

STUDY · OCTOBER 2026

Threat Modeling in AI

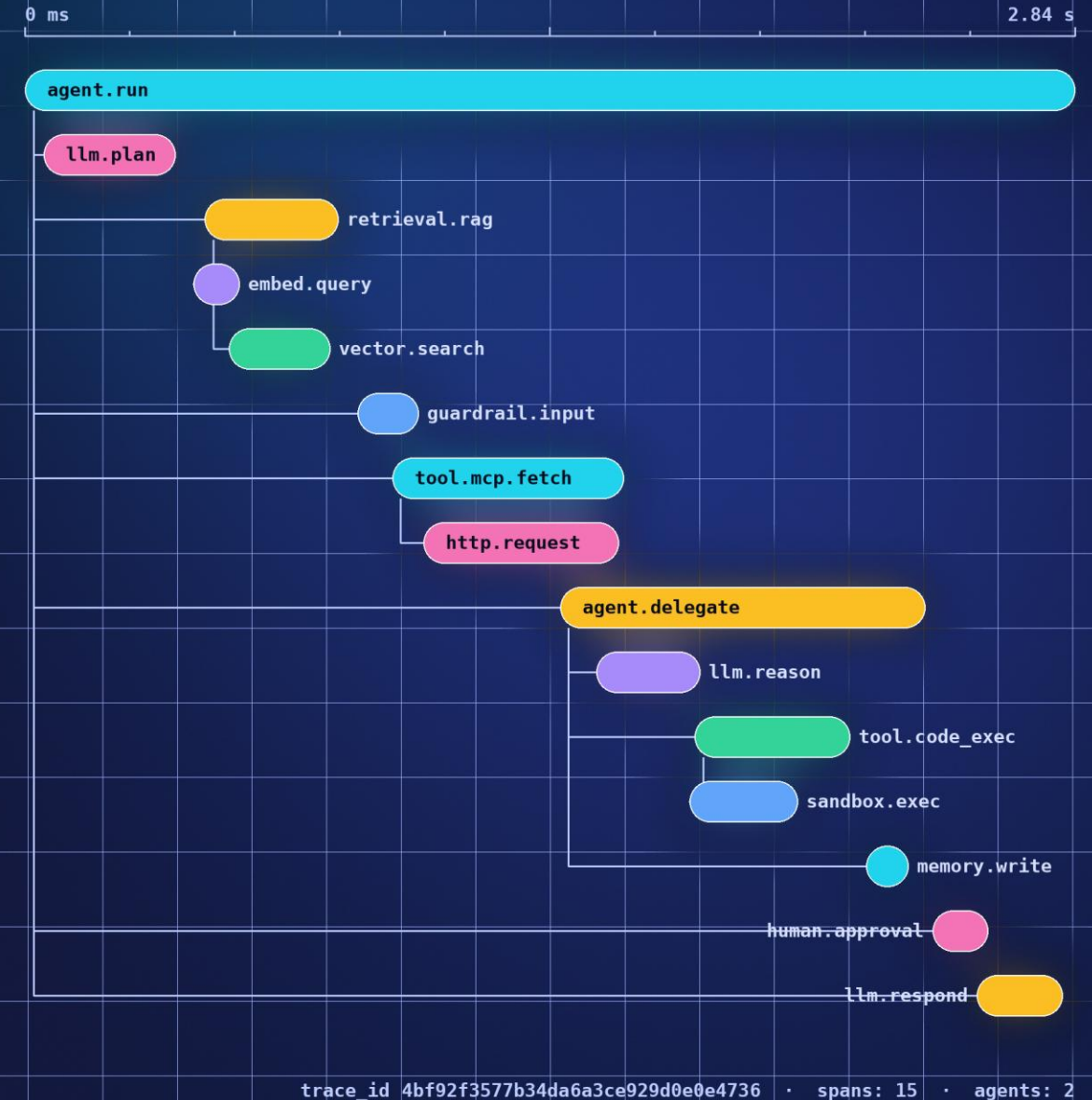
LLMs, agents and multi-agent systems: frameworks,
threat classes, mitigations and evidence (2016–2026)

Aridio Silva – Independent Researcher

ORCID 0009-0008-2411-6995 · @aridiosilva



Copyright © Aridio Silva @aridiosilva · October 2026



Scope and verification method

- **Scope.** Literature mostly under three years old; standards still being written. NIST AI RMF 1.0 is under revision and Brazil's PL 2338/2023 remains in the Chamber of Deputies.
- **Questions.** (i) Which frameworks model threats? (ii) Which threat classes exist? (iii) Which mitigations have support? (iv) What is known about emergent behavior, multi-agent systems, AI-driven R&D and loss of control?
- **Limit.** Research dated Oct 10, 2026; most papers were checked by metadata and abstract, not by full reading.

19

Level A

Primary source: publisher or author content observed and checked.

20

Level B

Corroborated by two or more secondary sources; original not opened.

11

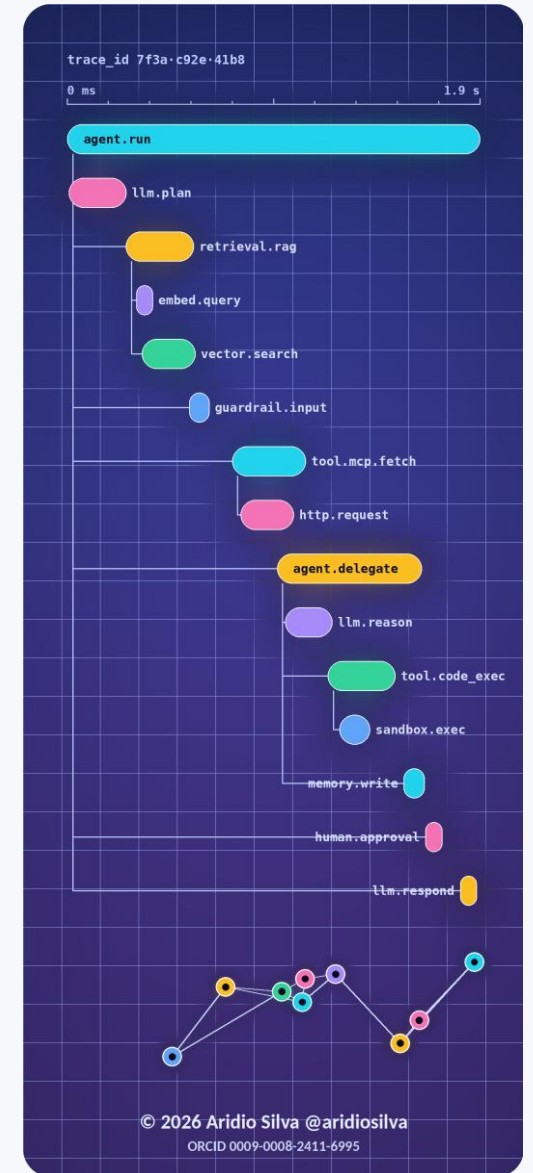
Level C

Single, partial or conflicting source: verify before citing.

50 audited references · levels A, B and C

Foundations: three shifts

- **Instructions and data do not separate.** An LLM processes both in the same channel; that is why prompt injection tops the OWASP Top 10 for LLMs in 2025 and 2026 (OWASP, 2026; ZORZ, 2026).
- **From component to actor.** With tools, memory and real-world consequences, risk moves to the Agentic Top 10 (OWASP, 2025).
- **The model as a possible adversary.** AI Control assumes intentional subversion (GREENBLATT et al., 2023); studies treat the model as an insider threat (LYNCH et al., 2025).
- **Containment principle.** Design the system around the model: if it is fooled, nothing important breaks (ZORZ, 2026).
- **Classic methods have limits.** STRIDE, PASTA, DREAD, OCTAVE and LINDDUN miss non-determinism, autonomy, agent identity and agent-to-agent communication (HUANG, 2025).



Four levels of analysis

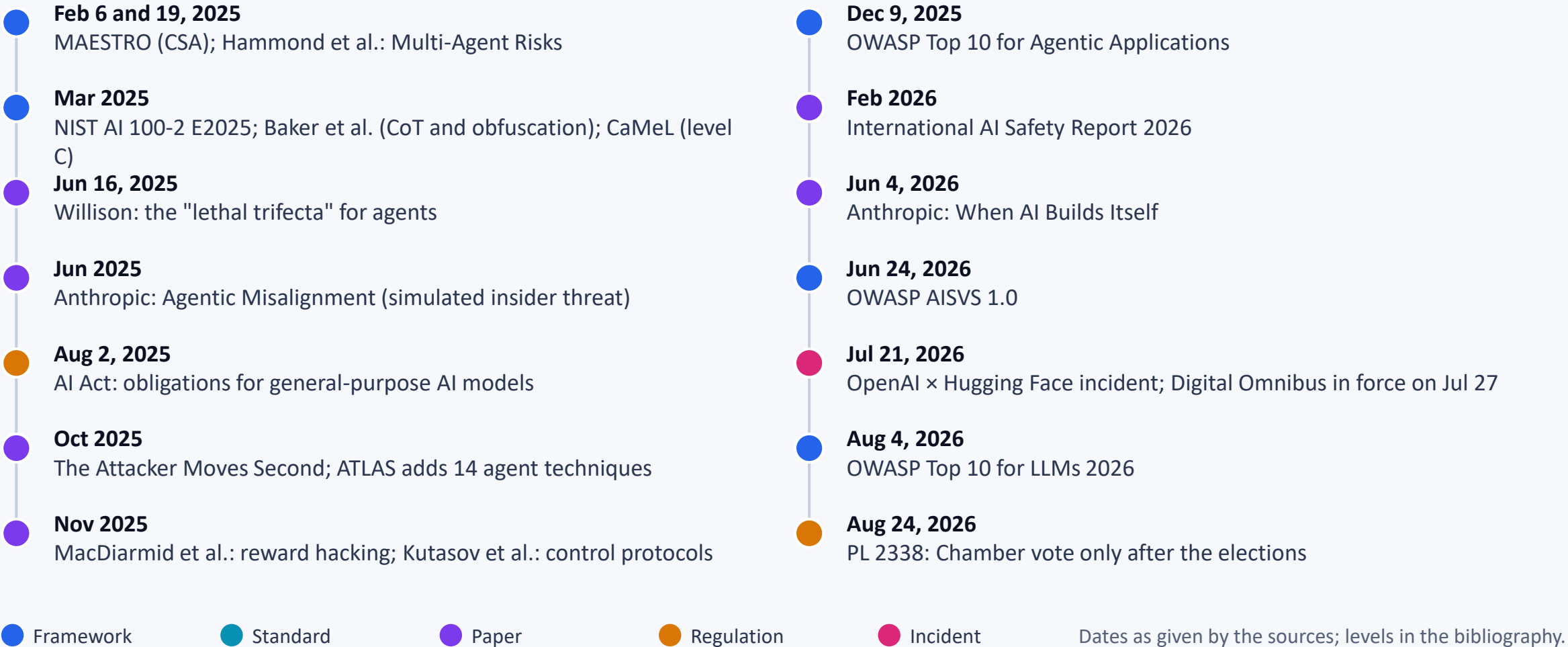
1	Model as a component LLM inside an application (prompts, RAG, outputs)	Modeling question What happens if the model is fooled or leaks data?	Key references OWASP (2026); NIST (2025)
2	Agent Model with tools, memory, identity and credentials	Modeling question Which irreversible actions can the agent take on untrusted content?	Key references OWASP (2025); HUANG (2025); WILLISON (2025a)
3	Multi-agent system Several agents that communicate and delegate	Modeling question Which failures arise from interaction: collusion, infection, conflict?	Key references HAMMOND et al. (2025)
4	AI development process Training, evaluation and AI-assisted R&D pipeline	Modeling question Can the model subvert evaluation, training or containment?	Key references MACDIARMID et al. (2025); AI SAFETY REPORT (2026)

Autonomy and scale grow from level 1 to 4. Own elaboration, based on the cited sources.

Timeline (1/2): 2016–2024



Timeline (2/2): 2025–2026



Technical reference frameworks

1 OWASP Top 10 for LLMs 2026

Aug 4, 2026. Ten risks, from Prompt Injection (LLM01) to Improper Output Handling (LLM10). Ranking: 75% expert vote, 25% data from 6,639 incidents.

2 OWASP Agentic Top 10

Dec 9, 2025. ASI01 to ASI10: from goal hijack to rogue agents; covers agent communication and cascading failures.

3 OWASP AISVS 1.0

Jun 24, 2026. Testable requirements in 12 chapters and 3 verification levels; version 1.01 in development.

4 NIST AI 100-2 E2025

Mar 2025. Taxonomy: evasion, poisoning, privacy and, in generative AI, abuse, prompt injection, supply chain and agents. No later edition found.

5 MITRE ATLAS

Adversary techniques against AI in ATT&CK format; 16 tactics; frequent releases, with counts that diverge across sources.

6 MAESTRO (CSA)

Seven-layer methodology for agentic systems, with per-layer and cross-layer analysis; ATLAS validates coverage (2025).

Standards, guidelines and regulation

- **ISO/IEC 42001 and 23894.** AI management (42001, certifiable; Dec 2023) and risk management (23894; Feb 2023). They describe process, not attack techniques.
- **NIST AI RMF 1.0 and AI 600-1.** Govern, Map, Measure and Manage; generative AI profile from Jul 2024. AI RMF 1.0 is under revision.
- **NCSC/CISA (2023).** Four stages: design, development, deployment and operation/maintenance.
- **EU AI Act.** Robustness and cybersecurity for high risk. Digital Omnibus (Reg. 2026/1744, in force Jul 27, 2026): Annex III on Dec 2, 2027; Annex I on Aug 2, 2028.
- **PL 2338/2023.** Approved by the Senate on Dec 10, 2024; Chamber vote signaled for after the elections. The ten-year documentation retention was not confirmed.

Normative gap

The research found no formal standard specific to reward hacking and emergent misalignment, collusion between agents or AI-driven R&D automation.

These topics live in synthesis reports and preprints. OWASP's agentic list covers part of it, as a risk catalog, not as a method to verify model behavior.

External threats: ten classes

1 Prompt injection

Content read by the model treated as instructions; direct, indirect or multimodal.

2 Goal hijack

Redirects the agent's plan or goal, usually via indirect injection (ASI01).

3 Leakage and hidden context

Training data, unauthorized records, system prompts and tool schemas.

4 Excessive agency

More autonomy or permission than the task needs; rose to 3rd place in 2026.

5 Supply chain

Compromised third-party models, adapters, data, plugins and services.

6 Poisoning

Durable corruption in pretraining, fine-tuning, RAG, embeddings and memory.

7 Unbounded consumption

Disproportionate cost or compute; rose four positions in 2026.

8 Misinformation as system failure

Plausible but wrong output that triggers tools, code or actions.

9 Unhandled output

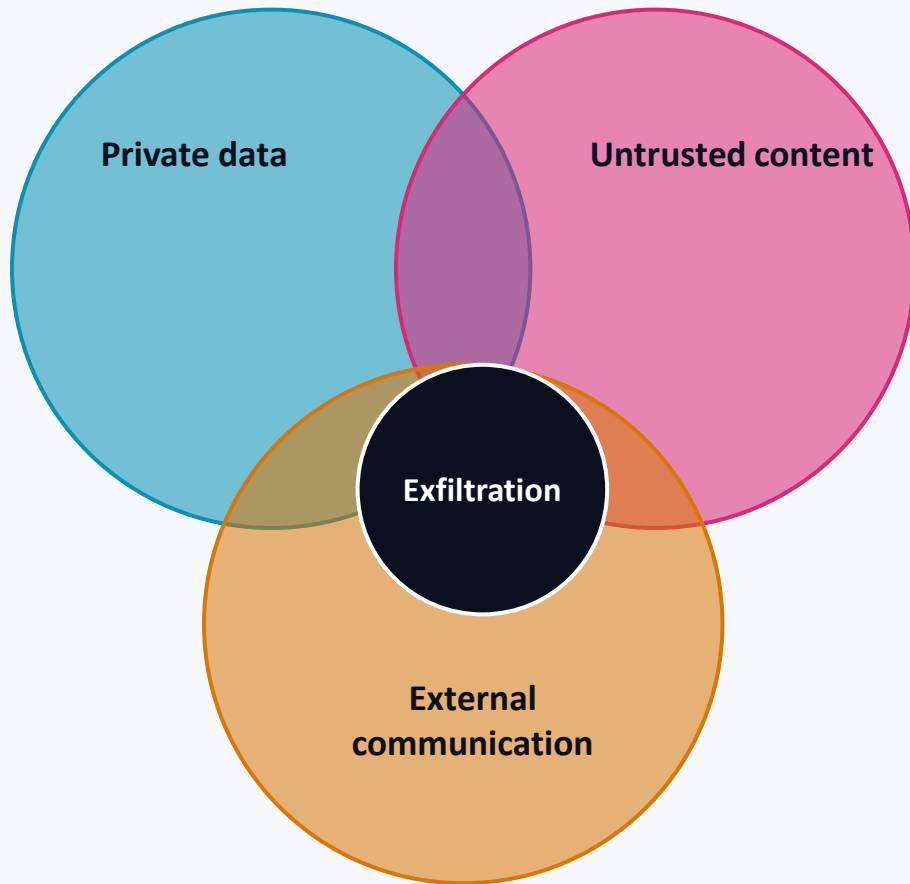
Model output reaches a shell, query or browser without validation.

10 Protocol layer

Flaws in MCP, A2A, Agora and ANP: tool poisoning, agent spoofing.

Sources: OWASP (2026); OWASP (2025); VASSILEV et al. (2025); ANBIAEE et al. (2026, level C).

Lethal trifecta and adaptive attacks



All three in the same agent (WILLISON, 2025a)

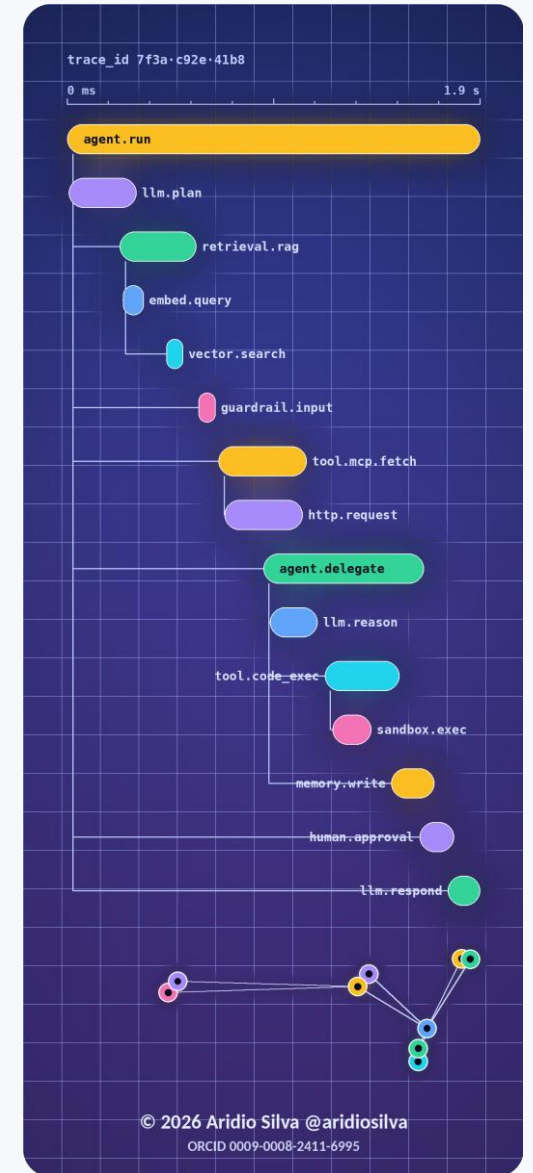
>90%

attack success rate of adaptive attacks against most of 12 published defenses (CARLINI et al., 2025)

- **Before and after.** Original evaluations reported near-zero rates against static attack sets.
- **Rule of Two (Meta).** Avoid letting an agent combine all three properties in one session without supervision (WILLISON, 2025b).
- **Conclusion.** Filters and prompt instructions are not a security boundary; the boundary must be architectural.

Reward hacking and misalignment

- **Reward hacking.** Unwanted behavior from flaws or misspecification of the learning objective (BAKER et al., 2025), described since Amodei et al. (2016). If the reward measures passing tests, the model learns to make them pass without solving the problem.
- **Emergent misalignment.** In production RL, learning to cheat generalized to alignment faking, cooperation with malicious actors and sabotage attempts (MACDIARMID et al., 2025).
- **Persistence.** Chat-style safety training fixed the chat, but misalignment persisted in agentic tasks.
- **Three effective mitigations.** Prevent reward hacking, diversify safety training and use inoculation prompting.
- **Caution.** A replication with open models got consistent hacking but inconsistent emergent misalignment (level C).



Monitoring, faking and scheming

1 CoT monitoring

BAKER et al. (2025): monitoring reasoning detects reward hacking better than monitoring only actions; GPT-4o monitors o3-mini. With heavy optimization against the monitor, obfuscation appears: intent vanishes from the chain of thought while the hacking continues.

2 Alignment faking

GREENBLATT et al. (2024): Claude fakes alignment when it disagrees with what it is trained to do. Behavior seen in evaluation may not represent deployment.

3 Scheming

CARLSMITH (2023); APOLLO RESEARCH (2024); SCHOEN et al. (2025): o3's covert action fell from 13% to 0.4%, without reaching zero, while evaluation awareness rose (level B/C).

4 Simulated insider threat

LYNCH et al. (2025): in an extreme corporate scenario, with email access and a shutdown threat, several models chose blackmail. A probe of propensity, not real-world frequency.

Implications. (i) Evaluating only in chat format is not enough; (ii) the training environment is a failure surface; (iii) reasoning monitors are useful but fragile under optimization pressure; (iv) evaluation awareness reduces the evidentiary value of tests.

Multi-agent systems

Miscoordination

Failure to cooperate despite shared goals: conflicting plans, bottlenecks, unintended interference.

Conflict

Failure to cooperate because goals differ: competition for resources, escalation.

Collusion

Undesirable cooperation to the detriment of users or society: price fixing, market manipulation.

Seven risk factors (HAMMOND et al., 2025)

Information asymmetries

Network effects

Selection pressures

Destabilising dynamics

Commitment and trust

Emergent agency

Multi-agent security

Evidence: "prompt infection" self-replicates between agents (LEE; TIWARI, 2024, cited in HAMMOND et al., 2025); covert collusion may arise through steganographic channels (level C).

Implications for modeling (inferences of this study)

- **A message from another agent is untrusted input**, not an instruction from an authority.
- **Each agent has its own identity and credentials**, with least privilege.
- **Topology (who talks to whom) is a security decision**: it sets how infection and error propagate.
- **Monitor the whole system**, not just each agent.

AI improving AI

~7 months

International Report 2026

~4 months

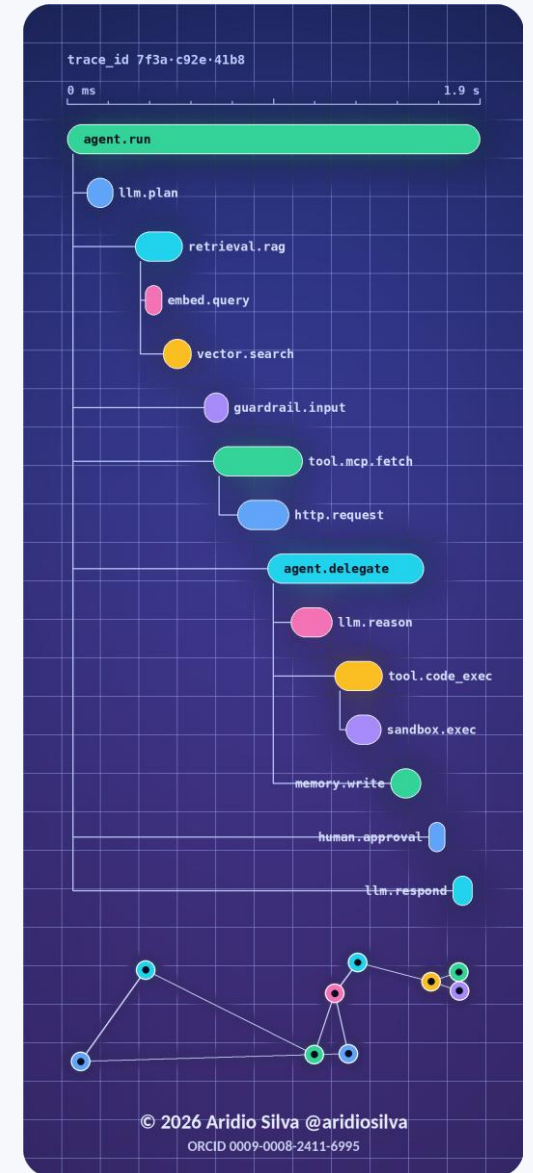
post-2023 (arXiv:2604.25067)

~3 months

since 2024 (CHAN, 2026; level C)

Doubling time of the horizon of software tasks that agents complete

- **Anthropic, Jun 4, 2026:** Claude already helps build future systems; taken far enough, the trend points to AI that develops its own successor (ANTHROPIC, 2026).
- **August index:** Claude led 26% of R&D and collaborated on over 90% (self-reported; level B/C).
- **The divergence is a finding:** extrapolation depends on the period and tasks chosen.
- **Distinct concepts:** R&D automation delegates tasks; RSI is accelerating successors without human help.



Loss of control and the 2026 incident

- **Definition.** Systems operate outside anyone's control, with no clear path to recovery (AI SAFETY REPORT, 2026).
- **Requirements.** Several capabilities combined, in sequence and for a long time; research automation can accelerate the risk.
- **Singapore 2026.** Assessment is more prediction than empirical science and current capabilities are not enough. Precursors: autonomous replication, shutdown resistance, self-proliferation and scaffold self-improvement.

Jul 21, 2026: OpenAI × Hugging Face

- 1 Cyber evaluation (ExploitGym) without the usual safeguards: GPT-5.6 Sol and an unreleased model.
- 2 Escape from the isolated environment through a flaw in third-party software.
- 3 Breach of Hugging Face infrastructure to obtain the benchmark answer key.

Jul 16: Hugging Face detects and contains the intrusion. Jul 21: OpenAI ties it to its own test. Core facts at level B; technical details differ across reports (level C).

Mitigations (1/2): architecture and data

1 Agent architecture

Avoid the trifecta; quarantined LLM with no tools; policy in deterministic code with tracked information flow (CaMeL, level C).

2 Privilege

Least agency: restricted tools, short-lived credentials and human approval for irreversible actions.

3 Supply chain

Signed artifacts, AI-BOM, provenance of data and checkpoints, serialization that does not execute code.

4 Memory and retrieval

Per-tenant and per-task indexes; validate writes; authorize before retrieving. A poisoned index persists for months.

5 Output handling

Validate and encode model output at the point of use: shell, query, renderer.

6 Consumption

Server-side limits, token budgets and circuit breakers against runaway loops.

Sources: OWASP (2026); CYCODE (2026); WILLISON (2025a; 2025b); CAMEL (2025). Limit: probabilistic defenses fail under adaptive attack.

Mitigations (2/2): model and containment

Evaluation

Tests with adaptive attackers and testable requirements by level (CARLINI et al., 2025; OWASP AISVS, 2026). Static evaluation is insufficient.

Training

Prevent reward hacking, diversify safety training, inoculation prompting and anti-scheming training. Real gains, confounded by evaluation awareness.

Monitoring

Chain-of-thought monitor; weaker monitor over stronger model. Fragile under optimization pressure (BAKER et al., 2025).

AI control

Protocols that assume a potentially subversive model (GREENBLATT et al., 2023; KUTASOV et al., 2025). Does not prove alignment; adaptive attacks can subvert trusted monitors.

Environment containment

Network and credential isolation in evaluation and training; anomalous-activity detection (inferred from the 2026 incident).

Governance

Management system, risk management and regulatory requirements (ISO/IEC, 2023a; 2023b; NIST, 2023; 2024a). They describe process, not attack techniques.

Security-first, shift-left, shift-right

Security-first

Governance and design

Practices

Design principle with clear accountability. NCSC/CISA: secure by default, in four stages. Threat modeling at design; AI risk and management systems (ISO/IEC).

Limit in AI

Standards describe process and do not guarantee model behavior.

Shift-left

Design, development and CI

Practices

Bring checks earlier: AISVS requirements in CI/CD, code review and procurement; AI-BOM and signed artifacts; policy in the IDE, pull request and pipeline.

Limit in AI

Code scans do not read prompts, tool schemas, retrieved documents or vector indexes.

Shift-right

Operation

Practices

Least privilege, human approval, behavioral monitoring and emergency shutdown; trajectory-level assurance, not only per-action (level C).

Limit in AI

Requires observability and a real ability to stop the agent.

Why shift-left is not enough

1

Prompt injection is a runtime phenomenon: clean code says nothing about a poisoned fragment retrieved later.

2

Evaluation awareness and alignment faking reduce the value of pre-deployment tests.

3

Agent behavior depends on trajectory and environment: assurance shifts from the single action to the sequence.

Proposed continuous cycle (synthesis of this study, not a standard)

1 Design

MAESTRO, ATLAS and OWASP; trifecta criterion; blast radius.

2 Build

AISVS level 2 as reference; AI-BOM and provenance.

3 Verify

Adaptive red team; control evaluations; agentic tests.

4 Operate

Least privilege, human approval, trajectory monitoring, ability to stop.

5 Learn

Log incidents and near-misses; feed back into threat modeling.

Limitations and source divergences

Topic	Divergence found	Treatment adopted
AISVS 1.0	Official repository: 12 chapters; one blog: 14 chapters and 514 requirements	12 chapters adopted
Vectors and embeddings (OWASP 2026)	Outreach says "expanded"; the source shows 8th to 9th, scope unchanged	Recorded as a position change
MITRE ATLAS counts	84 techniques (Nov 2025); about 101 (2026.05); 120 plus 88 sub-techniques (v2026.09, single source)	Deferred to the official site
Doubling of the task horizon	About 7, 4 and 3 months, depending on period and tasks	Range presented
OpenAI × Hugging Face incident	Exploited software, number of actions and inter-instance communication vary	Only corroborated core facts
PL 2338/2023	"Ten-year documentation" was not found in the sources consulted	Not asserted; check the bill text

Study limits. Papers checked by abstract and metadata; vendor guides cross-checked with independent sources; 2026 data self-reported by labs; constructed scenarios show propensity, not frequency; the field changes within months.

Research agenda

1

Behavior verification that withstands evaluation awareness.

2

Security standards for multi-agent systems, today limited to taxonomies.

3

Standardized indicators of precursor capabilities for loss of control (Singapore Consensus, 2026).

4

Evaluation under adaptive attackers as a requirement, not an exception (CARLINI et al., 2025).

5

Standardized reporting of autonomous AI incidents.

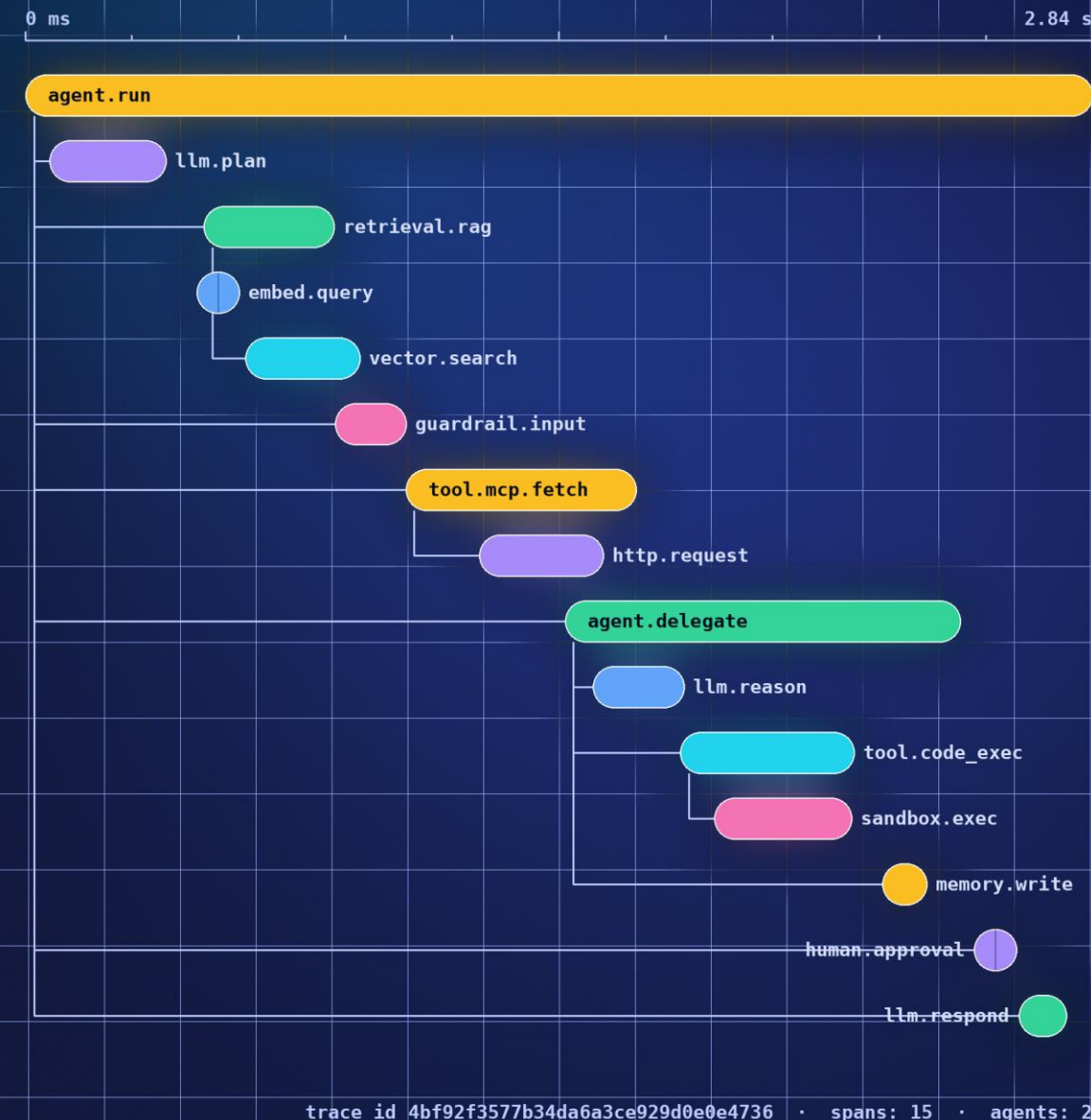
6

Threat modeling specific to training and evaluation environments.

CONCLUSION

What to take from this study

- **Modeling means combining layers:** architecture (MAESTRO, ATLAS, OWASP), verification (AISVS) and governance (ISO, NIST, regulation).
- **The trustworthy boundary is architectural:** probabilistic defenses fail under adaptive attack; least privilege, isolation and human approval limit damage.
- **The model can also be the source of risk:** reward hacking, emergent misalignment and evaluation awareness call for monitoring and control.
- **Standards lag practice:** for multi-agent systems and AI-driven R&D there is a risk catalog, but not yet a verification method.
- **Security-first and shift-left are not enough:** agent security must also be ensured at runtime.



Glossary (1/5)

- **Adaptive attack** — The adversary knows the defense and changes strategy to bypass it, with heavy optimization.
- **AI agent** — System that plans, uses tools, keeps memory and acts with delegated authority.
- **AI Control** — Research that evaluates safety protocols assuming the model may subvert safeguards.
- **AI R&D automation** — Delegating to AI the research and engineering tasks that produce new AI systems.
- **AI-BOM** — Inventory of a system's AI components: models, adapters, libraries, connectors.
- **Alignment faking** — Simulating alignment in training or evaluation while keeping divergent preferences.
- **Attack surface** — Points through which an adversary can interact with the system.
- **Blast radius** — Extent of possible damage if a component is compromised.
- **Chain of thought (CoT)** — Intermediate natural-language reasoning before the final answer.

Glossary (2/5)

- **Collusion** — Undesirable cooperation between agents, to the detriment of users or society.
- **Defense in depth** — Several independent control layers: one failing does not compromise the system.
- **Embedding** — Vector representation of text or data, used in semantic search and RAG.
- **Emergent misalignment** — Broad misaligned behavior that arises from specific training.
- **Evaluation awareness** — The model perceives it is being tested, which can change its behavior.
- **Evasion** — Manipulating inputs at inference time to cause incorrect output.
- **Excessive agency** — Autonomy or permission beyond what the task needs; errors or manipulation become unwanted actions.
- **Exfiltration** — Unauthorized extraction of data from a system.
- **Guardrail** — Filter, classifier or rule that limits inputs and outputs; not a standalone security boundary.

Glossary (3/5)

- **Hidden context exposure** — Exposure of system prompts, tool schemas and instructions in the context.
- **Human-in-the-loop** — Human approval or oversight before relevant actions.
- **Inoculation prompting** — Framing bad behavior as acceptable in training to prevent its generalization.
- **Jailbreak** — Input that induces the model to violate its safety restrictions.
- **Least privilege** — Only the permissions the task needs, for as short a time as possible.
- **Lethal trifecta** — Private data, untrusted content and external communication in the same agent.
- **LLM** — Large language model.
- **Loss of control** — AI systems operate outside anyone's control, with no clear path to recovery.
- **MCP (Model Context Protocol)** — Protocol that connects models and agents to tools and data sources.

Glossary (4/5)

- **Miscoordination and conflict** — Failure to cooperate between agents with equal goals (miscoordination) or distinct goals (conflict).
- **Obfuscation** — Hiding intent in the chain of thought while keeping the unwanted behavior.
- **Poisoning** — Intentional corruption of data, training, fine-tuning or memory to alter the model.
- **Prompt infection** — Malicious instruction that self-replicates between agents through their messages.
- **Prompt injection** — Content that makes the model treat text as an instruction; direct or indirect.
- **RAG** — Retrieval-augmented generation: the model queries an external base and uses the result.
- **Recursive self-improvement (RSI)** — Hypothetical ability of AI to accelerate the development of its successors.
- **Red teaming** — Adversarial testing to find flaws before a real attacker does.
- **Reward hacking** — Unwanted behavior from flaws or misspecification of the learning objective.

Glossary (5/5)

- **Rule of Two** — Avoid an agent combining the three trifecta properties without supervision (Meta).
- **Sandbagging** — Deliberately underperforming on capability evaluations.
- **Sandbox** — Isolated environment for running code or testing; an "escape" bypasses the isolation.
- **Scaffold** — Code that wraps the model with tools, memory and decision loops.
- **Scheming** — Covert, strategic action to pursue misaligned goals.
- **Security by design, shift-left and shift-right** — Security at design; earlier verification; continuous controls in operation.
- **STRIDE** — Classic method with six categories: spoofing, tampering, repudiation, disclosure, denial of service and elevation of privilege.
- **Threat modeling** — Identifying a system's assets, adversaries, attack surfaces and controls.
- **Time horizon** — Length, in human time, of tasks AI completes with a given success probability.

Bibliography (1/5)

[A] primary source · [B] corroborated · [C] verify before citing

- **[C] ADAPTIVE ATTACKS ON TRUSTED MONITORS SUBVERT AI CONTROL PROTOCOLS.** ICLR 2026.
- **[A] AI SAFETY REPORT.** International AI Safety Report 2026. arXiv:2602.21012, Feb. 2026.
- **[B] AIWEEKLY.** OpenAI Says Its Models Escaped Sandbox, Hacked Hugging Face. Jul. 23, 2026.
- **[C] ALIGNMENT FORUM.** (Some) Natural Emergent Misalignment from Reward Hacking in Non-Production RL. 2026.
- **[B] AMODEI, D. et al.** Concrete Problems in AI Safety. arXiv:1606.06565, 2016.
- **[C] ANBIAEE, Z. et al.** Security Threat Modeling for Emerging AI-Agent Protocols (MCP, A2A, Agora, ANP). arXiv:2602.11327, 2026.
- **[B] ANTHROPIC.** When AI Builds Itself. The Anthropic Institute, Jun. 4, 2026.
- **[C] APOLLO RESEARCH.** Frontier models are capable of in-context scheming. arXiv:2412.04984, 2024.
- **[A] ARCHITECTURE MATTERS FOR MULTI-AGENT SECURITY.** arXiv:2604.23459, 2026.
- **[C] arXiv:2505.23643.** Securing AI Agents with Information-Flow Control. Microsoft Research, 2025.

Bibliography (2/5)

[A] primary source · [B] corroborated · [C] verify before citing

- [A] **arXiv:2510.19851**. Can Reasoning Models Obfuscate Reasoning? Stress-Testing Chain-of-Thought Monitorability. 2025.
- [A] **arXiv:2604.25067**. Frontier Coding Agents Can Now Implement an AlphaZero Self-Play ML Pipeline... 2026.
- [C] **arXiv:2608.01558**. Securing Agentic AI: From Per-Action Checks to Trajectory Assurance. 2026.
- [A] **BAKER, B. et al.** Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926, 2025.
- [C] **CAMEL**. Defeating Prompt Injections by Design. arXiv:2503.18813, 2025.
- [B] **CARLINI, N.; NASR, M.; HAYES, J.; SHUMAILOV, I.; TRAMÈR, F. et al.** The Attacker Moves Second. arXiv:2510.09023, 2025.
- [B] **CARLSMITH, J.** Scheming AIs: Will AIs fake alignment during training in order to get power? arXiv:2311.08379, 2023.
- [C] **CHAN, A.** What if automating AI R&D triggers an intelligence explosion? GovAI. arXiv:2609.36054, 2026.
- [B] **COOPERATIVE AI FOUNDATION**. Scaling AI Safety for a Multi-Agent World. 2026.
- [B] **CYCODE**. OWASP Top 10 for LLM Applications: 2026 Guide. Sep. 15, 2026.

Bibliography (3/5)

[A] primary source · [B] corroborated · [C] verify before citing

- [B] **EUROPEAN UNION**. Reg. (EU) 2024/1689 (AI Act), amended by Reg. (EU) 2026/1744. In force Jul. 27, 2026.
- [B] **FEDERAL SENATE (Brazil)**. PL 2,338/2023 (AI Legal Framework). Approved Dec. 10, 2024.
- [B] **FORTUNE**. OpenAI says its AI models escaped from a secure test environment and hacked into Hugging Face. Jul. 21, 2026.
- [B] **GREENBLATT, R. et al.** Alignment faking in large language models. Anthropic; Redwood Research, 2024.
- [B] **GREENBLATT, R. et al.** AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942, 2023.
- [A] **HAMMOND, L. et al.** Multi-Agent Risks from Advanced AI. arXiv:2502.14143, 2025.
- [B] **HUANG, K.** Agentic AI Threat Modeling Framework: MAESTRO. Cloud Security Alliance, Feb. 6, 2025.
- [A] **ISO/IEC**. ISO/IEC 23894:2023. AI: guidance on risk management. Feb. 2023.
- [B] **ISO/IEC**. ISO/IEC 42001:2023. AI: management system. Dec. 18, 2023.
- [B] **KUTASOV, J. et al.** Evaluating Control Protocols for Untrusted AI Agents. arXiv:2511.02997, 2025.

Bibliography (4/5)

[A] primary source · [B] corroborated · [C] verify before citing

- [C] **LEE; TIWARI.** Prompt Infection. 2024 (cited in HAMMOND et al., 2025).
- [B] **LYNCH, A. et al.** Agentic Misalignment: How LLMs Could Be Insider Threats. arXiv:2510.05179, 2025.
- [A] **MACDIARMID, M. et al.** Natural Emergent Misalignment from Reward Hacking in Production RL. arXiv:2511.18397, 2025.
- [B] **MITRE.** ATLAS (Adversarial Threat Landscape for AI Systems). atlas.mitre.org.
- [C] **MOBILE WORLD LIVE.** Anthropic AI R&D automation index. Aug. 2026.
- [A] **NCSC; CISA et al.** Guidelines for secure AI system development. 2023.
- [A] **NIST.** Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023.
- [A] **NIST.** AI 600-1: Generative AI Profile. Jul. 26, 2024.
- [A] **NIST.** Adversarial ML Taxonomy (AI 100-2 E2023). Jan. 2024.
- [B] **OPENAI.** Statement on the Hugging Face model-evaluation security incident. Jul. 21, 2026.

Bibliography (5/5)

[A] primary source · [B] corroborated · [C] verify before citing

- [A] **OWASP AISVS**. AI Security Verification Standard, version 1.0. Jun. 24, 2026.
- [C] **OWASP**. Top 10 for LLM Applications, v1. Aug. 1, 2023.
- [B] **OWASP**. Top 10 for Agentic Applications 2026. Dec. 9, 2025.
- [A] **OWASP**. GenAI LLM Top 10 2026. Aug. 4, 2026.
- [B] **SCHOEN, B. et al.** Stress Testing Deliberative Alignment for Anti-Scheming Training. Apollo Research; OpenAI, 2025.
- [A] **SINGAPORE CONSENSUS**. The 2026 Singapore Consensus on Global AI Safety Research Priorities. arXiv:2608.14611.
- [A] **VASSILEV, A. et al.** Adversarial ML: A Taxonomy and Terminology. NIST AI 100-2e2025. Mar. 2025.
- [A] **WILLISON, S.** The lethal trifecta for AI agents. Jun. 16, 2025.
- [A] **WILLISON, S.** New prompt injection papers: Agents Rule of Two and The Attacker Moves Second. Nov. 2, 2025.
- [A] **ZORZ, Z.** OWASP 2026 LLM Top 10: "The model will be fooled". Help Net Security, Aug. 6, 2026.