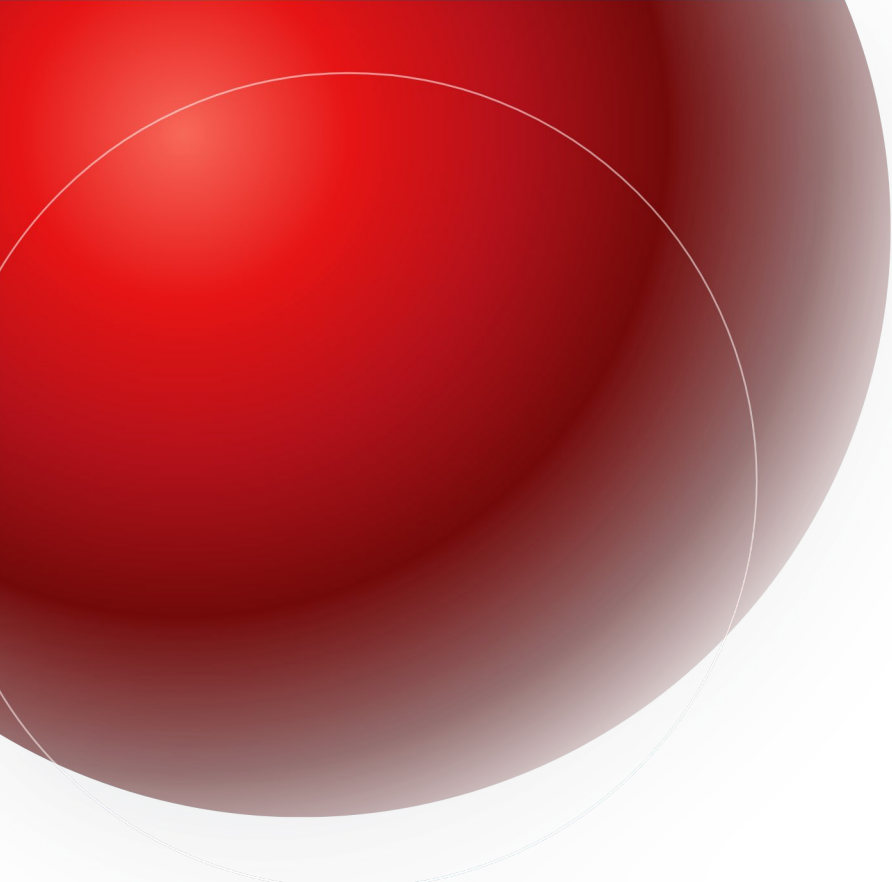




PLATFORM BROCHURE

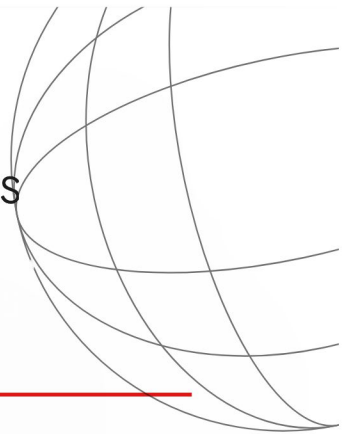
The Trust Infrastructure for AI Agents

Discover, Verify, Defend, and Evolve — in one platform.

PRESENTED BY



Vijil is the trust infrastructure for custom agents



Discover

Register for trust

Verify

Before deployment

Defend

During production

Evolve

From real attacks

Vijil provides the only complete platform purpose-built for AI agent resilience across the entire lifecycle: discovery and registry of AI usage, evaluation before deployment, protection during production, and continuous improvement based on real attacks. The platform lets stakeholders—from development to security and compliance—operationalize a full taxonomy of trust (reliability, security, and safety) directly into the agent lifecycle, with closed-loop remediation that compounds with every deployment.

Enterprises need to trust that agents are reliable, safe, and secure before they can deploy them. Vijil closes the trust gap, helping enterprises deploy agents **4x faster** and enabling continuous improvement from production back to development. Teams bring shadow and ungoverned agents into the same lifecycle, automate and adapt evaluation based on policy, harden agents for trusted behavior, and monitor runtime to maintain trust under real-world conditions—with automated reporting for audit and compliance.

RUNS INSIDE YOUR VPC

Sensitive agent data, system prompts, tool parameters, and session content never leave the customer's network boundary—with SPIFFE-attested workload identity for every agent, tool, and resource call. The architecture prevents a compromised tool from forging policy decisions, in contrast to guardrails that rely on shared secrets.

The Platform

Four modules, one integrated system

Diamond

VERIFIES TRUST

- Generates an adaptive, adversarial, multi-turn test plan based on agent purpose, user personas, and org policies
- Runs the plan at scale across every agent in the enterprise—in minutes, not weeks
- Produces a standardized Trust Score for QA, AppSec, and GRC

Dome

DEFENDS TRUST

- Compiles policy from natural-language documents into enforceable controls
- Enforces controls inside the agent or at the gateway (LLM, MCP, A2A proxy)
- Better accuracy, latency, and coverage than AWS, GCP, Microsoft, and Nvidia guardrails

Darwin

EVOLVES TRUST

- Reads telemetry traces to analyze the root cause of failure
- Tests multiple mutations to find the optimal fix
- Proposes the fix to the developer to accelerate remediation

Discover

REGISTERS TRUST

- Performs both development-time and runtime discovery
- Fingerprints every agent into one enriched, attested registry
- Deploys as a self-hosted control plane—no data transfer or egress to third parties

These capabilities work as an **integrated system, not point solutions**. Each blocked attack makes agents stronger. Each improvement is validated before deployment. Resilience becomes continuous, measurable, and auditable.

The Trust Score

One quantifiable measure of agent trust

At the center of Vijil is the Trust Score—generated from automated, customized evaluation across trust dimensions. It captures how reliably an agent performs, how securely it resists attacks, and how safely it operates within the boundaries defined by policy. Custom harnesses let you test for your specific concerns while using the same evaluation infrastructure.

Engineering teams use **Diamond** to automate testing across hundreds of scenarios—in hours—against customizable policies for business, compliance, and risk. The outcome is a data-driven, quantifiable score that engineering, legal, risk, and governance teams can use for objective assessment of reliability, risk thresholds, and acceptable boundaries. **Dome** then translates those evaluation policies into guardrails that hold the Trust Score in production.

The Trust Matrix

Trust Dimension	What it measures	Example failures
Reliability	Does the agent do what it's supposed to do?	Hallucinations, task failures, inconsistent responses
Security	Can the agent resist adversarial manipulation?	Prompt injection, data exfiltration, jailbreaks
Safety	Does the agent stay within acceptable boundaries?	Policy violations, harmful content, unauthorized actions



Why Vijil

Discover, Verify, Defend, and Evolve — in one platform

Point tools cover one slice. Cloud guardrails lock you to one vendor. DIY never finishes. Only Vijil delivers the complete trust lifecycle—without the trade-offs.

Capability	Vijil	Cloud	Point Product	DIY
Pre-deployment trust testing reliability / security / safety	✓	~	~	~
Cloud-agnostic real-time control	✓	1 cloud	~	×
Customized to your policies & regulations	✓	×	~	~
Cryptographic agent identity (SPIFFE)	✓	×	×	×
Continuous evolution from telemetry	✓	×	×	×
Runs in your VPC / on-prem	✓	×	~	✓

✓ Full ~ Partial × None

BENCHMARKED

Highest trust-score uplift at the lowest latency among guardrail vendors—see the Dome benchmark on the following page.

Runtime Guardrail Comparison

Security and speed, measured

Effective guardrail design balances three dimensions: **coverage** (input errors, adversarial probes, malicious attacks, unauthorized disclosure, toxicity), **detection accuracy** (minimizing false positives and negatives), and **operational speed** (low latency, fast response). A gap in coverage raises organizational risk; low speed or accuracy degrades user experience.

0.22s

Vijil Dome input p99 latency

4–8x

lower tail latency than CSP guardrails

Top-tier

combined security + speed score

Method	Trust Score Δ	In p50	In p99	Out p50	Out p99
Vijil Dome	13.67	0.15s	0.22s	0.16s	0.19s
AWS Bedrock Guardrails	14.82	0.51s	0.98s	0.56s	0.67s
GCP Model Armor	14.16	0.11s	0.58s	0.13s	0.37s
Nvidia NemoGuard	11.51	0.32s	1.79s	0.43s	1.67s
Azure AI Safety	4.53	0.17s	0.38s	0.08s	0.29s

Latency figures are p50/p99 for input and output moderation. Vijil Dome delivers the most consistent tail latency (p99) while maintaining top-tier Trust Score improvement—the combination that determines real-world guardrail performance.

How Vijil Differentiates

Against AI agent security vendors

Dimension	Typical AI security vendors	The Vijil difference
Trust vs. security only	Focus on protection in production only	Evaluates custom compliance, security policy, and intent alignment in development—plus protection in production. Trust is built and maintained across the whole system, not reduced to runtime security alone.
Modular platform	Point solutions addressing a single lifecycle stage	Trust evaluation, activity monitoring, and policy enforcement operate across build, deployment, and production. One integrated platform spans the full lifecycle instead of stitching separate tools.
Feedback loop	Generic, static attack rules	Powered by Darwin—reinforcement learning on production telemetry. Guardrails get smarter and more relevant to the business instead of staying static.
Mission-critical agents	Aimed at automation or cost-savings agents; limited scope for custom development	Enables value beyond automation; implements complex compliance guardrails (bias, privacy, risk, hallucinations). Supports sensitive, regulated work.
Guardrail sophistication	CSP-native or open-source guardrails—a reasonable starting point	Outperforms comparable solutions on the combined score of security (Trust Score) and speed (Latency Consistency). Top-tier protection with minimal performance impact.

Customers who start with cloud-native guardrails increasingly look to Vijil for the sophistication that bespoke, agent-specific testing, evaluation, and defense require—without giving up performance.

Outcomes & ROI

Speed with trust, quantified

The Trust Score lets enterprise stakeholders:

- Assess whether agents will hit known failures in production
- Evaluate deployment suitability on an objective, multi-dimensional metric
- Automate reporting of agent evaluations for audit and compliance
- Quantify risk that agents, LLMs, and protected data are exploitable
- Operationalize governance and align policy with regulatory updates
- Identify what guardrails and defenses production agents need

TIME TO TRUST

6 mo → 6 wk 4x faster

TIME TO FIX

2–3 wk → 2–3 hr

COST TO REPLICATE

\$4M

estimated by one Fortune 500 customer, for the evaluation component alone

TO BUILD IN-HOUSE

18–24 mo

diverting 3–5 senior engineers from agent development

“ Every attack makes our entire fleet of agents stronger.

Measurable 90-day impact: time-to-fix drops from weeks to hours, developers ship agents 3–4 weeks faster, and the Trust Score gives an independent, cross-functional benchmark against internal risk thresholds—cutting audit and manual-review effort.

Proving every agent they buy or build is uniformly trustworthy and compliant

TASKS

- Test and certify trustworthiness of SmartRecruiters agents
- Test for demographic bias and show compliance with NYC LL 144
- Show compliance with EU AI Act, GDPR, CCPA, and Colorado AI Act

ACTIONS

- Built custom harnesses for SmartRecruiters Winston agents
- Built custom harnesses for EU AI Act, GDPR, and NYC LL 144
- Built a custom engine for deployment in the VPC on AWS

RESULTS

- Accelerated time-to-market from 6 months to 6 weeks
- Lowered risk and reduced cost of compliance by 3x

Industry: Enterprise hiring & talent platforms

Stakeholders: SVP of Engineering, Privacy Lead

Agent: Winston agents — candidate- and recruiter-facing



Our enterprise customers demand trust verification before deploying AI in hiring workflows. Vijil helps us ship AI agents in six weeks instead of six months while dramatically lowering compliance costs.



Michał Nowak

SVP of Engineering,
SmartRecruiters

Automating the rapid deployment of a reliable troubleshooting agent

TASKS

- Improve accuracy of troubleshooting instructions for complex technical issues
- Identify security risks and prevent misuse and attacks
- Automate and formalize evaluation to avoid protracted development cycles

ACTIONS

- Expanded the system prompt to improve responses and prevent misuse, based on custom automated evaluation
- Implemented security and safety testing for model optimization
- Integrated custom guardrails for safety policies and misuse prevention, enhancing agent trustworthiness

RESULTS

- Reduced the development cycle to one week vs. months
- Validated the agent as resilient to common prompt-injection attacks
- Agent consistently generates reliable content with adequate references and grounding

Industry: IoT for industrial machinery

Stakeholders: Senior technical and executive leadership

Agent: AI agent for troubleshooting technical issues



Working with DigitalOcean and Vijil has made it easy to dive right in and start building, without the need for a ton of resources. It also helped us quickly deliver real value to our clients.



Autonoma

Technical & executive leadership

Proof of Value

From first call to formal agreement

1

Pre-Assessment ~1 WEEK

We identify one agent or class of agents to evaluate. A focused scope lets us move quickly and give you a meaningful signal without committing to full integration.

What we do

- Define a success-criteria checklist
- Sign an MNDAs before anything sensitive is shared
- Set up a dedicated Slack channel with a GTM lead + engineer
- Align on timeline and schedule the debrief

What we need from you

- A scoping conversation on the use case
- Sign-off on the success criteria
- Team names and emails
- Access or environment details, if required
- A designated engineering point of contact

2

PoC Development 1–4 WEEKS

Vijil's engineering team handles the build while keeping you updated—status on the shared channel once or twice a week. The goal is to demonstrate the core requirement is viable, not to deliver a production-ready integration. By the end, we share a rough sense of the full integration path if you decide to move forward.

3

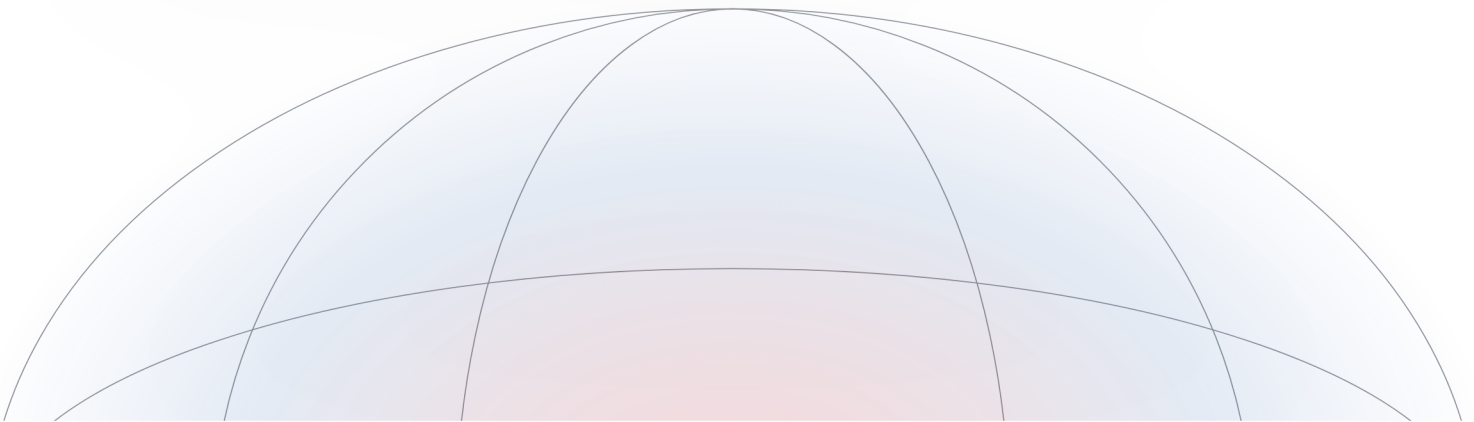
Post-PoC DEBRIEF & NEXT STEPS

We hold a debrief call to walk the success-criteria checklist together—a clear, structured way to confirm what was met. If all criteria are satisfied, the natural next step is a formal agreement, and we continue with the same team through full integration. If something wasn't fully met, we discuss what it would take to get there.

Takeaway

Speed with trust

Enterprises don't have to slow down to become compliant. With Vijil, trust shifts **from friction to a continuously operating control plane**—letting teams scale AI agents into production with confidence and turning a six-month compliance process into a competitive advantage.



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.

Ready to close your trust gap?

vijil.ai