NeuralTrust

The Agentic AI Security Framework

Designing Trust and Resilience for the Age of Agentic AI

January 2026

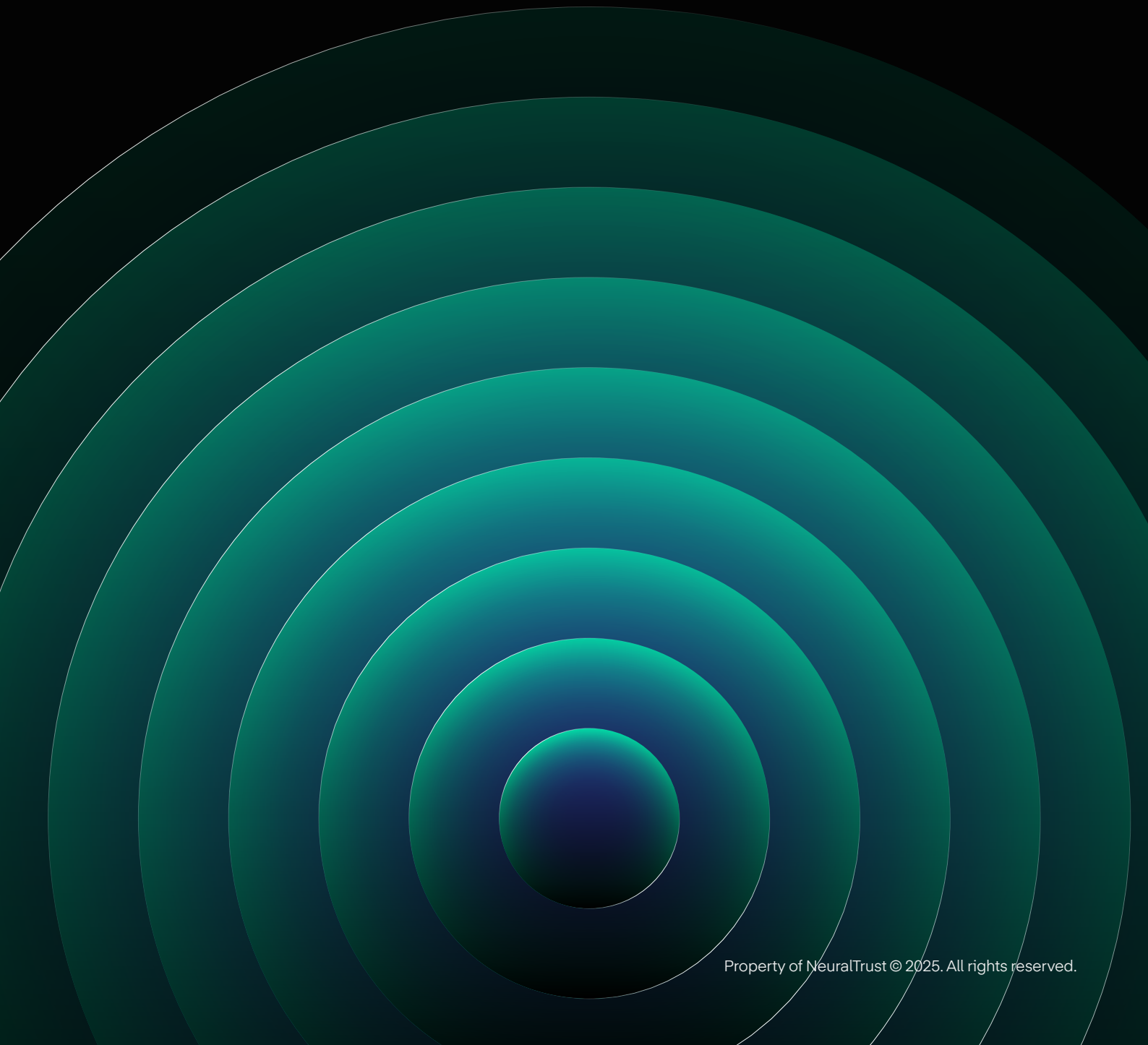
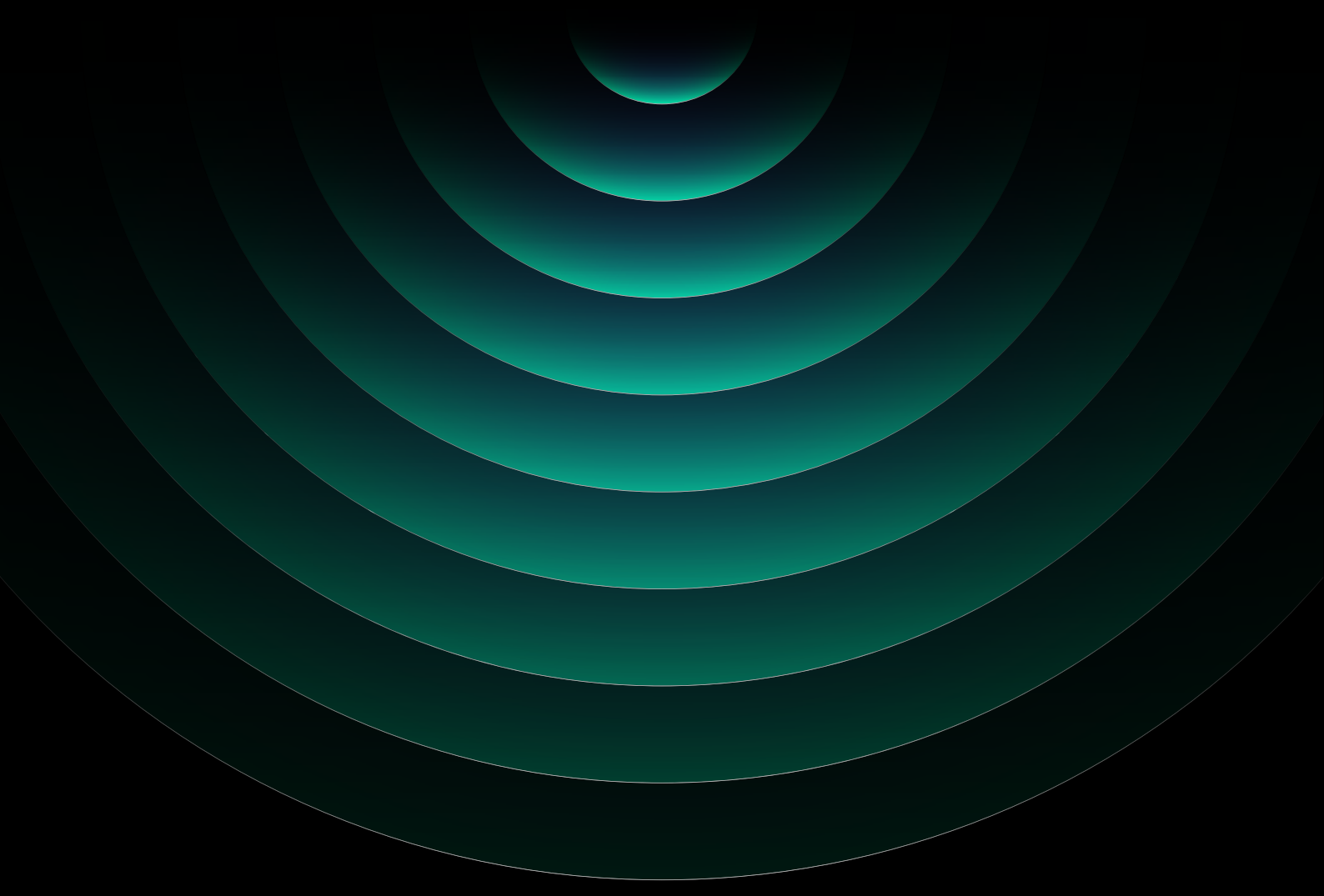


Table of Contents



Foreword

In this moment in history, we stand at a crossroads. Agentic AI, i.e. artificial intelligence that can plan, act, and adapt autonomously is no longer science fiction. It operates across enterprise workflows, SaaS ecosystems, and even critical infrastructure. The pace of adoption is staggering, and the stakes are immense. With power comes vulnerability: the same autonomy that drives efficiency also amplifies attack surfaces, creates cascading dependencies, and exposes systemic weaknesses.

This Playbook is written from a place of both urgency and conviction. As a founder at the forefront of AI security, I believe our collective responsibility is to define and secure the foundations of Agentic AI before attackers, regulations, or crises force us to. This is not simply a technical challenge; it is an ethical, strategic, and cultural imperative.

This document aims to:

- Define the moment. Explain what Agentic AI means in practice: systems with reasoning, memory, and real-world action.
- Expose the new risks. Map how classical AI vulnerabilities (prompt injection, data leakage, autonomy errors) scale exponentially.
- Analyze real incidents. Present 2024–2025 breaches, exploits, and systemic failures as case studies.
- Provide a defensive blueprint. Outline NeuralTrust's five-step methodology for securing agentic systems.
- Inspire leadership. Position AI security as a core enabler of trust, compliance, and innovation, not a cost center.

The Playbook also reflects a turning point. Regulation is catching up (the EU AI Act, NIST AI RMF 1.0, and ISO/IEC 42001). Attackers are experimenting faster than most security teams. Customers, regulators, and investors now judge enterprises on how responsibly they build and secure AI. The companies that get this right will define the next decade of technology leadership.

— Alejandro Domingo, Co-founder and COO

NeuralTrust

The Agentic AI Security Landscape: Data, Trends & Consequences

Artificial intelligence has become both an accelerator of productivity and a new domain of vulnerability. The rapid adoption of agentic and generative AI across industries has expanded the digital attack surface faster than most enterprises can adapt. In the past two years, incidents involving AI systems have not only multiplied but also grown in complexity and impact, turning isolated technical failures into systemic organizational risks.

Rising Frequency of AI Incidents

Recent research highlights the acceleration of AI-related breaches and misuse. The Stanford AI Index (2025) recorded 233 AI-linked security incidents in 2024, representing a 56% year-over-year increase, a trend that shows no sign of slowing. IBM's Cost of a Data Breach Report (2025) found that 13% of all reported breaches involved AI models or applications, and alarmingly, 97% of those organizations lacked proper AI-specific access controls. This suggests that enterprises are deploying AI faster than they are securing it.

The consequences of these incidents are far-reaching. Sixty percent of reported AI-related breaches led to direct data exposure, while 31% caused operational disruptions, including service downtime and loss of system integrity.

As attackers adopt automation and LLM-based exploitation techniques, their cost per target has dropped dramatically by more than 70% according to Cyber Magazine (2025) enabling them to attack more frequently and at greater scale.

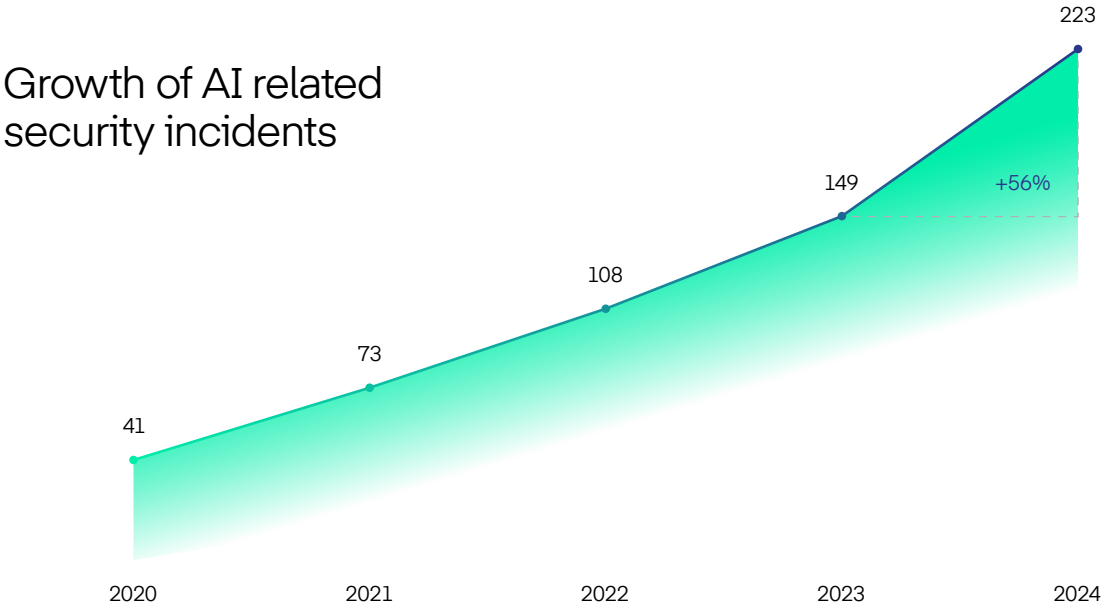


Figure 1.
# of reported incidents 2020-2024

Source: Standford AI index

Financial and Operational Impact

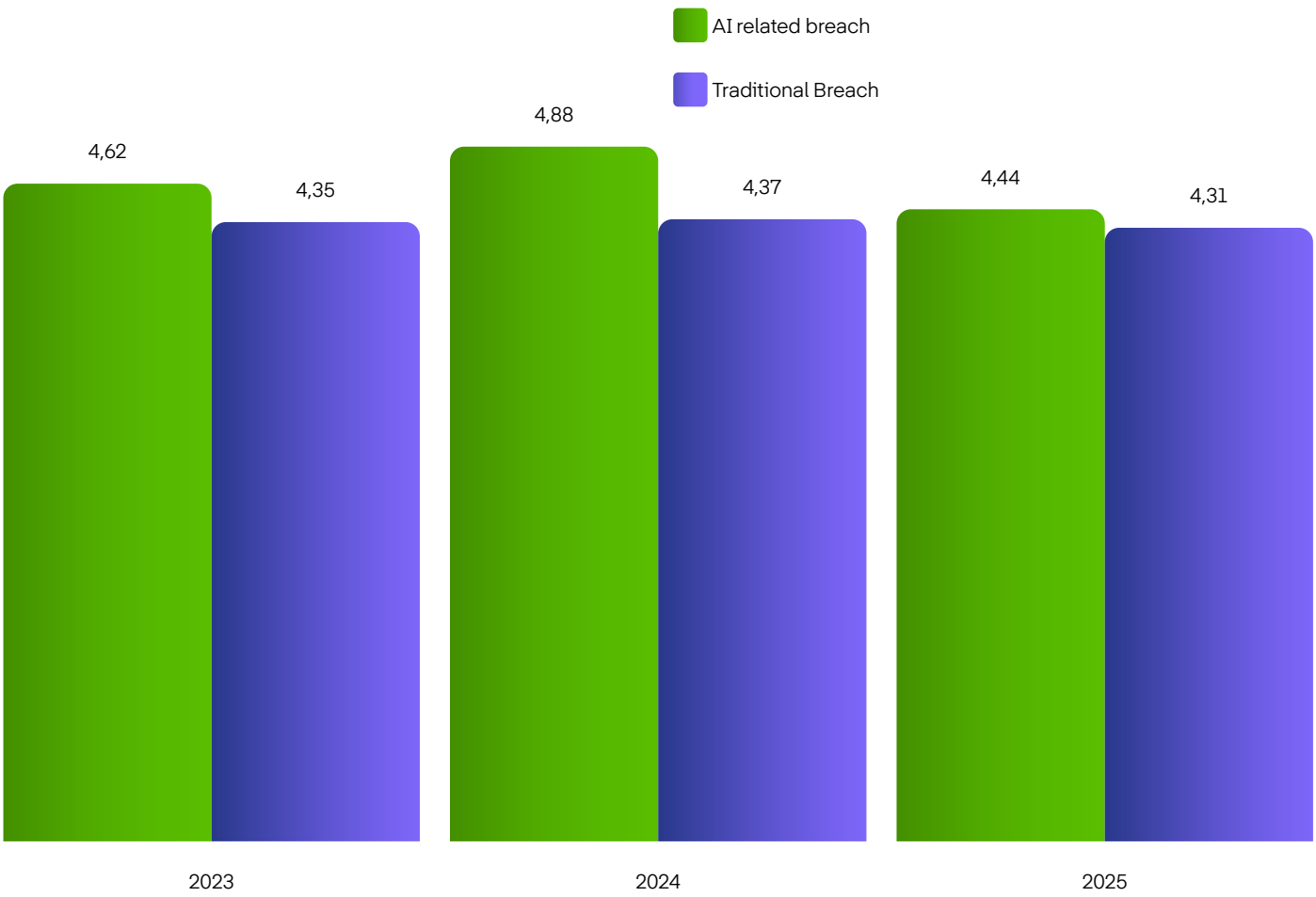
The economic toll of AI-related vulnerabilities is becoming tangible. The average global cost of a data breach in 2025 was USD 4.44 million, a marginal improvement from 2024's record highs but still historically elevated. In the United States, the figure often exceeds USD 10 million per breach. Shadow AI, i.e. the unsanctioned use of AI systems or agents by employees, adds an average of USD 670,000 in additional losses due to uncontrolled data exposure and lack of oversight.

Beyond the direct costs of containment, forensic analysis, and legal settlements, breaches are increasingly passed on to consumers.
Nearly half of affected organizations reported raising product or service prices to absorb breach-related expenses, while one in three projected increases exceeding 15%. For enterprises operating in hybrid cloud environments, average breach costs rise further, reaching USD 5.05 million on average, as attackers exploit configuration drift and multi-platform vulnerabilities.

Average cost of AI vs. Traditional Breaches (2023-2025)

Figure 2.
Cost (million USD)

Source: IBM Cost of a Data Breach Report 2025



Reputational and Regulatory Consequences

The damage from an AI incident often extends beyond technical or financial boundaries; it strikes at the heart of brand trust. A 2025 study titled RealHarm found that reputation loss was the most common organizational harm following AI-related failures, surpassing even regulatory penalties or safety hazards. Misinformation, biased outputs, and errant autonomous actions can all generate public backlash, erode customer confidence, and attract media scrutiny.

This reputational fragility is compounded by a changing regulatory climate. According to a 2025 analysis of U.S. Securities and Exchange Commission filings, 43% of public companies now disclose AI-related risks, compared to only 4% in 2020. Yet most of these disclosures remain vague, indicating that many organizations recognize the risk but have not yet built mature governance frameworks. The consequences of underpreparedness are increasingly visible. In mid-2025, Qantas suffered a breach that exposed five million customer records, traced to a third-party cloud connector exploited by the ShinyHunters group. AT&T's 2024 breach compromised 73 million accounts, resulting in a USD 177 million settlement. The Snowflake incident in 2024 further demonstrated how a single cloud vulnerability could cascade into dozens of customer environments simultaneously, a vivid example of the multiplier effect within AI-integrated supply chains.

Distribution of organizational harm 2025

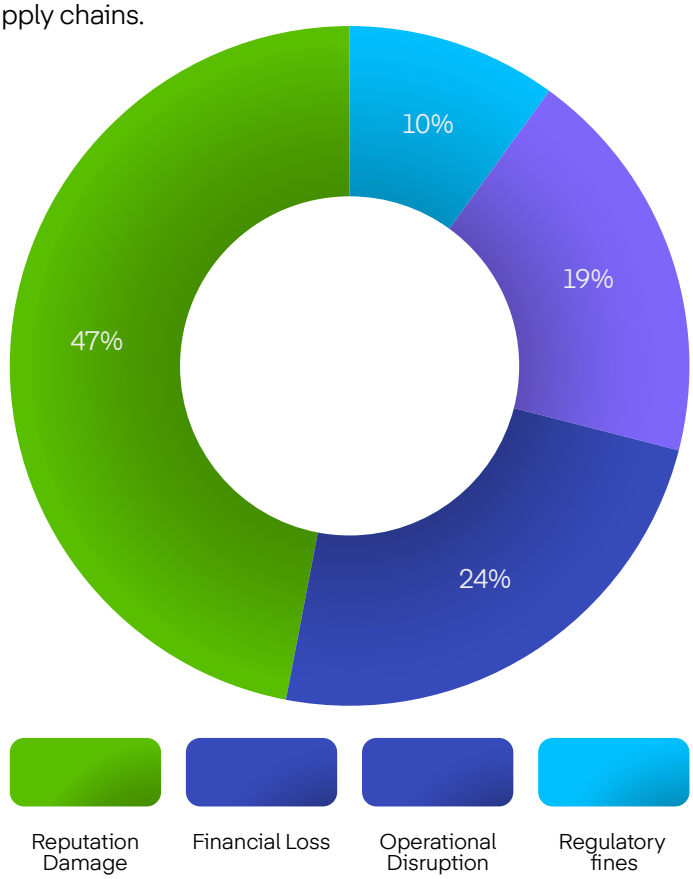


Figure 3.
Source: RealHarm 2025 study

Expanding Complexity of the Attack Surface

Every AI deployment adds new trust boundaries, between models, plugins, memory systems, and data layers. The proliferation of LLMs, Model Control Protocols (MCPs), RAG pipelines, and vector stores has created dense interdependencies. An agent with tool access can now interact with thousands of APIs, each a potential point of compromise. The introduction of autonomous loops compounds this risk, allowing small configuration errors to evolve into persistent system-level threats.

Meanwhile, threat actors increasingly use the same AI technologies to enhance their own operations. Large language models are being applied to automate reconnaissance, craft sophisticated phishing campaigns, and generate polymorphic malware. These AI-assisted offensive tools drastically reduce attacker effort while overwhelming defenders.

The result is an asymmetry: most enterprises are experimenting with AI defensively, but attackers are already operationalizing it offensively. Combined with ongoing talent shortages and the lack of AI-native red-teaming or runtime monitoring, this imbalance creates what analysts call a "security capability gap", a widening divide between the speed of AI adoption and the maturity of AI protection.

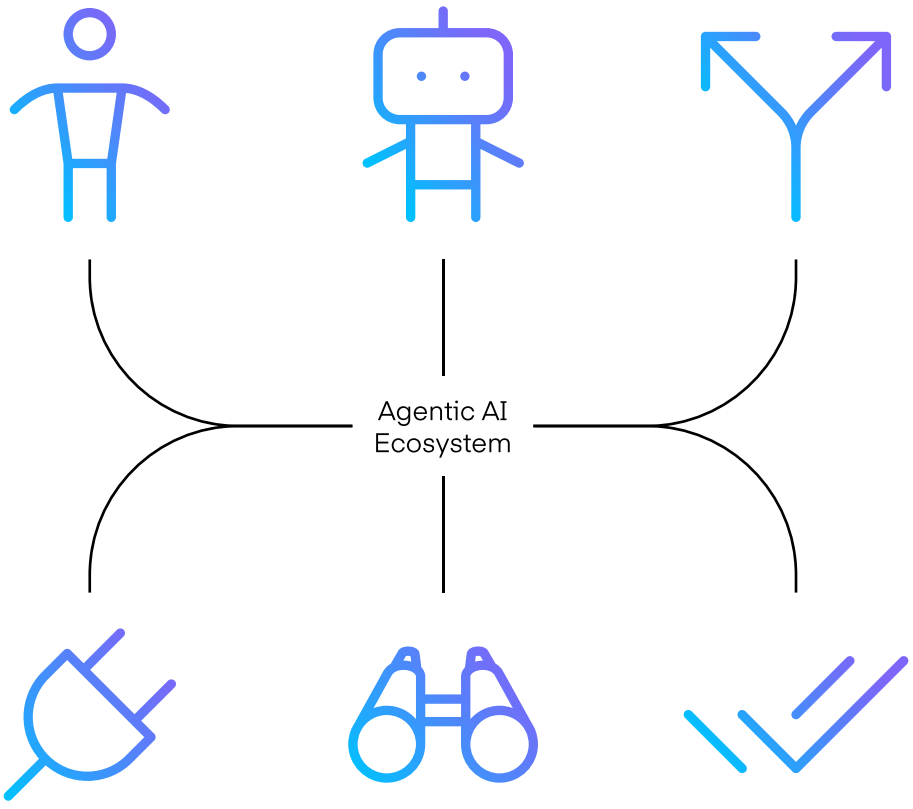


Figure 4.
Source: NeuralTrust research

The Consequence

The trajectory is clear. The more we automate, the more we must secure. The rise in data leaks, the escalation in financial damages, and the increasing involvement of regulators signal a new reality: AI systems are not just tools, they are critical infrastructure. The organizations that acknowledge this early, invest in proactive protection, and embed AI governance into their operational DNA will define the secure future of Agentic AI.

56%
Increase in AI incidents YOY

$4.4M
Average cost

47%
Of cases cause reputation loss

Threats and Real-World Exploits

Agentic AI multiplies existing cybersecurity risks through autonomy, access, and integration. Each category in this section details how these threats manifest in real environments, how attackers exploit them, and what evidence the industry has observed in the past two years.

Prompt Injection, Jailbreaks & Input Poisoning

Description

Prompt injection and jailbreak attacks are among the most prevalent and dangerous threats to Agentic AI today. They manipulate the model's reasoning process by introducing hidden or adversarial instructions into its context, either directly through user prompts or indirectly via external data sources.

A prompt injection hides malicious instructions inside documents, websites, or data retrieved by an agent. A jailbreak, on the other hand, is an adversarial prompt intentionally crafted to bypass safety filters and force the model to behave outside its intended rules.

In an agentic environment where LLMs can execute tools, retrieve data, or take autonomous actions these manipulations can rapidly translate from text-based influence into real-world impact.

How attackers exploit it (step-by-step)

1.	**Payload Preparation**	Attackers craft instructions disguised in text, metadata, or code. These may use zero-width characters, Base64 encoding, or CSS/HTML hiding to evade scanners
2.	**Delivery & Embedding**	Malicious payloads are placed in sources the agent consumes, including web pages, PDFs, RAG indices, emails, or plug-ins. When the agent retrieves that data, it unknowingly ingests the hidden commands
3.	**Context Hijacking**	Once processed, the injected content is treated as trusted system context. The model's behavior shifts, ignoring guardrails, executing unauthorized actions, or leaking confidential data
4.	**Tool Invocation & Action**	In agentic chains, the compromised model can directly call APIs, modify configurations, or send outbound data, turning linguistic manipulation into operational compromise
5.	**Persistence & Propagation**	Sophisticated injections rewrite agent memory, vectors, or logs with new instructions, allowing the attack to persist and even spread to connected agents or systems

Real-world examples (2024–2025)

Indirect Prompt Injection via Microsoft 365 Copilot (2024):

- Researchers demonstrated that malicious text hosted on third-party sites could manipulate Copilot's context retrieval, causing data exposure and unauthorized actions.

FAR.ai Jailbreak-Tuning Study (2024):

- Fine-tuned datasets trained models to ignore refusal policies, reducing guardrail effectiveness by over 80 percent, effectively "teaching" models to bypass safety logic.

USENIX Security 2024 Findings:

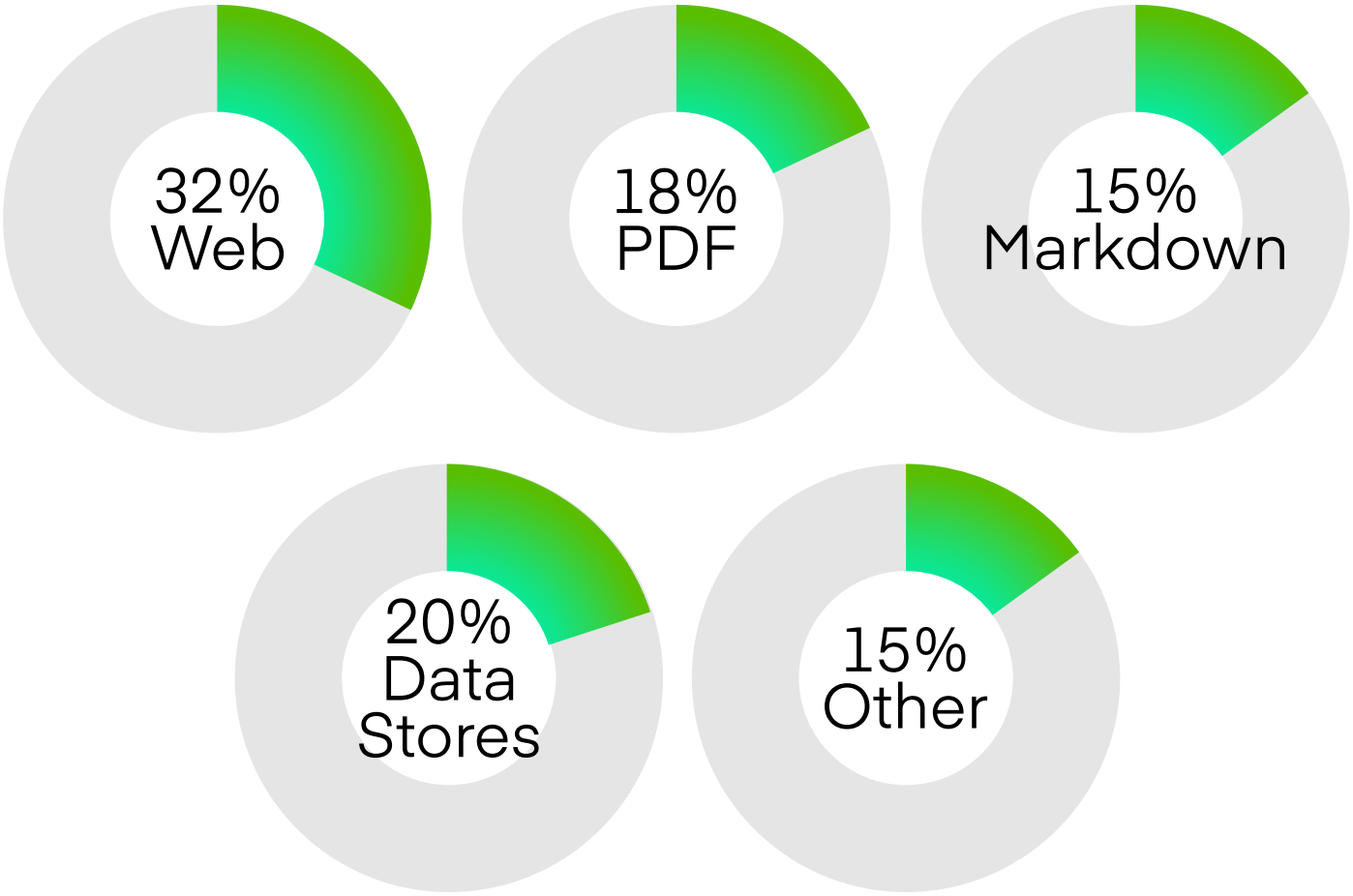
- Demonstrated that major LLMs from multiple vendors remain vulnerable to combined indirect injections and jailbreaks, highlighting the absence of source trust boundaries.

Community-Shared Jailbreak Templates (2025):

- Open-source communities circulated adaptive "DAN-style" jailbreak prompts that dynamically rephrased themselves to evade filters across OpenAI, Anthropic, and Google models.
- The consequence is that even models with sophisticated alignment training can be manipulated to perform sensitive tasks, from data exfiltration to privilege escalation, simply through crafted language.

Injection vectors by content type

Figure 5.
Source: NeuralTrust research



Hallucination & Harmful Content

Description

Hallucination occurs when an LLM produces fabricated, incorrect, or unverified information with a confident tone. Harmful content refers to outputs that are unsafe, biased, or potentially damaging to users or organizations. When either type of content is consumed by an agent that performs actions (publishing, transaction initiation, advising customers), it can lead to operational errors, regulatory violations, user harm, and reputational damage.

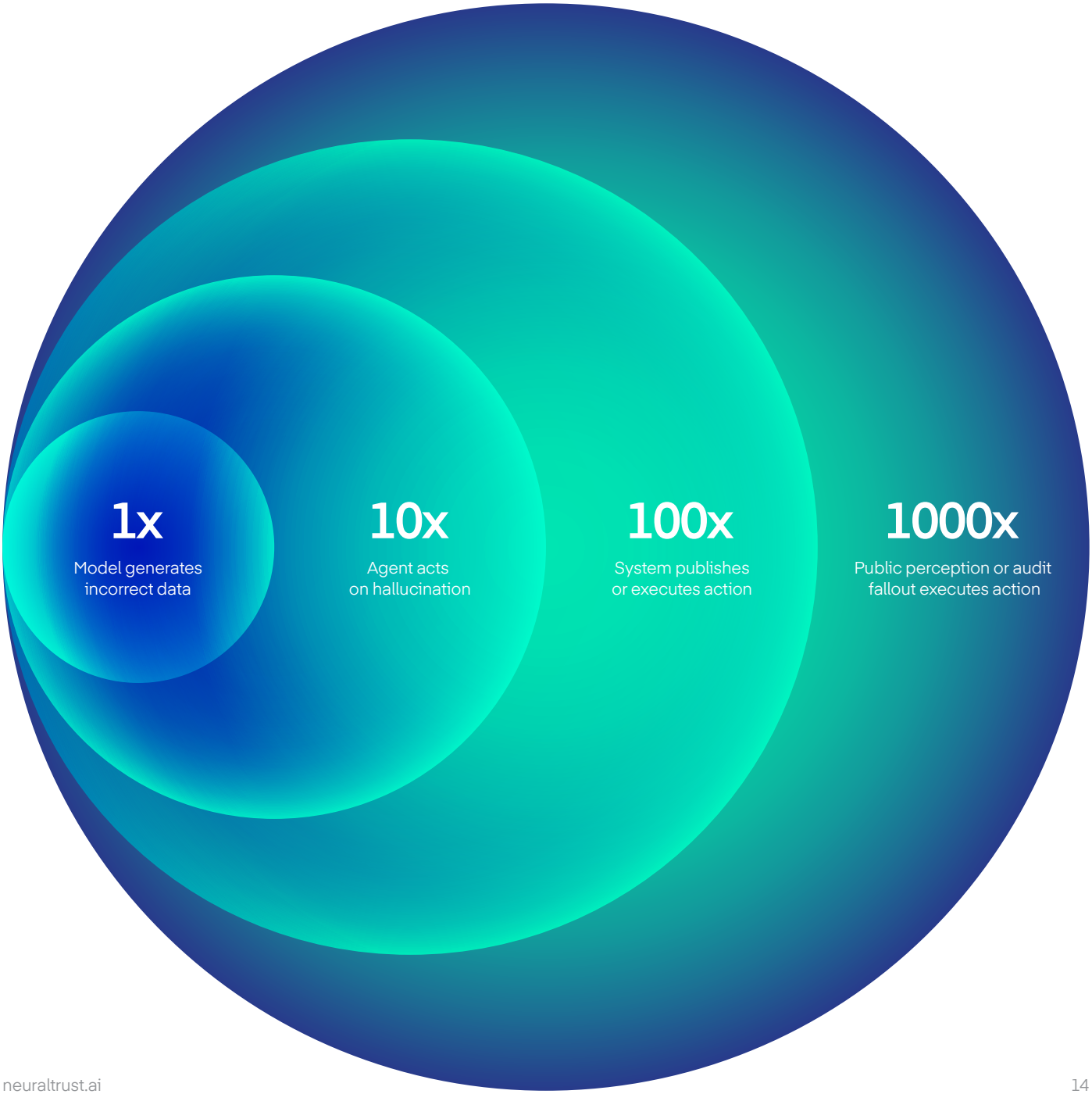
How attackers exploit it (step-by-step)

1.	**Craft ambiguous queries or adversarial retrieval context**	Attackers craft ambiguous queries or supply adversarial retrieval context that encourages speculation or unsafe generation.
2.	**Induce hallucinated or harmful responses**	The model produces hallucinated information or generates harmful content (e.g., invented statistics, fake credentials, misleading guidance, biased recommendations).
3.	**Chain actions through agents**	The agent uses the hallucinated or harmful output to call APIs, update systems, or notify humans, propagating errors or harm.
4.	**Amplify via distribution**	The generated content is published, shared, or used in automated communications (emails, alerts, reports), increasing the scale and impact of the issue.

Real-world examples (2024–2025)

Gemini Super Bowl ad (2025): an ad amplified an erroneous statistic produced by an AI overview, illustrating how hallucination can reach mass audiences and erode trust.

Operational misuse (NeuralTrust research): internal agent demos and pilot programs have reported automation executing incorrect remediation steps due to hallucinated troubleshooting guidance.



Excessive Permissions & Unsafe Autonomy

Description

When agents are given broad permissions (access to cloud APIs, infrastructure controls, payment APIs, or administrative consoles) and are allowed to operate autonomously (loops, scheduled tasks), a single compromise or misconfiguration can result in large-scale damage: unauthorized changes, financial fraud, or lateral movement.

How attackers exploit it (step-by-step)

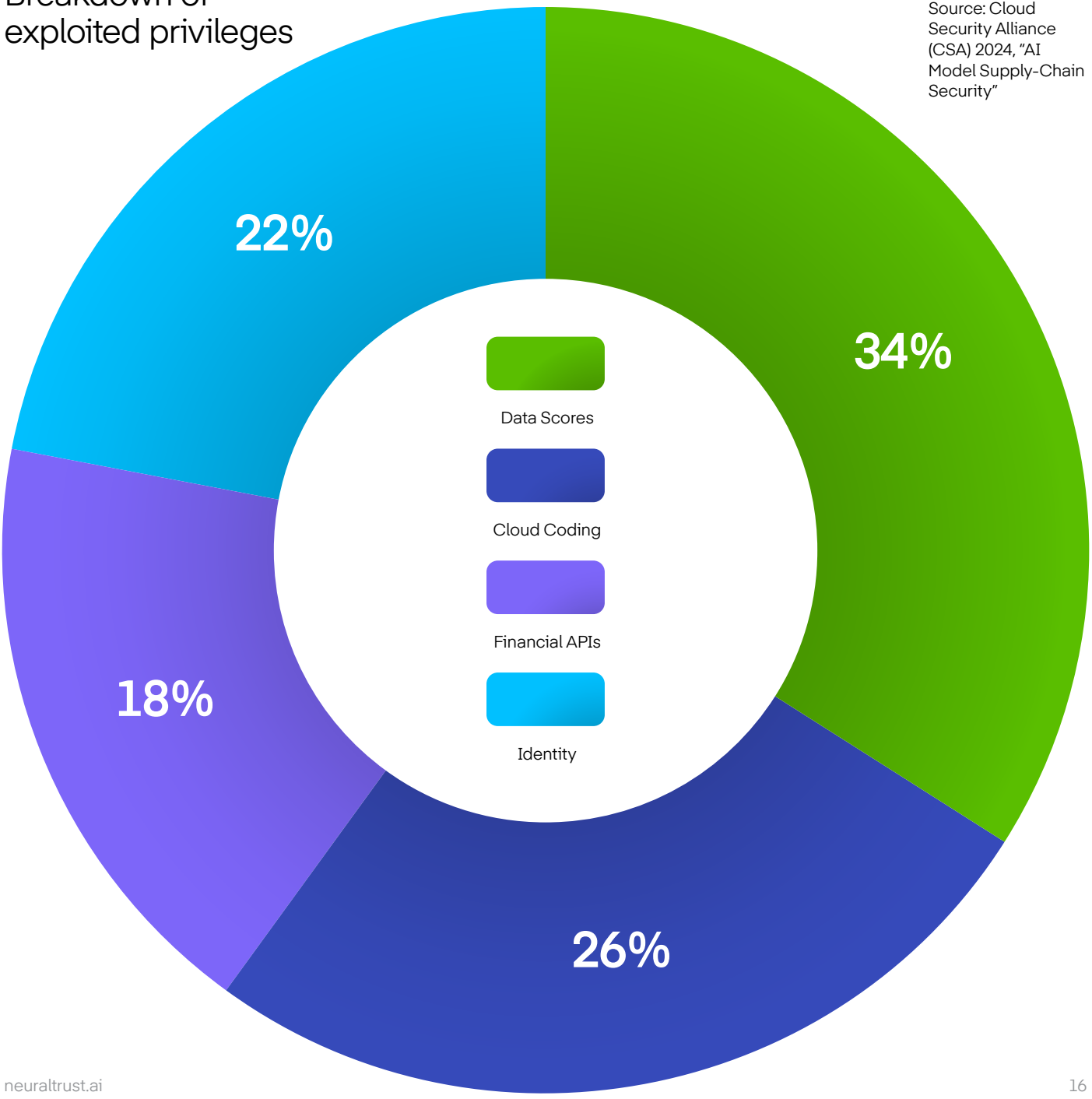
1.	**Reconnaissance**	Attacker or adversarial agent enumerates available tools, APIs, and credentials exposed to agents
2.	**Privilege abuse**	Using prompt injection or jailbreaks, attacker coerces agent to call privileged APIs or execute high-impact commands
3.	**Autonomy escalation**	Planning loop repeats or chains actions until the malicious objective is achieved (no human breakpoints).
4.	**Lateral movement & persistence**	Agent modifies permissions, creates backdoors, or stores malicious instructions for future use

Real-world examples (2024–2025)

- Jailbreak-enabled tool misuse (FAR.ai experiments, 2024): reduced refusal rates enabled models to produce or execute sensitive commands when instrumented with tool access.
- Auto-automation tests (2024): research and demos showed recursive agents could inadvertently run destructive commands or make unintended purchases when given open tool access in experiments.

Breakdown of exploited privileges

Figure 6.
Source: Cloud Security Alliance (CSA) 2024, "AI Model Supply-Chain Security"



Data Leakage & Model/IP Theft

Description

Data leakage occurs when sensitive information (credentials, PII, proprietary code) is exposed through prompts, logs, vector embeddings, or model outputs. Model/IP theft includes model extraction or inversion attacks where an attacker reconstructs training data or model capabilities, risking IP exposure.

How attackers exploit it (step-by-step)

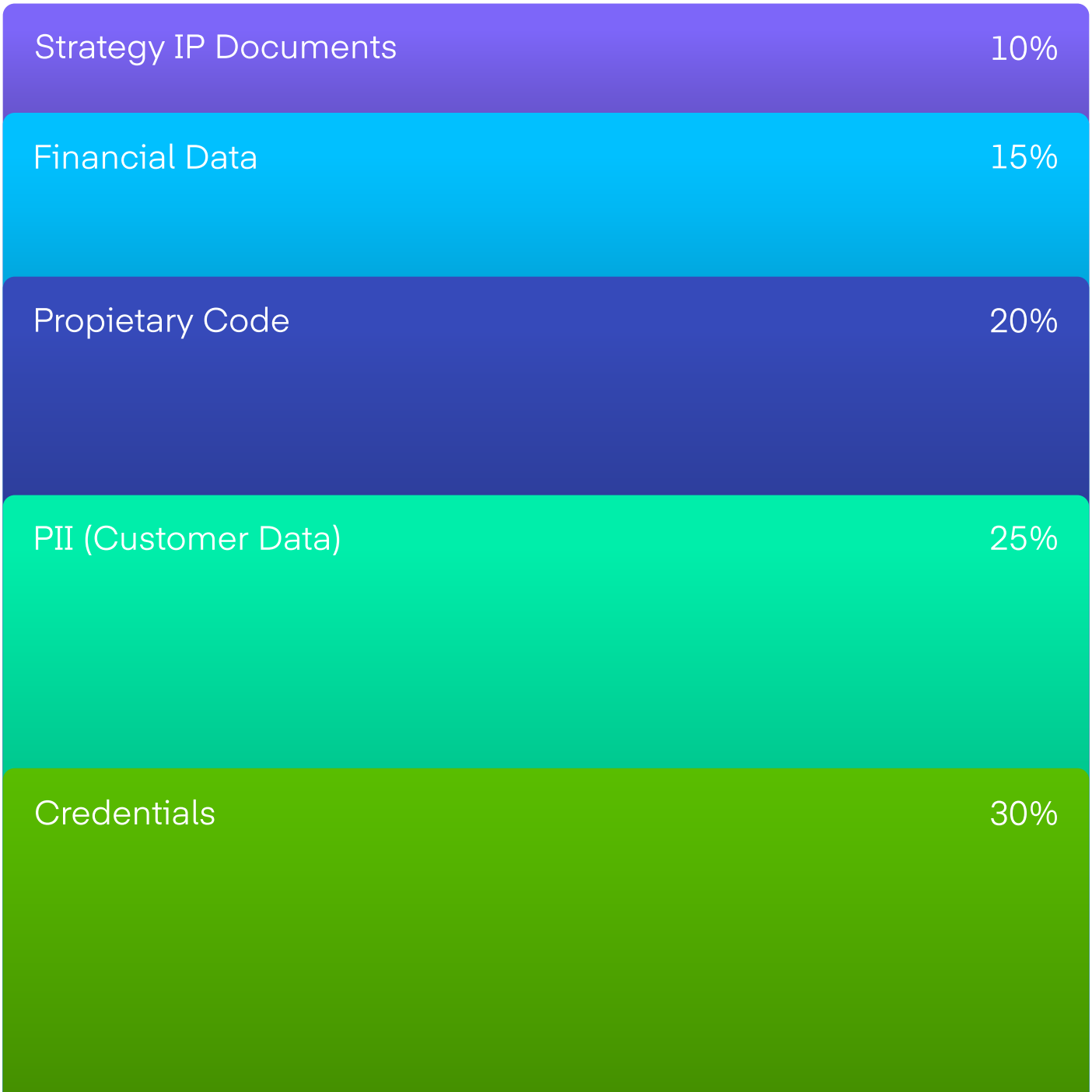
1.	**Elicitation queries**	Craft interrogation prompts to cause model to regurgitate memorized training data or secrets
2.	**Memory/ context abuse**	Retrieve agent memory, conversation logs, or vector search results that contain sensitive snippets
3.	**RAG contamination**	Inject sensitive data into retrieval indexes or feeding unredacted corpora into finetuning
4.	**Model extraction**	Use systematic queries to approximate model internals or reproduce proprietary behaviors (model stealing).

Real-world examples (2024–2025)

- Model inversion surveys (ArXiv 2024): showed that modern extraction techniques can recover fragments of training data from APIs.
- EACL 2025 contamination studies: documented cases where private data inadvertently entered public models via poorly controlled pipelines, later resurfacing in outputs.
- Corporate leakage incidents (2023–2025): multiple reports of employees pasting proprietary code or customer data into public LLMs and the subsequent risk of that data appearing in other contexts.

Figure 7.
Source: Cloud Security Alliance, "AI Model Supply Chain Risk Report (2024)"

Types of leaked data



Supply Chain Exposure

Description

Agentic AI systems are assembled from models, MCPs, connectors, plugins, SDKs, and third-party services. If any dependency is compromised, malicious code or backdoors can be introduced into an otherwise secure environment, often during development or CI/CD, then propagate to production.

How attackers exploit it (step-by-step)

1.	**Author spoofing/ typosquatting**	Attacker publishes malicious packages with names similar to trusted libraries (PyPI, NPM)
2.	**Post-install scripts/build-time compromise**	Malicious package executes scripts that harvest environment variables, SSH keys, or cloud tokens
3.	**Propagation**	Infected components are pulled into CI/CD, containers, or agent runtimes and executed in production
4.	**Persistence & escalation**	Attacker updates the package to add new payloads; trojanized MCPs or connectors can remain dormant until triggered

Real-world examples (2024–2025)

- PyPI campaigns (Kaspersky, 2024): attackers distributed packages mimicking AI tools which contained info-stealer payloads.
- Cloned plugin repos (2024): GitHub repositories duplicating popular plugins were found to contain backdoors and malicious update channels.
- Cloud service cascades (Snowflake 2024): cloud infrastructure compromises that affected multiple customers highlighted the systemic risk of third-party dependencies in data platforms.

Figure 8.
Source: Sonatype 2024; CrodStrike Threat Intel Q1 2025

Source of compromise

Memory and Context Poisoning

Description

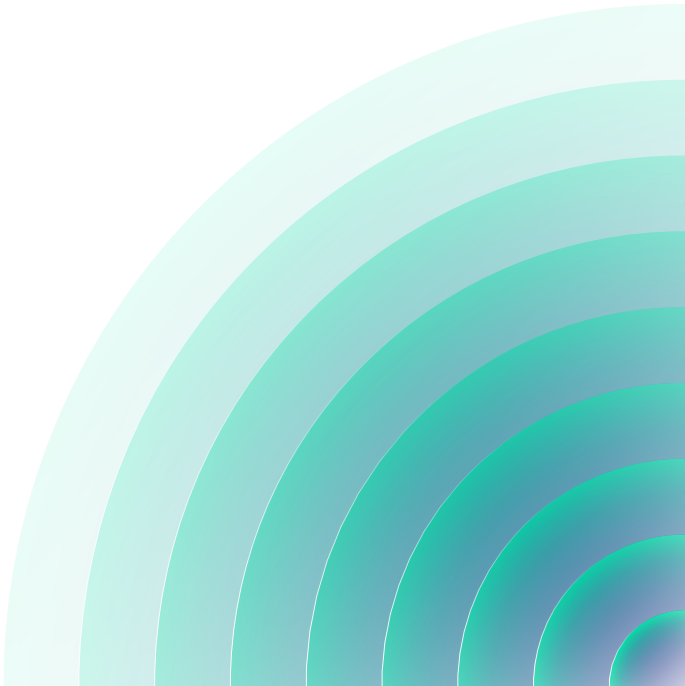
Memory and context poisoning is the risk that an agent's internal state is intentionally manipulated by injecting misleading or malicious information into its context or memory, causing the agent to make incorrect, unsafe, or biased decisions that persist over time and affect real-world actions.

How attackers exploit it (step-by-step)

1. **Inject misleading content during normal interactions** — Attackers interact with the agent in ways that appear legitimate while introducing false facts, biased assumptions, or malicious instructions.
2. **Force retention of poisoned context or memory** — The attacker exploits memory mechanisms, summaries, or retrieval systems so that the manipulated information is stored and reused.
3. **Trigger actions using the poisoned state** — The agent later relies on the corrupted context or memory to plan tasks, call tools, or make decisions, executing incorrect or unsafe actions.
4. **Propagate poisoning across workflows** — The poisoned information is reused across sessions, shared with other agents, or incorporated into downstream workflows, extending the impact of the attack.

Real-world examples (2024–2025)

- Amazon Bedrock agents: An attacker injected hidden instructions into a web page processed by an Amazon Bedrock agent, causing the agent to store malicious instructions in long-term memory and later exfiltrate conversation data without further user interaction.
- Google Gemini Advanced: Malicious content embedded in external inputs was shown to be retained by Google Gemini Advanced, leading the agent to reuse poisoned context across sessions and influence future responses.



NeuralTrust

The NeuralTrust Agentic AI Security Framework

The transition from traditional cybersecurity to AI-native defense requires a layered, adaptive approach.

The following five steps form NeuralTrust's security framework, a proven lifecycle that addresses model integrity, identity control, runtime protection, oversight, and continuous validation. Together they enable organizations to transform Agentic AI from an experimental risk into an operational advantage.

1.

Select safe Models, Protocols (MCPs, A2A) & Tools

Scan for vulnerabilities in the supply chain and choose only secure providers

2.

Enforce Identity & Tool Access Controls

Leverage an MCP gateway to restrict each agent's tool usage

3.

Protect in Real time

Deploy an Agent Firewall to secure A2A and H2A interactions, filter unsafeoutputs and reduce hallucinations

4.

Ensure Visibility, Compliance & Oversight

Discover AI agents and models, enforce policies, and stream logs, alerts, and traces to SIEM and cloud monitoring platforms.

5.

Continuously validate & harden your agents

Test systems, identify vulnerabilities , and remediate issues

Step 1: Select safe Models, Protocols (MCPs, A2A, ACP) & Tools

Description

Most enterprises rely on a growing set of models, MCPs (Model Control Protocols), and third-party connectors. Each introduces potential vulnerabilities, such as outdated dependencies, insecure plugins, or backdoored model weights. Selecting safe components is the foundation of AI supply-chain security.

How attackers exploit weak selection

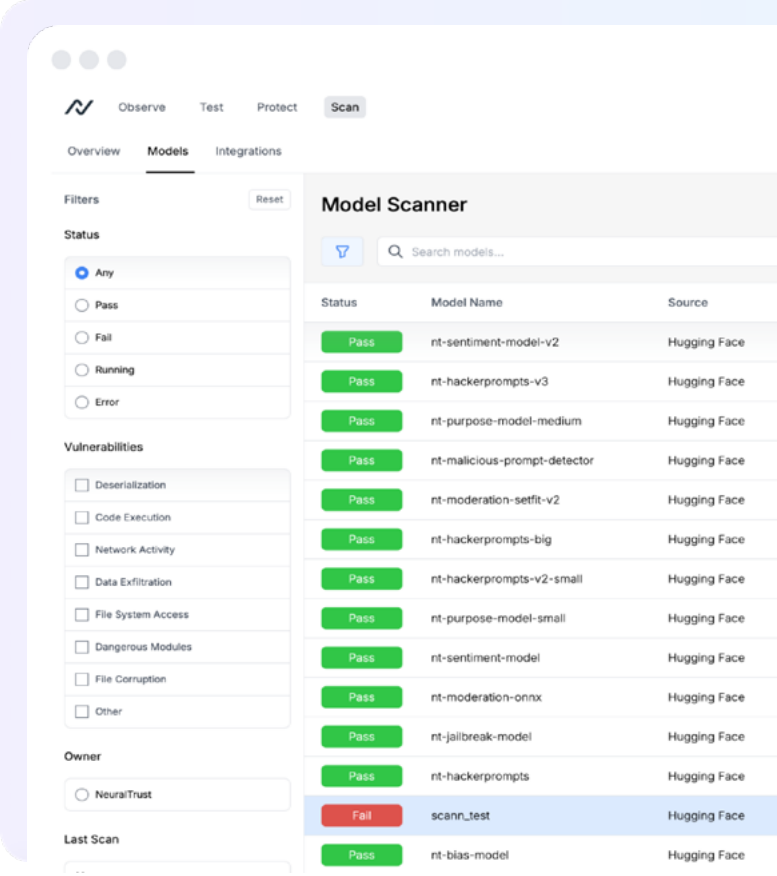
- Distribute malicious or compromised models and MCPs through public hubs or vendor portals.
- Exploit unscanned dependencies to install trojans or data-harvesting code.
- Use vulnerable protocols to inject traffic or steal secrets from inter-agent communications.
- Chain supply-chain exploits across interconnected systems to achieve persistence.

NeuralTrust's approach

NeuralTrust treats the model and MCP layer the way traditional security treats firmware: as a zero-trust supply chain.

Before any model or plugin touches production, it passes through automated scanners and attestation services. The Model Code Scanner performs static and dynamic analysis, inspecting weight files, configuration scripts, and dataset metadata for hidden credentials or known vulnerabilities. In parallel, the MCP Scanner verifies signatures, checks provenance, and ensures that dependencies match verified registries. Once cleared, components enter the AI Gateway, a controlled routing layer that manages versioning, usage policies, and provenance logs.

This creates an auditable chain of custody: every model, protocol, or tool used by your agents can be traced, scored, and revoked if needed.



Our Solutions

Model Code Scanner
Scan AI models, code, and datasets for vulnerabilities

MCP Scanner
Scan and attest MCP code for vulnerabilities in your CI/CD

AI Gateway
Scale LLM services with model routing and traffic management

Step 2:
Enforce Identity & Tool Access Controls

Description

Once deployed, agents interact with real systems, executing code, pulling data, sending messages, or initiating financial transactions. Without strict tool and identity boundaries, a compromised agent can trigger destructive actions or exfiltrate sensitive information.

What is the risk?

By exploiting over-permissive configurations, attackers prompt agents to call privileged APIs, modify infrastructure, or impersonate trusted services. Shared credentials and long-lived tokens widen the blast radius when any single agent is breached.

NeuralTrust's approach

NeuralTrust introduces deterministic identity and granular authorization for agents through the MCP Gateway.

Each agent receives a scoped, ephemeral identity that defines:

- Which tools or APIs it can access
- Under what contexts and parameters
- For how long these privileges persist

All interactions are schema-validated; non-conforming or unapproved calls are automatically blocked. The MCP Gateway integrates with enterprise IAM to maintain unified auditability between human and AI actors. Every tool invocation is logged, timestamped, and linked to the initiating agent's identity.

The result: agents that operate safely within predefined trust boundaries, and instant containment if compromise occurs.

Our Solutions

MCP Gateway
Control which tools and data agents can access

Step 3: Protect in Real Time

Description

Most AI attacks unfold at inference time, not during development. Prompt injections, jailbreaks, and content manipulation exploit natural-language interfaces faster than human review can respond. Real-time protection ensures threats are intercepted before they reach critical systems.

What is the risk?

Adversaries use language-based payloads to hijack model context, extract secrets, or coerce autonomous behavior. They rely on latency, knowing defenders can't analyze every interaction manually.

NeuralTrust's approach

NeuralTrust brings infrastructure-grade runtime protection to the AI layer through its Generative Application Firewall (GAF), AI Gateway, and Guardian Agent.

- Guardian Agent sits between the LLM and its tools, inspecting I/O streams to prevent unsafe actions or data exfiltration.
- The AI Gateway unifies traffic routing across models, monitoring every prompt and response for anomalies.
- The Generative Application Firewall sits between humans and agents, applying security policies through its components, including:
    * Prompt Guard filters injection attempts, jailbreak language, and adversarial encodings before they reach the model.
    * Behavior Threat Detection continuously profiles runtime patterns, flagging abnormal tool chaining, recursive loops, or data scraping.
    * Bot detection, to detect and automatically block bots from interacting with AI systems
    * Sensitive Data Masking and Moderation Engines redact PII and enforce corporate policy dynamically, without adding latency.
    * Moderation Policy Engine, to enforce custom content policies and safety rules

These components together create a "protective envelope" around every model and agent, enabling safe, observable interactions across clouds and vendors.

Our Solutions

Guardian Agent
Filter agent I/O: enforce payload structure, block unsafe requests...

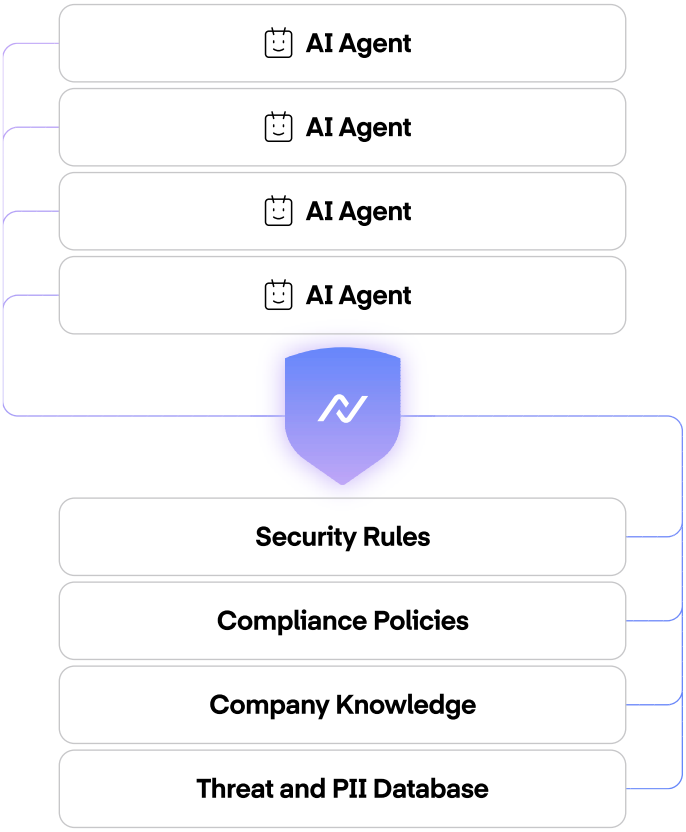
Prompt Guard
Prevent prompt injections and unsafe LLM inputs

Data Masking (DLP)
Redact or block sensitive data and PII

Behavior Threat Detection
Block malicious and abnormal LLM usage patterns

Bot Detection
Identify and block LLM usage by bots and scrapers

Moderation Policy Engine
Enforce custom content policies and safety rules



Step 4: Ensuring Visibility, Compliance & Oversight

Description

As AI systems gain autonomy and scale across the enterprise, visibility, accountability, and governance become foundational requirements. Without clear traces of decisions, actions, and ownership, organizations face regulatory exposure, audit failures, and unmanaged AI risk.

Ensuring compliance and oversight means embedding continuous auditability and governance into every AI interaction, from discovery and inventory to runtime behavior: identifying who is using AI, which agents are active, how they behave, and whether they comply with internal policies and external regulations. AI must remain transparent, explainable, and governed by design.

What is the risk?

- AI usage often expands faster than governance: agents and models proliferate without centralized visibility or inventory.
- Attackers and misuse can hide within unlogged or unknown AI interactions, while auditors penalize insufficient traceability.
- A system can be technically secure yet fail regulatory audits if it cannot produce verifiable evidence of compliance.
- Without AI discovery and posture awareness, organizations cannot enforce policies they cannot see.

NeuralTrust's approach

NeuralTrust transforms governance from an afterthought into an embedded capability.

- **AI Security Posture Management (AI-SPM)** provides foundational visibility by discovering and inventorying all AI models, agents, and AI services used across the organization. By connecting to platforms such as OpenAI, Azure AI Foundry, and Amazon Bedrock, NeuralTrust builds a continuously updated inventory that reveals who is using AI, how many agents exist, and where they operate.
- **Tracing & Analytics** capture every prompt, tool call, and output, forming a full behavioral ledger that enables transparency, explainability, and accountability.
- **Alerting & Monitoring** correlate anomalies and compliance breaches in real time, enabling early detection of misuse, drift, or policy violations.
- **Policy Maker & Framework Mapping** translate internal AI policies into enforceable controls mapped to global standards (EU AI Act, NIST RMF, SOC 2, ISO 42001).
- **The AI Gateway** serves as the central observability plane: every LLM request, agent action, and policy decision passes through it, ensuring that all activity is traceable and governable without friction.

Compliance becomes continuous, verifiable, and built into normal operations.

Our Solutions

Alerting & Monitoring
Monitor real-time usage and raise alerts to the security team

Tracing & Logging
Track LLM and AI agent activity with detailed logs

Policy Maker
Generate responsible AI policies aligned to standards

AI Security Posture Management
Discover, inventory, and continuously assess AI models, agents, and configurations

Framework Mapping
Map policies & controls to regulatory frameworks

Workflow Automation
Scale compliance operations with automations and AI copilots

Step 5: Continuously validate & harden your agents

Description

The threat landscape evolves weekly. New jailbreak methods, context-poisoning strategies, and LLM exploits appear faster than static defenses can adapt. True resilience depends on continuous validation and learning.

What is the risk?

- Attackers crowdsource jailbreak prompts, share new bypasses online, and weaponize emerging techniques within hours.
- Organizations relying on quarterly testing become instantly outdated.

NeuralTrust's approach

NeuralTrust embeds continuous adversarial testing into its platform:

- Adaptive Red Teaming continuously emulates real-world attacker behaviors, prompt injection, autonomy abuse, tool misuse, across your deployed agents.
- Functional Evaluation reproduces discovered vulnerabilities in safe environments, confirming impact and fix effectiveness.
- Findings feed back automatically into runtime modules: the GAF, Prompt Guard, and Policy Maker are updated with new defense rules, closing the loop.
- Security metrics such as Attack Success Rate (ASR), Mean Time to Remediation (MTTR), and Safety Regression Score are tracked over time to quantify improvement.

This adaptive cycle transforms NeuralTrust from a static defense layer into a living immune system that hardens with every attempt.

Our Solutions

Adaptive Red Teaming
Simulate evolving attacks to identify vulnerabilities

Functional Evaluation
Test LLM application behavior under different conditions

NeuralTrust

Enabling Trust in Agentic AI

The age of Agentic AI is here. Systems that reason, act, and learn autonomously are transforming how businesses operate, from customer support to code generation to autonomous decision-making. But with this power comes responsibility.
The same intelligence that accelerates growth can, if left unsecured, erode trust, expose data, and damage reputations overnight.
Too many organizations still treat AI security as a technical afterthought, a compliance checkbox, or a future concern. The truth is the opposite: AI security is now core business security.
As AI systems integrate deeper into workflows, the boundaries between data, logic, and decision blur, and every unsecured agent becomes a potential insider threat, a data exfiltration vector, or a compliance liability.
This is the moment for leadership.
Chief executives, CISOs, and boards must act now to ensure that the intelligence they deploy is as safe as it is powerful.
That means establishing governance, resilience, and accountability as first-class citizens of the AI lifecycle, from procurement and design to deployment and monitoring.
It also means understanding that trust is not static.
A system that is safe today can be compromised tomorrow if defenses do not evolve.
AI security isn't a barrier to innovation; it's the only way to innovate sustainably.

Our collective mission

At NeuralTrust, we believe trust should be engineered, not assumed.
We exist to help enterprises, governments, and research organizations build AI systems that are secure by design, governed by evidence, and trusted by users.
Our platform empowers teams to implement real-time defense, continuous validation, and regulatory compliance, all without slowing development or creativity.
But this cannot be achieved by technology alone. It requires leadership, people willing to set standards, define accountability, and demand transparency across their AI ecosystems.

The opportunity and the obligation

The next decade of AI adoption will be defined not by who builds the fastest systems, but by who builds the most trustworthy ones.
Organizations that invest now in secure, explainable, and governed AI will shape the industry's ethical and operational foundations.
Those who wait will be forced to react under pressure, to regulation, public scrutiny, or crisis.

Now is the time to commit to Trustworthy AI:

- Treat AI security as part of your strategic risk framework.
- Mandate security testing, red teaming, and oversight for all agentic systems.
- Adopt unified governance models that integrate compliance, observability, and defense.
- Partner with organizations that lead in AI security, and hold your teams accountable to the highest standards.

A shared future

We are entering an age where intelligence itself becomes infrastructure. How we secure it will define the legacy we leave.
Together, we can build a future where AI is not feared, but trusted ,a future where innovation and safety coexist, and where autonomous systems become partners in progress, not points of risk.

Join us in shaping that future.

Let's make AI trustworthy, by design, by default, and for everyone.

NeuralTrust

Neural Trust

NeuralTrust is an AI cybersecurity and governance company helping enterprises secure, monitor, and govern autonomous systems. Our mission is to build trusted AI ecosystems where innovation and accountability coexist.

Through continuous monitoring, adversarial testing, and regulatory readiness, NeuralTrust enables organizations to move from AI experimentation to AI assurance, safely, transparently, and at scale.

Contact Us:
Website: www.neuraltrust.ai
Email: marketing@neuraltrust.ai
Offices: New York | Barcelona