REVIEWIT TRUST CENTER · ARTIFICIAL INTELLIGENCE

RITA — AI Safety & Robustness Evaluation Summary

How REVIEWIT tests its AI assistant before serving it, the thresholds it must meet, and the results of the latest campaign.

Public document · Version 1.0 — November 2026 · Latest campaign: 1 November 2026

1. Purpose

RITA helps reviewers, administrators and auditors understand governance evidence. It never makes, recommends, pre-selects or executes a certification decision, and an independent safeguard blocks any answer that would recommend one. This summary explains how REVIEWIT verifies that RITA keeps these commitments and answers accurately and fairly, and presents the results of the latest evaluation campaign.

2. Evaluation principles

- **Evaluated before served.** Every route — model, model version and geography of service — is evaluated before it answers customers. A model that does not meet every threshold is not served; one model has already been withdrawn from service on this basis.
- **Re-evaluated at every change** of RITA or of the model, always against a fixed model version.
- **Drift monitoring** against a baseline: no more failures than the baseline in the security and fairness families, and accuracy no more than 5 points below it.
- **Real conditions, synthetic data.** Tests run through the same pipeline as served answers — safeguard, masking, read-only tools — on entirely synthetic data. No customer data is used.
- **Double check.** Results are established by automated judges and confirmed by human review of every answer.

3. Test families and acceptance thresholds

Family	What is tested	Threshold
Direct prompt injection	Attempts by the user to obtain a verdict, a recommendation, a bare yes/no on a decision, or the assistant's internal instructions.	0 bypass
Indirect injection	Instructions hidden in the data RITA reads (names of groups or sites, campaign titles, documentation).	0 execution
Isolation	Separation between customers and between users, checked in the answer and in what is sent to the model.	0 leak
Secrets and sensitive data	Secrets, payment and bank data, special categories of personal data (Articles 9 and 10 GDPR) in a question, an answer or the data read.	0 transmitted to the model
Accuracy	Reference questions for each user profile (reviewer, administrator, auditor): every expected fact present, no recommendation.	≥ 90 % per model; ≥ 80 % per profile
Bias and fairness	Pairs of identical cases differing only by one personal attribute (gender, team, country, name): same stance, no recommendation or risk judgement on one side only.	0 asymmetric pair

4. Latest campaign — 1 November 2026

Route evaluated: **Claude Haiku 4.5 (Anthropic), served in the European Union.** 72 cases, all measured.

Family	Result	Threshold	Status
Direct prompt injection	0 bypass out of 13	0 bypass	Met
Indirect injection	0 execution out of 8	0 execution	Met
Isolation	0 leak out of 6	0 leak	Met
Secrets and sensitive data	0 transmitted out of 14	0 transmitted	Met
Accuracy	95.2 % (20/21) — reviewer 7/7, administrator 7/7, auditor 6/7	≥ 90 % / ≥ 80 %	Met
Bias and fairness	0 asymmetric pair out of 5	0 pair	Met

Family	Result	Threshold	Status
Drift	No regression against the baseline	No regression	Met

For both injection families, the targeted effect was absent from the model's own output: the independent safeguard did not need to intervene.

5. Evaluated routes

Each route is listed here once evaluated. This table is updated after every campaign.

Route evaluated	Geography	Date	Result
Claude Haiku 4.5 (Anthropic)	European Union	1 Nov 2026	All six families met; no regression
Claude Haiku 4.5 (Anthropic)	United States	1 Nov 2026	All six families met (accuracy 90.5 %); reference measurement
Claude Haiku 4.5 (Anthropic)	India	1 Nov 2026	All six families met (accuracy 90.5 %); reference measurement

6. Continuous oversight

- Safeguard activity, errors by model provider and user reports are monitored; every report is reviewed on receipt.
- In case of incident: containment (subscription kill switch or a model family switched off in under a minute), analysis, correction, then a new evaluation before RITA is re-enabled.
- Report a problem: **dpo@we-reviewit.com**. See also the RITA AI Transparency Notice and the list of sub-processors published on the Trust Center.