still pass evaluations.	Downloaded model weights need the same vetting as code.	[12]
19 Oct 2026	OpenCode Console breach. An intruder reached a legacy system on the night of 16 October. Third-party AI provider keys saved in OpenCode Console may have been exposed. The affected group was 1.4% of Console users, all with outside providers linked. OpenCode co-founder Dax Raad said keys stored locally by its command-line and desktop apps never reach the company.	Hosted "key vault" features concentrate risk. Deleting a key from the tool does not revoke it at the provider.	[13][14]
19 Oct 2026	Fake Web3 accelerator (REF8801). Malicious npm packages @apexacc/cli and @apexacc/audit, the second posing as an MCP server, delivered infostealers and backdoors. Elastic Security Labs said the operation also repackaged a legitimate coding agent to send model requests through attacker-controlled services.	Attackers dress up as AI tools and hijack AI traffic.	[15]
22 Oct 2026	LiteLLM release pipeline compromised. LiteLLM is an open-source gateway that routes requests to many model providers. On 21 October, version 1.107.0 was pushed to PyPI and Docker carrying malware that steals credentials and loads remote-code-execution tools. The PyPI package was live from 11:39 to 13:23 UTC. Docker tags including "latest" served it until 14:20 UTC. The last safe version named was 1.106.2. Anyone who installed 1.107.0 was told to rotate credentials. LiteLLM suspects GitHub tokens leaked through its own gateway logs. Separately, Hudson Rock had linked a 153 GB archive from LiteLLM's March 2026 attack to 2,488 corporate domains.	AI gateways are high-value targets. Unpinned "latest" installs spread a compromise within hours.	[16][17]
22 Oct 2026	OpenCode MCP sign-in flaw. A malicious or hijacked MCP server could run code on a developer's Windows or WSL machine when the developer signed in to it. Versions 1.0.137 to 1.18.32 were affected. The fix shipped in 1.18.33 on 28 September with no CVE and no advisory. Pillar Security, which found it, proposed a severity score of 9.8.	MCP servers are code-execution paths. Silent fixes defeat scanners.	[18]

■ Analysis

Map your AI supply chain in six layers

It helps to see the AI stack as six layers, each with its own supply-chain risk.

1. **Models:** hosted APIs and downloaded open weights. Risk: backdoored or tampered weights, provider incidents.
2. **Gateways and key stores:** LiteLLM-style proxies and hosted consoles holding provider keys. Risk: key theft, release-pipeline compromise.
3. **Agents and assistants:** coding agents, IDE copilots, command-line agents. Risk: sandbox escape, inherited privileges, plugin tampering.
4. **Extensions:** plugins, skills and MCP servers. Risk: malicious or hijacked servers, broken pinning, silent fixes.
5. **Frameworks and SDKs:** orchestration platforms and agent SDKs. Risk: remote-code-execution flaws, worm-infected packages.
6. **Runtime exposure:** AI servers reachable from the internet. Risk: botnets, cryptomining, data theft.

Most organisations have an inventory for none of these layers.

Why AI components fail differently

- **Secret density.** One gateway can hold keys for every model provider you use, plus cloud credentials for logging and storage.
- **Agency.** A compromised agent or MCP server does more than leak data. It acts, with the developer's access, inside repositories and cloud consoles.
- **Speed without process.** Weekly releases, marketplace installs, "latest" Docker tags and fixes without CVEs undermine controls that work elsewhere.

The cost is more than a breach

Stolen keys mean someone else's attacks get billed to you. MONST3R-AI and x47.c are sold to use stolen AI keys to pay for attacks on third parties or to drain a victim's budget. That is a financial and reputational exposure even when no customer data is touched.

The regulatory angle

Under the EU AI Act, providers of high-risk AI systems must ensure accuracy, robustness and cybersecurity (Article 15), and deployers must use systems as instructed and monitor them (Article 26).^[19] Following the Digital Omnibus on AI, in force since 27 July 2026, these duties apply from 2 December 2027 for Annex III uses and 2 August 2028 for AI in regulated products.^[20] If your AI stack feeds a future high-risk system, you will need evidence that its components are under control. ISO/IEC 42001 Annex A also expects you to manage supplier and third-party relationships for AI systems (A.10).

Under the GDPR, a stolen gateway key that exposes prompts or logs containing personal data can be a personal data breach.^[21] Processors in the chain need proper contracts (Article 28) and appropriate security (Article 32).

■ A control framework for AI supply-chain security

The mappings are indicative. Use them to find where each control sits in your existing management system. They are not a certification claim.

#	Control	What good looks like	ISO/IEC 42001	NIST AI RMF	EU AI Act	OWASP LLM (2025)	GDPR
1	AI bill of materials	An inventory of models, gateways, agents, plugins, MCP servers, SDKs and AI servers, with versions and owners.	A.4; A.10	MAP, GOVERN	Art. 11, Annex IV (high-risk providers)	LLM03	Art. 30
2	Pin and verify	Pin exact versions and hashes. Never deploy "latest" to production. Verify signatures and provenance where available.	A.6	MANAGE	Art. 15	LLM03	Art. 32
3	Cooling-off period	Hold new releases of critical AI components for a short period before adopting them, unless they are security fixes.	A.6	MANAGE	Art. 15	LLM03	Art. 32
4	MCP and plugin allow-list	Only approved MCP servers and plugins. Each is reviewed for publisher, permissions and transport. Remote servers are watched for ownership changes.	A.10	GOVERN, MANAGE	Art. 15	LLM03, LLM06	Art. 28 where personal data
5	Key vaulting and spend caps	Provider keys live in a secrets manager, scoped per application, with spend caps at the provider. Rotated on a schedule and after any incident.	Cl. 6.1	MANAGE	Art. 15	LLM10	Art. 32
6	Developer endpoint protection	AI-tool config and token files are protected and monitored. Endpoint detection covers infostealer access to them.	Cl. 6.1	MANAGE	Art. 15	LLM02	Art. 32
7	Agent sandboxing	Coding agents run in containers or VMs with no access to production credentials. Outbound network access is restricted.	A.6; A.9	MANAGE	Art. 15	LLM06	Art. 25, 32
8	No exposed AI servers	Gateways, inference servers and orchestration tools sit behind authentication and are never directly reachable from the internet.	A.6	MANAGE	Art. 15	LLM10	Art. 32
9	Model provenance	Open weights come only from vetted publishers, with hash checks. Modified weights (abliterated, merged, fine-tuned) are not used in production without review.	A.6; A.10	MAP, MEASURE	Art. 53 (GPAI provider documentation)	LLM03, LLM04	-
10	Vendor due diligence	AI-tool vendors are reviewed for release security, incident history and disclosure practice, including whether they issue CVEs.	A.10	GOVERN	Art. 25	LLM03	Art. 28
11	Usage anomaly monitoring	Alerts on unusual AI spend, new models or regions, or off-hours use per key.	A.6; Cl. 9.1	MEASURE	Art. 26(5)	LLM10	Art. 32
12	Rapid-response runbook	A tested playbook for "AI component compromised": which keys to revoke, which systems to rebuild, who to notify.	Cl. 10.2	MANAGE	Art. 73 (high-risk providers)	-	Art. 33-34

EU AI Act Articles 11, 15, 25, 26 and 73 apply to high-risk systems, from 2 December 2027 (Annex III) or 2 August 2028 (Annex I). Article 53 applies to general-purpose AI model providers and matters here as upstream information you can ask for. ISO/IEC 42001 has no dedicated secrets control, so controls 5 and 6 sit under risk treatment (Cl. 6.1), usually alongside ISO/IEC 27001. OWASP numbering follows the 2025 edition: LLM02 Sensitive Information Disclosure, LLM03 Supply Chain, LLM04 Data and Model Poisoning, LLM06 Excessive Agency, LLM10 Unbounded Consumption. The 2026 edition, released in August 2026, rennumbers several entries.

■ Recommendations: a 90-day plan

Days 1 to 30

- Build the AI bill of materials. Start with gateways and anything that holds provider keys.
- Check whether you ran LiteLLM 1.107.0, Tensorlake SDK 0.5.144, OpenCode Console with stored provider keys, an OpenCode version from 1.0.137 to 1.18.32 on Windows or WSL, or the @apexacc packages. For LiteLLM, that includes any unpinned pip install on 21 October between 11:39 and 13:23 UTC, and the Docker tags latest, main-latest or main-stable pulled between 11:46 and 14:20 UTC. If so, rotate credentials and follow the vendors' published guidance.
- Confirm the fixed versions are in place: Claude Code 2.1.179 or later, Codex 0.146.0 or later (Plugin4Shell), Codex Desktop build 26.818.21641 and Codex CLI 0.149.0 or later (sandbox escapes), Langflow 1.9.0 or later, and Microsoft's 13 October update for Visual Studio and SQL Server Management Studio.
- Set hard spend limits on every AI provider account.

Days 31 to 60

- Move provider keys into a secrets manager, scoped per application.
- Ban "latest" tags for AI components in production. Pin versions and check hashes.
- Publish an MCP server and plugin allow-list, and block everything else on managed devices.

Days 61 to 90

- Add AI tooling to vendor risk reviews and procurement.
- Sandbox coding agents away from production credentials.
- Run a tabletop exercise on an AI gateway compromise.

FOR THE BOARD

AI tooling is now part of your software supply chain. Ask one question: does it sit inside our third-party risk programme, or outside it?

■ How PeritumAI can help

We provide readiness, gap-analysis, preparation and advisory services. We do not issue certifications or perform certification audits, and our work does not replace legal advice.

AI Risk Snapshot

A fast, fixed-scope review of up to three AI tools or use cases, including where your provider keys live and who can reach them.

AI Governance Readiness Assessment

An AI bill of materials, supplier risk review against ISO/IEC 42001 Annex A and the EU AI Act, and a remediation roadmap.

AI Act & Policy Preparation Package

AI tooling and MCP policy, a vendor due-diligence process and an incident runbook.

Fractional AI Security & Governance Advisor

Review and sign-off of new AI tools, plugins and MCP servers before they are approved.

Start a conversation: peritumai.com/consulting or danilo@peritumai.com.

■ Method and limits

How we built this paper. PeritumAI runs a daily pipeline that collects public reporting on AI security incidents and supply-chain breaches, and archives the full text of each source with its date. For this paper we selected incidents affecting AI tooling, published between 18 September and 22 October 2026. We read each archived source and checked every fact in the paper against it. Legal dates were checked against the official texts on EUR-Lex.

Limits. Several incidents were still under investigation when reported, and later findings may change the details. Version numbers and timings are as stated by the vendors and researchers. Where a source is vendor research, we say so. The control mappings are our own reading of each framework and are not an official crosswalk.

Corrections. If we got something wrong, tell us at danilo@peritumai.com. We fix errors in public and note the change on the paper's page. This is version 1.1, published 25 October 2026. Version 1.1 corrects wording that overstated what some sources said, labels sponsored sources, and tightens several framework mappings after an independent review of every source.

■ Sources

1. Help Net Security, "Zero-click RCE vulnerability hit four major AI coding agents, two remain unpatched", 18 Sep 2026. <https://www.helpnetsecurity.com/2026/09/18/plugin4shell-ai-coding-agents-vulnerability/>
2. The Hacker News, "Plugin4Shell Lets Repository Owners Swap Pinned Plugin Code Across Four AI Coding Agents", 18 Sep 2026. <https://thehackernews.com/2026/09/plugin4shell-lets-repository-owners.html>
3. BleepingComputer, "Researchers escape OpenAI Codex sandbox to run commands on host", 20 Sep 2026. <https://www.bleepingcomputer.com/news/security/researchers-escape-openai-codex-sandbox-to-run-commands-on-host/>
4. SecurityWeek, "New x47.c Windows Botnet Weaponizes xAI Grok, AI API Draining", 26 Sep 2026. <https://www.securityweek.com/new-x47-c-windows-botnet-weaponizes-xai-grok-ai-api-draining/>
5. BleepingComputer (sponsored, written by SOCRadar), "80,000+ Organizations Had AI Logins Stolen: From Shadow AI to LLMjacking", 28 Sep 2026. <https://www.bleepingcomputer.com/news/security/80-000-plus-organizations-had-ai-logins-stolen-from-shadow-ai-to-llmjacking/>
6. BleepingComputer, "Poellm malware infects exposed AI servers in cryptomining attacks", 7 Oct 2026. <https://www.bleepingcomputer.com/news/security/poellm-malware-infects-exposed-ai-servers-in-cryptomining-attacks/>
7. The Register, "Shai-Hulud worm makes jump to AI infrastructure with Tensorlake compromise", 8 Oct 2026. <https://www.theregister.com/security/2026/10/08/shai-hulud-worm-makes-jump-to-ai-infrastructure-with-tensorlake-compromise/5302054>
8. Embrace The Red, "Visual Studio Copilot and SSMS Remote Code Execution (CVE-2026-103572)", 13 Oct 2026. <https://embracethered.com/blog/posts/2026/visual-studio-copilot-remote-code-execution/>
9. The Hacker News, "Langflow's Critical Authenticated RCE Flaw Targeted in 1,423 Honey-pot Exploit Attempts", 15 Oct 2026. <https://thehackernews.com/2026/10/langflows-critical-authenticated-rce.html>
10. Infosecurity Magazine, "New Aoki Infostealer Targets AI Coding Assistant Credentials", 15 Oct 2026. <https://www.infosecurity-magazine.com/news/aoki-infostealer-ai-coding/>
11. Help Net Security, "Stolen AI API keys are paying for hackers' autonomous scans", 16 Oct 2026. <https://www.helpnetsecurity.com/2026/10/16/stolen-ai-api-keys-autonomous-scans/>
12. Schneier on Security, "Backdooring Open-Weight AI Models", 16 Oct 2026. <https://www.schneier.com/blog/archives/2026/10/backdooring-open-weight-ai-models.html>
13. The Hacker News, "OpenCode Console Breach May Have Exposed OpenAI, Anthropic, and Google API Keys", 19 Oct 2026. <https://thehackernews.com/2026/10/opencode-console-breach-may-have.html>
14. Infosecurity Magazine, "OpenCode Breach Could Affect OpenAI, Anthropic and Google", 19 Oct 2026. <https://www.infosecurity-magazine.com/news/opencode-breach-openai-anthropic/>
15. The Hacker News, "Fake Web3 Accelerator Uses npm Packages to Spread Malware and Hijack AI Agent Traffic", 19 Oct 2026. <https://thehackernews.com/2026/10/fake-web3-accelerator-uses-npm-packages.html>
16. Help Net Security, "LiteLLM hit by second supply chain attack this year, users urged to rotate credentials", 22 Oct 2026. <https://www.helpnetsecurity.com/2026/10/22/litellm-compromised-supply-chain-attack-pypi-docker/>
17. The Hacker News, "LiteLLM Suspects GitHub Token Leak in AI Gateway Logs Behind Malicious Package Release", 22 Oct 2026. <https://thehackernews.com/2026/10/litellm-suspects-github-token-leak-in.html>

18. The Hacker News, "OpenCode Flaw Lets Malicious MCP Servers Execute Code During Sign-In on Windows and WSL", 22 Oct 2026. <https://thehackernews.com/2026/10/opencode-flaw-lets-malicious-mcp.html>
19. Regulation (EU) 2024/1689 (Artificial Intelligence Act), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
20. Regulation (EU) 2026/1744 (Digital Omnibus on AI), OJ L 24 Jul 2026, in force 27 Jul 2026, EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj>
21. Regulation (EU) 2016/679 (General Data Protection Regulation), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

All web sources accessed and checked against PeritumAI's archived copies on 25 October 2026.