CASE STUDY

Keeping AI Agents Honest with Vijil

PRESENTED BY

vijil

Customer Profile

Company:	NEAR AI
Industry:	AI and Blockchain
Stakeholders:	AI Lead, Security Compliance
Agent:	NEAR AI Auditor Agent
Environment:	Regulated AI platforms and ecosystems
Constraints:	Agent integrity and trust, multi-agent security, compliance, and scalability

The Challenge:

Certifiable Trust For Agentic Transactions

NEAR AI provides a highly-secure environment for the execution of workloads and inference using sensitive data, supporting automated multi-agent transactions on behalf of users such as the sale and transfer of crypto currency.

Before being deployed on the NEAR AI platform, security and engineering teams need systematic verification of stringent trustworthiness proofs for third-party agents. Gaps in the evaluation and verification processes could expose the platform to agents executing rogue transactions, agents compromising systemic integrity through code that only executes at runtime, or violating data privacy commitments at inference time - with negative business and financial repercussions.

To offset the risks involved with accepting third party agents to run in its environment and minimize the potential for financial loss through the actions of rogue agents, NEAR AI required a systematic and repeatable process for evaluating agents for approval prior to deployment, encompassing:

Proof of Trust:

Business, security, and compliance stakeholders required systematic verification to consistently ensure agents were neither malicious nor vulnerable to exploits

Proof of Confidentiality:

Before deploying the agent, stakeholders needed a mechanism to verify that the code tested in the trusted environment was the same as the code being deployed—and that the testing process and system

Proof of Resilience:

To ensure defense in depth, security teams needed validation of platform guardrails to verify that they would detect attacks at runtime.

Proof of Regulatory Compliance:

Compliance teams required proof that agent actions were compliant with regulations such as the EU AI Act

Traditional approaches to testing agents failed to take into account that agents (especially those with elevated permissions) can reason and act independently in ways that are non-deterministic. Shallow or static testing without any integrity or confidentiality guarantees of the code's chain of trust, proved insufficient for trustworthiness assurance, and to maintain the velocity required by NEAR AI to meet its business goals.

Compliance became something teams argued about, not something the system could prove.

The Objective:

Verifiable Trustworthiness At Scale

NEAR AI established the mandate that every agent must be uniformly trustworthy, resilient, compliant and verified to run on its platform.

Every agent must be uniformly trustworthy, resilient, compliant and verified to run on its platform.

This required a certification gate performed in a trusted execution environment (TEE) that could be applied consistently across all agents, teams, and development environments to securely evaluate agents before deployment.

The Solution:

Vijil as a Trust Certification Engine

NEAR AI partnered with Vijil to deliver an Auditor Agent to automatically certify submitted agents in a TEE, using a set of bespoke test harnesses. The Diamond agent provided behavioral, static, and dynamic testing. The output from the evaluation process produced a report that stakeholders could assess to determine trustworthiness.

Using Vijil, the NEAR AI platform could provide systematic proof output of agent trustworthiness to internal and external human stakeholders, as well as agent stakeholders.

As the NEAR AI Auditor Agent, Vijil Diamond operates as a Certificate Authority, issuing and revoking on-chain certificates assigned to trusted agents.

The report, and the process itself, running in a confidential computing environment, could be embedded in the attestation evidence the environment provides to prove:

- the environment is secure and verifiable
- the real Vijil auditing code (agent) itself is running (not a spoof)
- the trace and results being a part of the attestation evidence provided from the TEE provide mathematical backup for certification
- the agent code or data itself is never compromised

1. Agent Trustworthiness Testing

Custom test harnesses to certify the behavior of the NEAR AI Auditor Agent. Evaluations included:

- Behavioral consistency across various scenarios
- Reliability under adversarial conditions
- Compliance with declared intent
- Dynamic analysis of behavior for potential malicious functions

2. Proof of Integrity

The Auditor Agent performed evaluation in a trusted execution environment, allowing security and engineering teams to verify the chain of proof:

- Evidence that the code tested in the trusted execution environment is the same code that is deployed
- Certificate chaining, with the Audit Agent serving as the certificate authority

3. Regulatory Compliance by Design

Enabled demonstration of compliance with multiple regulatory frameworks using a shared testing infrastructure, minimizing blind spots for security and legal teams.

- Automated production of a report for audit and review by compliance stakeholders

For security and legal teams, this meant fewer blind spots and less regulatory fragmentation.

4. Platform Guardrail Validation

The evaluation extended to testing platform guardrails for accuracy of detections under hostile conditions. Since the platform facilitates agent-to-agent transactions, the test harness was specifically designed to identify attacks or exploits that would grant the agent excessive permissions to perform unauthorized transactions.

The Results:

Speed with trust

The impacts of the collaboration were immediate and measurable:

Uniform Trustworthiness Assessment

Security and compliance teams gained defensible evidence of agent security, reliability, and safety prior to deployment on the platform, allowing the AI team to operate within acceptable risk thresholds.

- Engineering and security teams could minimize the potential for financial loss through rogue agents, malicious code, or tampering slipping through manual audits

90% Reduction in Review Cycles

Review cycles were reduced from weeks to days, allowing NEAR AI to deploy agents faster and more confidently.

Automated Trust Attestation

The automation of testing processes for compliance requirements replaced manual audits, significantly lowering costs and demands on the AI engineering teams.

For the CISO, this shifted risk posture from reactive to provable. Instead of relying on policy attestations or post-hoc audits, NEAR AI could demonstrate—at any point in time—how agents were tested, what failure modes were evaluated, and how compliance evidence was generated prior to deployment.

Why This Matters to AI Teams

For AI teams developing agents, the key takeaway is:

- Trustworthiness of agents can be systematically evaluated and verified at scale — even for cutting edge and highly sensitive use cases
- The audit system itself runs confidentially and its integrity is verifiable through a chain of trust
- Continuous, provable trust enhances deployment efficiency while maintaining responsible innovation

NEAR AI succeeded by treating trustworthiness as infrastructure—not documentation.

For Heads of AI:	For SVPs of Engineering:	For CISOs:
This approach preserves velocity while ensuring responsible innovation.	It creates repeatability and removes deployment bottlenecks.	It provides systematic evidence, not just assurances.

Takeaway

Speed with trust

NEAR AI didn't slow down to become compliant. They re-architected trust so speed and safety reinforced each other.

NEAR AI defined and implemented an agent certification process that balanced business objectives against the risks of financial losses. By operationalizing verification and making the process of trustworthiness evaluation uniform and systematic, NEAR AI could confidently scale the process of auditing AI agents for the stringent conditions of its execution environment.

About Vijil

Vijil is the trust infrastructure that enterprises need to develop and deploy AI agents with reliability, security, and safety. Vijil compresses the time and effort to deploy trusted agents by 4x.