

Presseinformation

State of AI Security-Studie: KI-Angriffe treffen im ersten Halbjahr 2026 nicht das Modell, sondern sein Umfeld

Exponierte MCP-Server, veraltete KI-Infrastruktur und kompromittiertes Entwicklervertrauen sind im ersten Halbjahr 2026 die grössten Risikotreiber laut einer Studie von TrendAI

Neue Studienergebnisse von TrendAI zeigen: KI-Sicherheitsvorfälle im ersten Halbjahr 2026 entstehen selten durch Angriffe auf Modelle selbst. Sie entstehen durch bekannte Schwachstellen – unauthentifizierte Zugänge, ungepatchte Software, schwaches Secret-Management, missbrauchtes Vertrauen – die sich im Tempo von KI-Systemen schneller verbreiten als je zuvor.

Der Bericht [*Aftershocks: The Hidden Cost of Speed at Scale — 1H 2026 State of AI Security*](#) basiert auf TrendAI-Telemetriedaten, eigener Schwachstellenforschung, einer Befragung der Teilnehmer des Pwn2Own-Wettbewerbs Berlin 2026 sowie der Analyse öffentlich gemeldeter Sicherheitsvorfälle.

Das sind die wichtigsten Erkenntnisse des Berichts im Überblick:

- 43.175 KI-bezogene Dienstinstanzen in 25 Branchen identifiziert, davon 1.468 Tools, die ausschliesslich in einer einzigen Branche eingesetzt werden.
- 1.467 exponierte MCP-Server, darunter Instanzen mit unauthentifiziertem Zugriff auf Patientendaten.
- 12.461 frei erreichbare Ollama-Instanzen, überwiegend mit veralteten Softwareständen.
- 85 Prozent der Pwn2Own-Teilnehmer nutzen selbst LLMs in ihrer Sicherheitsforschung; der Wettbewerb brachte 47 Zero-Days und rund 1,3 Millionen US-Dollar Preisgeld hervor – ein neuer Rekord.
- Über 70 Prozent der im Untergrund nachgefragten Exploits zielen auf Schwachstellen, die älter als zwei Jahre sind.

Bekannte Schwachstellen, neues Tempo

Der Bericht widerspricht der Annahme, KI bringe grundlegend neue Angriffsarten hervor. Stattdessen beschleunigt sie bestehende: Die Analyse des Pwn2Own-Wettbewerbs zeigt, wie KI-gestützte Schwachstellenanalyse die Zeit von der Offenlegung einer Lücke bis zur einsatzfähigen Exploitation von Wochen auf Stunden verkürzen kann. Gleichzeitig bleibt die Exploit-Ökonomie im Untergrund träge – mehr als 70 Prozent der nachgefragten Exploits richten sich gegen Schwachstellen, die älter als zwei Jahre sind. Für Verteidiger bedeutet das, dass

Offenlegungsdatum, CVSS-Score oder Exploit-Wahrscheinlichkeit sich nicht mehr getrennt und im gewohnten Tempo bewerten lassen. Das Zeitfenster zwischen Patch und Ausnutzung schliesst sich schneller, unabhängig davon, ob eine Lücke neu oder Jahre alt ist.

KI-Infrastruktur wird zur kritischen Infrastruktur und bleibt oft ungeschützt

MCP-Server, die als Schnittstelle zwischen KI-Agenten und Unternehmenssystemen fungieren, werden vielerorts wie gewöhnliche Utilities behandelt, obwohl sie privilegierten Zugriff auf Datenbanken, Dateien und Anwendungen haben. Ein aktueller Scan fand 1.467 exponierte Instanzen; auf 70 Hosts war die Funktion `execute_sql` erreichbar, mindestens drei Server erlaubten unauthentifizierten Zugriff auf Patientendaten. Eine Analyse von mehr als 19.000 MCP-Server-Quellcodes ergab zudem, dass 48 Prozent der Projekte empfehlen, Zugangsdaten in unsicheren `.env`-Dateien oder Klartext-Konfigurationen zu speichern.

Dasselbe Muster zeigt sich bei Inferenz-Engines und Vektordatenbanken. Von 12.461 exponierten Ollama-Servern lief nur eine Handvoll auf der aktuellsten Version, die meisten deutlich dahinter. Die Vektordatenbank Chroma, 2025 und 2026 bereits bei Pwn2Own erfolgreich angegriffen, findet sich am häufigsten in einer veralteten Version im Netz. Popularität korreliert dabei nicht mit Update-Disziplin – bei Ollama gilt eher das Gegenteil: Je verbreiteter die Software, desto grösser die Zahl veralteter, angreifbarer Instanzen.

Blinde Flecken entstehen schon bei der Einführung: Ein Grossteil des identifizierten Risikos liegt nicht bei bekannten Anbietern wie OpenAI, Microsoft Copilot oder Google Gemini, sondern bei spezialisierten Branchenlösungen wie etwa in der Rechtsrecherche, Betrugserkennung oder medizinischen Dokumentation, die viele Governance-Programme schlicht nicht erfassen.

Der eigentliche Angriffsvektor: Entwicklervertrauen statt Modellangriff

Die folgenreichsten Vorfälle des Halbjahres bestätigen dieses Bild: Sie entstanden nicht durch Angriffe auf KI-Modelle, sondern durch kompromittierte Software-Lieferketten. Präparierte Versionen des KI-Gateways LiteLLM enthielten einen mehrstufigen Mechanismus zum Diebstahl von Zugangsdaten und zur lateralen Bewegung in Kubernetes-Umgebungen; entdeckt wurde der Angriff nur durch einen Programmierfehler in der Schadsoftware selbst. Die kompromittierte VS-Code-Erweiterung Nx Console lieferte einen Credential-Stealer aus. Nach einem Verpackungsfehler bei Claude Code tauchten binnen 24 Stunden gefälschte GitHub-Repositories mit angeblich „geleakten“ Dateien auf, die tatsächlich Schadsoftware verbreiteten.

Angrifer benötigen also keinen Durchbruch gegen ein KI-Modell, wenn Paketregistries, IDE-Erweiterungen, GitHub-Mechanismen oder OAuth-Berechtigungen denselben Zugang zuverlässiger liefern. Dasselbe Prinzip zeigt sich bei Cyberkriminalität und Agentic AI: Kriminelle Akteure industrialisieren zunehmend den Missbrauch öffentlicher Identitäten, Lieferantenbeziehungen und Plattformvertrauen, statt eigene Angriffswerkzeuge zu entwickeln.

Über die Studie

Der Bericht wertet TrendAI-Telemetriedaten aus dem ersten Halbjahr 2026, eigene Schwachstellenforschung, eine Befragung der Teilnehmer des Pwn2Own-Wettbewerbs Berlin 2026 sowie öffentlich dokumentierte Sicherheitsvorfälle aus. Er knüpft an den vorherigen TrendAI-Report [*Fault Lines in the AI Ecosystem*](#) an.

Weitere Informationen

Den vollständigen Bericht *Aftershocks: The Hidden Cost of Speed at Scale — 1H 2026 State of AI Security* finden Sie in englischer Sprache hier: <https://www.trendaisecurity.com/de/resources-insights/deep-research/1h-2026-state-of-ai-security>

Über TrendAITM

TrendAITM, weltweit führender Anbieter von KI-Sicherheit und der Enterprise-Geschäftsbereich von Trend Micro, ermöglicht Organisationen vollständige Transparenz und Governance ihrer KI-Nutzung sowie konsolidierte Sicherheit. AI Fearlessly.

Pressekontakt:

Brand Affairs AG

Mischa Keller

Mühlebachstrasse 8

8008 Zürich

Telefon: +41 (0) 79 380 95 86

E-Mail: mischa.keller@brandaffairs.ch

Internet: www.brandaffairs.ch