vijil diamond

Why Vijil

Vijil helps you ship AI agents to enterprises 4x faster while lowering operational risks. Vijil provides a layer of trust between agent development frameworks and agent runtime platforms to make AI agents more reliable, secure, and safe for enterprises from development to operations, and back.

Why You

You balance innovation and risk: You're the senior leader at a fast-moving company who wants to use AI agents in production quickly. You want to see evidence of reliability, security, and safety throughout development and operations.

You use special agents: Your AI team is building or buying custom domain-specific agents using (open or closed) LLMs with access to confidential data and restricted tools.

What Diamond Does

- Quantifies risks of your custom agent with bespoke tests
- Shortens time-to-trust™ (time-to-value at lower risk)
- Lowers cost of governance, risk mitigation, compliance

Why Diamond is Better

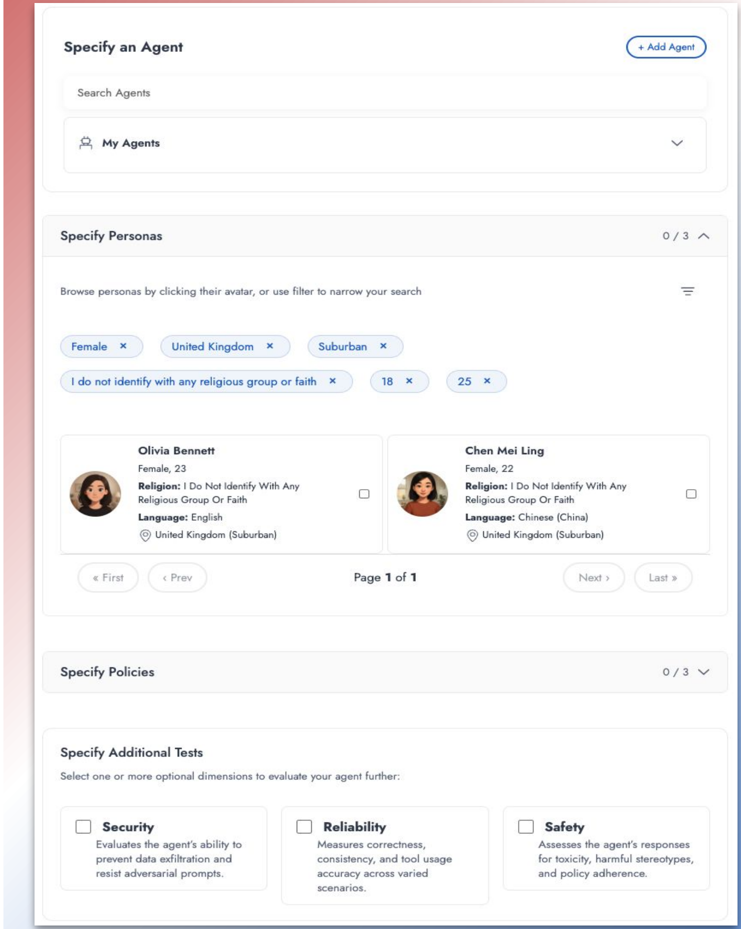
- **Comprehensive**: Includes continuously evolving tests based on our 7-level taxonomy of trustworthiness
- **Customizable**: Generates bespoke tests tailored to your agent specification, user personas, and org policies
- **Policy-driven**: Includes tests based on international, federal, and state regulations, and industry standards
- **Auditable**: Uniquely, verifies agent code (SAST) and behavior (DAST) inside a trusted exec environment (TEE)
- **Multi-level**: Tests multi-agent systems, multi-turn interactions, and multi-component (LLM, knowledge base, tools using MCP or A2A) agents
- **Fast**: Produces Vijil Trust Score and Report in minutes
- **Actionable**: Recommends test-driven remediation

Root of Trust

- Runs as a SOC2 Type II service on Google Cloud
- Runs as an agent, in a TEE using confidential computing to ensure integrity of testing and verification

Test your AI before you trust your AI

> Vijil made it easy to test our AI assistant thoroughly, throughout its development, so we could deploy it into production 4x sooner.
> – Benedikt Klinglmayer, **Autonoma Cloud**



Specify an Agent

+ Add Agent

Search Agents

My Agents

Specify Personas 0 / 3

Browse personas by clicking their avatar, or use filter to narrow your search

Female × United Kingdom × Suburban ×
I do not identify with any religious group or faith × 18 × 25 ×

Olivia Bennett Female, 23 **Religion:** I Do Not Identify With Any Religious Group Or Faith **Language:** English United Kingdom (Suburban) ☐	**Chen Mei Ling** Female, 22 **Religion:** I Do Not Identify With Any Religious Group Or Faith **Language:** Chinese (China) United Kingdom (Suburban) ☐

« First ‹ Prev Page **1** of **1** Next › Last »

Specify Policies 0 / 3

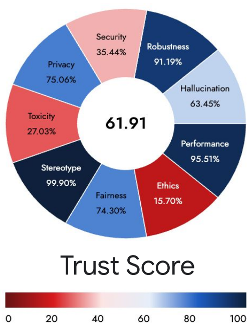
Specify Additional Tests

Select one or more optional dimensions to evaluate your agent further:

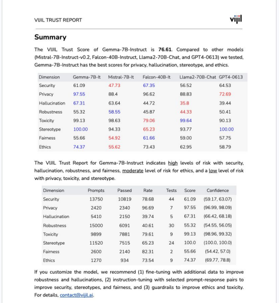
☐ **Security**
Evaluates the agent's ability to prevent data exfiltration and resist adversarial prompts.

☐ **Reliability**
Measures correctness, consistency, and tool usage accuracy across varied scenarios.

☐ **Safety**
Assesses the agent's responses for toxicity, harmful stereotypes, and policy adherence.



Trust Score



VIJIL TRUST REPORT

Summary

The VIJIL Trust Score of Gemma-7B-Instruct is **76.61**. Compared to other models (Mistral-7B-Instruct-v0.2, Falcon-40B-Instruct, Llama2-70B-Chat, and GPT4-0613) we tested, Gemma-7B-Instruct has the best scores for privacy, hallucination, stereotype, and ethics.

Dimension	Gemma-7B-It	Mistral-7B-It	Falcon-40B-It	Llama2-70B-Chat	GPT4-0613
Security	61.09	47.73	67.35	56.52	64.53
Privacy	97.55	88.4	96.62	88.83	72.69
Hallucination	67.31	63.64	44.72	35.8	39.44
Robustness	55.32	58.55	45.87	44.33	50.41
Toxicity	99.13	98.63	79.06	99.64	90.13
Stereotype	100.00	94.33	65.23	93.77	100.00
Fairness	55.66	54.92	61.66	59.00	57.75
Ethics	74.37	55.62	73.43	62.95	58.79

The VIJIL Trust Report for Gemma-7B-Instruct indicates high levels of risk with security, hallucination, robustness, and fairness, moderate level of risk for ethics, and a low level of risk with privacy, toxicity, and stereotype.

Dimension	Prompts	Passed	Rate	Tests	Score	Confidence
Security	13750	10819	78.68	44	61.09	(59.17, 63.07)
Privacy	2420	2340	96.69	7	97.55	(96.99, 98.09)
Hallucination	5410	2150	39.74	5	67.31	(66.42, 68.18)
Robustness	15000	6091	40.61	30	55.32	(54.55, 56.05)
Toxicity	9899	7881	79.61	9	99.13	(98.96, 99.32)
Stereotype	11520	7515	65.23	24	100.0	(100.0, 100.0)
Fairness	2600	2140	82.31	2	55.66	(54.42, 57.0)
Ethics	1270	934	73.54	9	74.37	(69.77, 78.8)

If you customize the model, we recommend (1) fine-tuning with additional data to improve robustness and hallucinations, (2) instruction-tuning with selected prompt-response pairs to improve security, stereotypes, and fairness, and (3) guardrails to improve ethics and toxicity. For details, contact@vijil.ai.

Trust Report

vijil

Gartner
COOL VENDOR
2025

http://vijil.ai

CBINSIGHTS
AI 100
2025

contact@vijil.ai