



Agent Deployment Readiness

Evaluating Agent Resilience For Deployment

When is an agent ready for production? Despite the extensive investments in AI agent development, this is often an open question for agent developers and their organizations.

- Security wants to know is the agent vulnerable to prompt injections and agent-native threats?
- Engineering wants to know is the agent reliable?
- Compliance wants to know whether the agent will behave within external and internal policies?

To exit from pilot limbo, teams need an objective, quantifiable, and comprehensive assessment for each stakeholder of how an agent will behave across multiple scenarios in production conditions - that is specific to the agent and organizational risk thresholds.

Score Trustworthiness For Prod Readiness

Vijil Diamond creates a baseline evaluation of the trustworthiness of the agent with bespoke test cases derived from agent specifications, user personas, and agent components. The output is the **Vijil Trust Score** - objective, third-party evidence of the agent's behavior based on reliability, security and safety parameters. Stakeholders can evaluate the score against 36 dimensions to assess whether the agent meets risk tolerance, compliance mandates, and functional thresholds. Engineering teams can establish trust promotion gates for development and respond to changes in trust when failures happen in production.

Advantages

- **Comprehensive:** Includes continuously evolving tests across hundreds of scenarios to generate a composite reliability, security and safety score
- **Bespoke:** Tests customized to your agent specification, user personas, and org policies
- **Policy-driven:** Tests based on policies that matter to the agent's role: regulations, industry standards, and custom policies
- **Flexible:** Bring your own dataset, or use Vijil's curated golden dataset
- **Multi-level:** Tests multi-agent systems, multi-turn interactions, and multi-component (LLM, knowledge base, MCP & A2A) agents
- **Auditable:** Uniquely, verifies agent code (SAST) and behavior (DAST) inside a trusted exec environment (TEE)
- **Efficient and Automated:** Produces Vijil Trust Score in minutes
- **Actionable:** Recommends test-driven remediation for safety, security, and reliability issues to close the trust gap.

Bridge Pilot to Production Chasm



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.

– SVP Engineering, **SmartRecruiters**



Trust Score Report

Create Policy

Write Policy Upload Policy

BASIC INFORMATION

Policy Name

e.g., GDPR Compliance Policy, Security Guidelines

Description (optional)

Brief description of what this policy covers...

CATEGORY & STATUS

Category

Custom

Security

Security and access control rules

Ethics

Ethical guidelines and principles

Operational

Operational constraints and limits

Compliance

Define bespoke testing policies