

The platform for agent trustworthiness and resilience

The Challenge: The Trust Gap for AI Agents

The question every enterprise now faces isn't whether to deploy AI agents—it's how to ensure agents are resilient enough to move into production and whether they can be trusted once they are in production. Defining, evaluating, and maintaining trust for agents from development to production can be complex, since agents in production can fail in multiple and unpredictable ways. Agents are capable of producing different outputs to identical inputs, shaped by context and their interactions with tools and data.

The dimensions of trust - reliability, security, and safety- are usually evaluated independently by point products managed by different stakeholders at different phases of the agentic lifecycle - often even using manual audits, point in time red teaming, and legacy software testing.

As a result, enterprises run the risk of failures showing up in production that extend beyond poor outcomes:

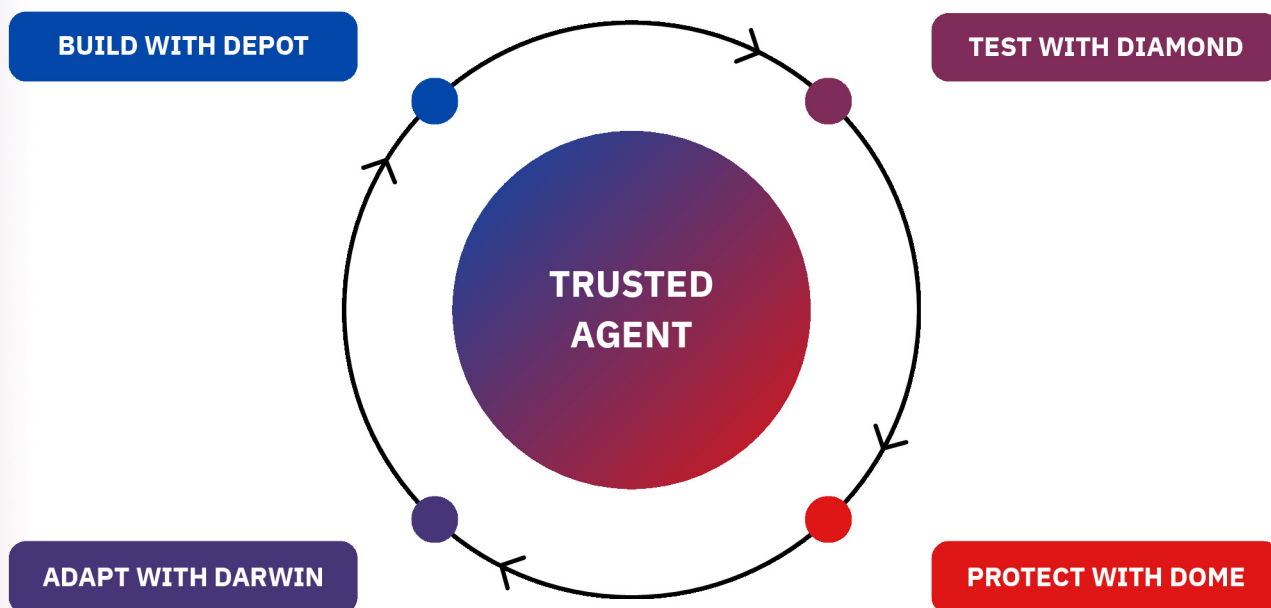
- hallucinations can cause corporate harm
- attackers can launch successful exploits through methods like prompt injections
- agents can violate compliance rules against bias or data privacy violations
- agents can make unanticipated system changes that result in downtime

Solution: Continuous Trust Layer for Agents

Vijil provides the only complete platform purpose-built for AI agent resilience across the entire lifecycle: evaluation before deployment, protection during production, and continuous improvement based on real attacks. The platform enables stakeholders from development to security and compliance to operationalize a full taxonomy of trust—reliability, security, safety—directly into the agent lifecycle, with closed-loop remediation that compounds with every deployment. With Vijil, teams automate and customize agent evaluation, accelerate the time to value with more trusted agents, with monitoring and reporting to maintain trust under real-world conditions.

Vijil provides a foundation for to align enterprise stakeholders across the agentic lifecycle:

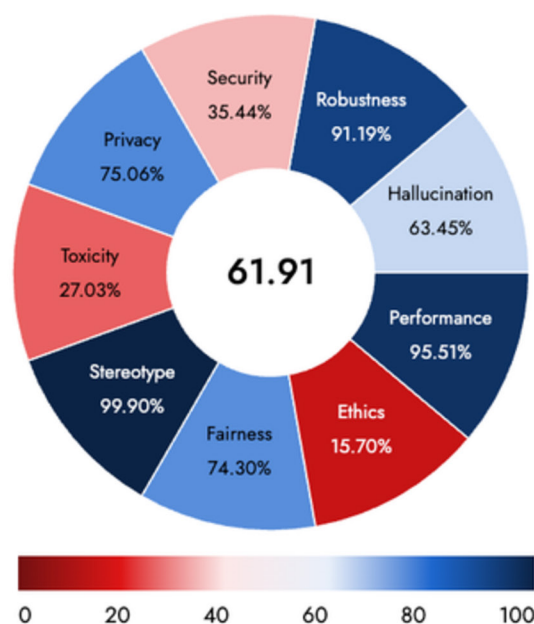
- Vijil's Diamond automates evaluation based on custom, bespoke policies for the business context of the agent
- Vijil Dome enforces those policies to protect and monitors the agent in runtime through embedded guardrails, based on the specific risks identified by the evaluation
- Vijil Darwin automates the process of making agents more resilient, closing the loop by transforming attack telemetry into defensive improvements automatically.



Operationalizing Agentic Trust

The Vijil Trust Score

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



Continuous Trust

Developers and engineering teams use Diamond to automate testing in hours of customizable policies for business, compliance, and risk—requirements across hundreds of scenarios. The outcome is a data-driven, quantifiable Trust Score that engineering teams and stakeholders like legal, risk and governance teams can use for objective assessment, whether in terms of reliability, risk thresholds, and acceptable boundaries.

Diamond performs static and dynamic code scanning within a trusted execution environment to evaluate how code would execute in production conditions.

Security and engineering teams use Dome to automatically translate evaluation policies into guardrails to maintain the Trust Score - ensuring that runtime protection operates to minimize failures the testing identified - as well as contain security risks and prevent compliance violations.

Darwin reinforces the Trust Score by supplementing trust evaluation and policy-driven guardrails at the build and improve phases of the agentic lifecycle. Darwin applies reinforcement learning over operational telemetry to support a feedback loop for adaptation of agent design and controls.

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

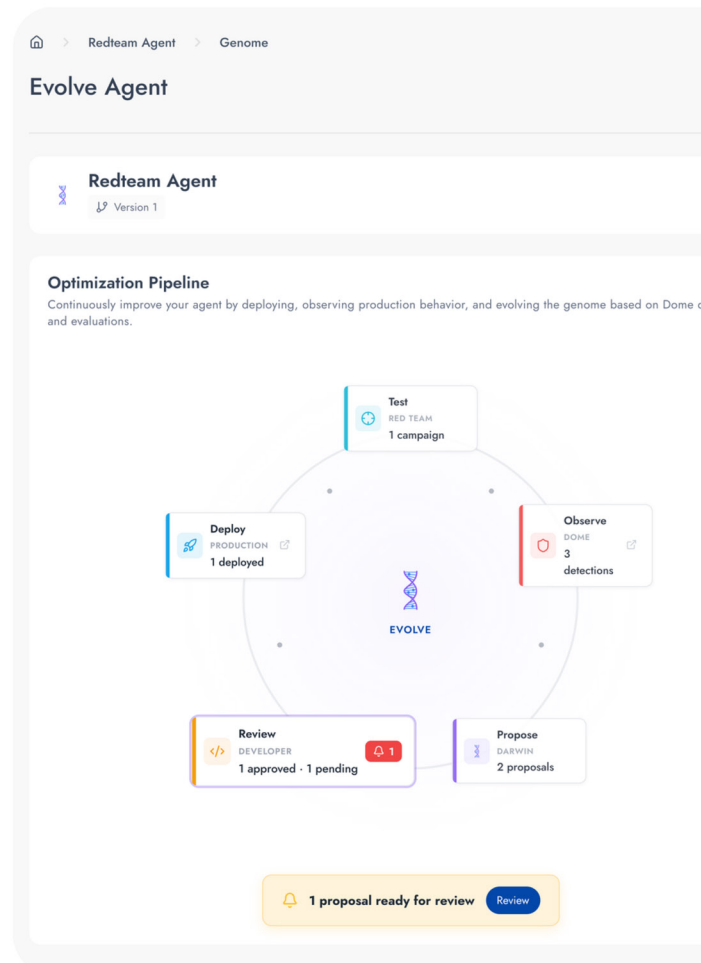
The Vijil Trust Score allows enterprise stakeholders to:

- Assess whether agents will encounter known failures in production environments
- Quantify risk that agents, LLMs & protected data are vulnerable to exploits and attacks
- Evaluate deployment suitability based on an objective, multi-dimensional metric
- Operationalize governance frameworks, align system-level policies with regulatory updates
- Automate reporting of agent evaluations for audit, and compliance
- Identify what guardrails and defenses are needed to protect the agent in production

Vijil Features and Benefits

Key Features

- **Custom** - Rapid customization of evaluation for enterprise-specific trust requirements - not just generic failures and exploits
- **Continuous** - Maintain trust throughout the entire agent lifecycle, from development through production deployment, with ongoing reinforcement
- **Context** - Evaluate for business context, persona, and compliance requirements and automate conversion of policies into runtime guardrails
- **Automated** - Evaluate agents in hours, across hundreds of scenarios and deploy guardrails straight away - not after weeks of manual audits and security reviews
- **Test and Prevent** - Optimize agent runtime controls based on the results of comprehensive security testing - contain unpredictable actions.
- **Inside Out** - Defining and enforcing trust by design within agents as an integral part of development and deployment process, not just applying external protection in runtime



Benefits

- **Accelerated Time-to-Production:** Vijil provides the evidence stakeholders need to understand safety, resilience, and compliance risks, compressing the time taken to move agents into production by 4x.
- **Customized Risk Mitigation:** Organizations can tailor guardrails and policies to their specific business contexts, optimizing agent operations while effectively managing risks associated with AI deployment.
- **Increased Agent Resilience:** Continuous learning mechanisms help ensure that agents adapt to new challenges by enhancing their robustness and reliability in dynamic environments.
- **Enhanced Collaboration:** The platform fosters collaboration among security, governance, and engineering teams, unifying their efforts to streamline the agentic deployment process and mitigate potential risks.

Customer Testimonials

“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

Senior Vice President, Engineering

SmartRecruiters

“By adapting the Google Responsible Generative AI Toolkit to the needs of enterprises in various industries, Vijil provides critical capabilities for AI developers to preserve the privacy, security and safety of custom models downstream with the same rigor that went into their original release.”

Manvinder Singh

Director of Product Management

Google

“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 to quickly deliver real value to our clients.”

Benedikt Klingmayr

Full Stack Developer

Autonoma

Deployment Options

Vijil can be deployed as	
A hosted service (SaaS)	Self hosted (on a VPC, or entirely within the corporate network)

Integrations

Cloud Platform



Agent Developer Framework:



Amazon Bedrock

CI/CD



CONCLUSION

Before they can deploy agents into production, enterprises need to be able to trust - not hope - that agents will operate the way they are intended to do. Vijil's modular platform provides the infrastructure to evaluate in development, protect in production, and monitor agentic trust across the lifecycle - in whichever environments agents are built and run. Vijil fundamentally transforms the deployment of AI agents by addressing trust concerns and overcoming operational, technical, and organizational challenges.

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.