

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

*ELYSÉA mesure sa propre couche 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 : nous cherchons des équipes pour la vérifier ou la contredire.*

## L'ENJEU

- « most benchmarks do not report statistical significance of their results » ; la validité de construit de nombreux bancs est fragile.<sup>1,2</sup>
- Les juges automatiques biaisent les taux sans correction ; les bancs cliniques récents se réannotent mal.<sup>3,4</sup>

## CE QUE LA CAMPAGNE MET À DISPOSITION

- **Quatre configurations comparées** : modèle seul, consigne simple, ELYSÉA, consigne d'un tiers ; protocole écrit et scellé le 28/08/2026, amendé, nouveau scellement avant la campagne.
- **Trois juges de trois familles**, validés sur jeux tenus à l'écart ; accord avec l'humain mesuré après le pilote sur l'annotation du seul fondateur, d'où l'intérêt d'une réannotation indépendante.
- **Neuf changements postérieurs au premier scellement**, chacun en attente de sa pièce de décision ; document remis sur demande, publié avec le nouveau scellement.
- Face à AILuminate, K-Bench et VERA-MH : conversations longues, conduite envers la personne, réannotation ouverte.<sup>5,4,6</sup>

## VÉRIFIABLE AUJOURD'HUI

- **Sur demande** : le protocole et ses engagements (jeu caché, accord juge-humain, taux corrigés, différences appariées). **Public** : le comparateur sur [elysea.app/comparateur](https://elysea.app/comparateur) (données illustratives, marquées).<sup>7,8,3,9</sup>
- **Ce qui manque** : les résultats (publiés après la campagne, favorables ou non) ; le banc public, décidé et à construire.
- Collaborations proposées, aucune engagée, gratuites ; transcriptions de bancs synthétiques seulement ; annotation externe soumise au cadre éthique de l'établissement.

## PROCHAINE ÉTAPE

Un échange de 45 minutes sur le protocole ; après le nouveau scellement, une proposition écrite de réannotation. Détail : document *Laboratoires et universités*, deux pages. [contact@elysea.app](mailto:contact@elysea.app) · [elysea.app](https://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 MLCommons — AILuminate — page sécurité 6 Belli et al., 05/2026 — VERA-MH: Validation of Ethical and Responsible AI in Mental... 7 NIST / CAISI, janv. 2026 — NIST AI 800-2 ipd 8 Bentley et al., 02/2026, v4 07/2026 — Human Validation Study of the Open-Source VERA-MH Safety... 9 11/2024 — Miller, Adding Error Bars to Evals