

A measurement of conversational AI that others will be able to redo and challenge.

ELYSÉA measures its own layer under a protocol written before any measurement; it has decided to publish its annotations in a public benchmark, to be built, and to open them to re-annotation: we are looking for teams to check it or contradict it.

WHAT IS AT STAKE

- “most benchmarks do not report statistical significance of their results”; the construct validity of many benchmarks is fragile.^{1,2}
- Automatic judges bias the rates without correction; recent clinical benchmarks are hard to re-annotate.^{3,4}

WHAT THE CAMPAIGN MAKES AVAILABLE

- **Four configurations compared:** model alone, simple prompt, ELYSÉA, third-party prompt; protocol written and sealed on 2026-08-28, amended, new sealing before the campaign.
- **Three judges from three families**, checked on held-out sets; agreement with humans measured after the pilot against the sole founder's annotation, hence the value of independent re-annotation.
- **Nine changes made after the first sealing**, each awaiting its decision record; document provided on request, published with the new sealing.
- Compared with AILuminate, K-Bench and VERA-MH: long conversations, conduct towards the person, open re-annotation.^{5,4,6}

VERIFIABLE TODAY

- **On request:** the protocol and its commitments (hidden set, judge-human agreement, corrected rates, paired differences).
Public: the comparator at elysea.app/comparateur (illustrative data, labelled).^{7,8,3,9}
- **What is missing:** the results (published after the campaign, favourable or not); the public benchmark, decided and to be built.
- Proposed collaborations, none under way, free of charge; synthetic benchmark transcripts only; external annotation subject to the institution's ethics framework.

NEXT STEP

A 45-minute exchange on the protocol; after the new sealing, a written proposal to re-annotate. Details: document *Laboratories and universities*, two pages. contact@elysea.app · elysea.app

SOURCES (web pages read on 2026-10-01 unless dated) 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, rev. 05/2026 — Lee et al., How to Correctly Report LLM-as-a-Judge Evaluations ⁴ Vowels et al., 2026-09-14/15 — K-Bench: a clinically calibrated benchmark... ⁵ MLCommons — AILuminate — safety page ⁶ Belli et al., 05/2026 — VERA-MH: Validation of Ethical and Responsible AI in Mental... ⁷ NIST / CAISI, Jan. 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