

Une mesure de l'IA conversationnelle que d'autres pourront refaire et contester.

ELYSÉA mesure sa propre couche, placée entre une application et le modèle d'IA, selon un protocole écrit avant toute mesure ; elle a décidé de publier ses annotations dans un banc public à construire et d'en ouvrir la réannotation, et cherche des équipes qui voudront la vérifier, la refaire ou la contredire.

VOTRE PROBLÈME

- La **significativité manque** : « most benchmarks do not report statistical significance of their results ».¹
- La **validité de construit est fragile** : une revue de 445 bancs relève des pratiques qui « undermine the validity of the resulting claims ».²
- Les **juges automatiques biaisent les taux** tant qu'ils ne sont pas corrigés par leur sensibilité et leur spécificité mesurées.³
- Les **bancs cliniques récents se réannotent mal** : K-Bench ne publie ni ses consignes, ni les exemples de son juge, ni ses transcriptions ; VERA-MH est ouvert, et son premier volet porte sur l'idéation suicidaire.^{4,5,6}

CE QUE LA CAMPAGNE MET À DISPOSITION

Une même conversation longue, envoyée aux deux



Le comparateur : une même conversation longue, envoyée au modèle seul et au modèle avec ELYSÉA, jugée tour par tour.

ÉLÉMENT	ÉTAT AU 01/10/2026
Protocole à quatre configurations : modèle seul, consigne de sécurité simple, ELYSÉA, consigne d'un tiers	Écrit et scellé le 28/08/2026, amendé depuis ; nouveau scellement avant la campagne. Remis sur demande, publié avec le nouveau scellement.
Refus à tort mesurés sur des bancs dédiés	Définis dans le protocole.
Trois juges de trois familles de modèles, aucun d'Anthropic ni d'OpenAI	Configurés.
Validation des juges	Juges mécaniques et panel validés sur des jeux tenus à l'écart (amendement proposé) ; l'accord avec l'humain sera mesuré après le pilote sur l'annotation du seul fondateur, d'où l'intérêt d'une réannotation indépendante.
Changements postérieurs au premier scellement	Neuf, chacun en attente de sa pièce de décision ; document remis sur demande, publié avec le nouveau scellement.
Résultats et dépôt à identifiant pérenne	Après la campagne ; procédure de dépôt écrite.

CE QUE LA CAMPAGNE S'ENGAGE À RESPECTER

VÉRIFIABLE SUR LE PROTOCOLE, AVANT LES RÉSULTATS

- Cible de mesure et définitions écrites avant la mesure ; jeu caché, publié seulement après.^{7,2}
- Accord juge-humain comparé à l'accord entre humains ; taux corrigés du biais du juge.^{6,3}
- Différences appariées entre configurations, avec leurs erreurs-types ; refus à tort et latence à côté de la sécurité.^{8,9}
- Code, échantillon stratifié de transcriptions et annotations à publier dans le banc public, décidé et à construire ; auto-évaluation selon BetterBench.^{7,1}

Les résultats seront publiés après la campagne, qu'ils soient favorables ou non. Aucun chiffre n'est publié avant.

CE QUI DISTINGUE CETTE CAMPAGNE

BANC EXISTANT	CE QU'IL MESURE, SELON SA SOURCE	CE QUE LA CAMPAGNE AJOUTE
AILuminate (MLCommons)	Réponses dangereuses en un tour, jeu privé, notation relative.	Conversations longues, conduite envers la personne, annotations à publier. ^{10,11}
K-Bench	Banc calibré par des cliniciens ; matériel non publié.	Réannotation ouverte de l'échantillon publié. ⁴
VERA-MH	Banc ouvert, validé contre l'humain, premier volet sur l'idéation suicidaire.	Quatre interdits de conduite et refus à tort, configuration par configuration. ^{5,6}

COLLABORATIONS, CONDITIONS ET RÈGLES

Collaborations proposées, aucune engagée

- **Réannotation indépendante** d'un échantillon, par des annotateurs qualifiés (des cliniciens pour la crise).
- **Audit statistique** du protocole ; **co-publication** sur le banc rejouable.
- **Gratuit** : aucun paiement entre ELYSÉA et un laboratoire ; un projet peut être porté par un chèque d'innovation Innosuisse.¹²

Règles

- **Seules des transcriptions de bancs**, faites de cas synthétiques, sont partagées ; jamais de données d'utilisateurs réels (règlement général sur la protection des données, article 35).
- **Désaccord publiable** : un laboratoire peut réannoter et publier un désaccord avec nos mesures.
- **Annotation humaine externe** : soumise au cadre éthique de votre établissement avant tout recrutement d'annotateurs.

PROCHAINE ÉTAPE

Un échange de 45 minutes avec votre équipe sur le protocole ; après le nouveau scellement, une proposition écrite de réannotation d'un échantillon ou d'audit statistique. contact@elysea.app · elysea.app

SOURCES (pages web lues le 01/10/2026, sauf date indiquée) 1 Reuel et al., Stanford, NeurIPS 2024 — BetterBench 2 Bean et al., NeurIPS 2025 — Measuring what Matters: Construct Validity in LLM Benchmarks 3 11/2025, rév. 05/2026 — Lee et al., How to Correctly Report LLM-as-a-Judge Evaluations 4 Vowels et al., 14-15/09/2026 — K-Bench: a clinically calibrated benchmark... 5 Belli et al., 05/2026 — VERA-MH: Validation of Ethical and Responsible AI in Mental... 6 Bentley et al., 02/2026, v4 07/2026 — Human Validation Study of the Open-Source VERA-MH Safety... 7 NIST / CAISI, janv. 2026 — NIST AI 800-2 ipd 8 11/2024 — Miller, Adding Error Bars to Evals 9 Artificial Analysis, 11/06/2026 — Benchmarking guardrail models for safety, refusal, and latency 10 MLCommons — AILuminate — page sécurité 11 Ghosh et al., 02/2025, rév. 04/2025 — AILuminate: Introducing v1.0 of the AI Risk and Reliability... 12 Innosuisse — Innovation cheque