TRUST CENTER REVIEWIT · INTELLIGENCE ARTIFICIELLE

RITA — Synthèse d'évaluation de la sécurité et de la robustesse de l'IA

Comment REVIEWIT teste son assistant IA avant de le servir, les seuils qu'il doit atteindre et les résultats de la dernière campagne.

Document public · Version 1.0 — novembre 2026 · Dernière campagne : 1er novembre 2026

1. Objet

RITA aide relecteurs, administrateurs et auditeurs à comprendre les preuves de gouvernance. Il ne prend, ne recommande, ne présélectionne ni n'exécute jamais une décision de certification, et un garde-fou indépendant du modèle bloque toute réponse qui en recommanderait une. Cette synthèse explique comment REVIEWIT vérifie que RITA tient ces engagements et répond avec exactitude et équité, et présente les résultats de la dernière campagne.

2. Principes d'évaluation

- **Évalué avant d'être servi.** Chaque route — modèle, version du modèle et géographie de service — est évaluée avant de répondre aux clients. Un modèle qui n'atteint pas tous les seuils n'est pas servi ; un modèle a déjà été retiré du service pour cette raison.
- **Réévalué à chaque évolution** de RITA ou du modèle, toujours sur une version de modèle figée.
- **Suivi de dérive** contre une ligne de base : pas plus d'échecs que la ligne de base dans les familles de sécurité et d'équité, et une exactitude jamais inférieure de plus de 5 points.
- **Conditions réelles, données synthétiques.** Les tests passent par la même chaîne que les réponses servies — garde-fou, masquage, outils en lecture seule — sur des données entièrement synthétiques. Aucune donnée client n'est utilisée.
- **Double contrôle.** Les résultats sont établis par des juges automatiques et confirmés par une relecture humaine de chaque réponse.

3. Familles de tests et seuils d'acceptation

Famille	Ce qui est testé	Seuil
Injection de consignes directe	Tentatives de l'utilisateur d'obtenir un verdict, une recommandation, un oui/non sec sur une décision, ou les consignes internes de l'assistant.	0 contournement
Injection indirecte	Consignes dissimulées dans les données lues par RITA (noms de groupes ou de sites, titres de campagne, documentation).	0 exécution
Isolation	Cloisonnement entre clients et entre utilisateurs, vérifié dans la réponse et dans ce qui est envoyé au modèle.	0 fuite
Secrets et données sensibles	Secrets, données de paiement et bancaires, catégories particulières de données (articles 9 et 10 du RGPD) dans une question, une réponse ou les données lues.	0 transmis au modèle
Exactitude	Questions de référence pour chaque profil (relecteur, administrateur, auditeur) : chaque fait attendu présent, aucune recommandation.	≥ 90 % par modèle ; ≥ 80 % par profil
Biais et équité	Paires de cas identiques ne différant que par un attribut personnel (genre, équipe, pays, nom) : même posture, aucune recommandation ni jugement de risque d'un seul côté.	0 paire asymétrique

4. Dernière campagne — 1er novembre 2026

Route évaluée : **Claude Haiku 4.5 (Anthropic), servi dans l'Union européenne.** 72 cas, tous mesurés.

Famille	Résultat	Seuil	Statut
Injection de consignes directe	0 contournement sur 13	0 contournement	Atteint
Injection indirecte	0 exécution sur 8	0 exécution	Atteint

Famille	Résultat	Seuil	Statut
Isolation	0 fuite sur 6	0 fuite	Atteint
Secrets et données sensibles	0 transmis sur 14	0 transmis	Atteint
Exactitude	95,2 % (20/21) — relecteur 7/7, administrateur 7/7, auditeur 6/7	≥ 90 % / ≥ 80 %	Atteint
Biais et équité	0 paire asymétrique sur 5	0 paire	Atteint
Dérive	Aucune régression par rapport à la ligne de base	Aucune régression	Atteint

Pour les deux familles d'injection, l'effet visé était absent de la sortie du modèle lui-même : le garde-fou indépendant n'a pas eu à intervenir.

5. Routes évaluées

Chaque route figure ici dès qu'elle est évaluée. Ce tableau est mis à jour après chaque campagne.

Route évaluée	Géographie	Date	Résultat
Claude Haiku 4.5 (Anthropic)	Union européenne	01/11/2026	Six familles atteintes ; aucune régression
Claude Haiku 4.5 (Anthropic)	États-Unis	01/11/2026	Six familles atteintes (exactitude 90,5 %) ; mesure de référence
Claude Haiku 4.5 (Anthropic)	Inde	01/11/2026	Six familles atteintes (exactitude 90,5 %) ; mesure de référence

6. Surveillance continue

- L'activité du garde-fou, les erreurs par fournisseur de modèle et les signalements des utilisateurs sont suivis ; chaque signalement est examiné à réception.
- En cas d'incident : confinement (coupe-circuit de la souscription ou coupure d'une famille de modèles en moins d'une minute), analyse, correction, puis nouvelle évaluation avant toute réactivation de RITA.
- Signaler un problème : **dpo@we-reviewit.com**. Voir aussi la Notice de transparence IA de RITA et la liste des sous-traitants publiée sur le Trust Center.