

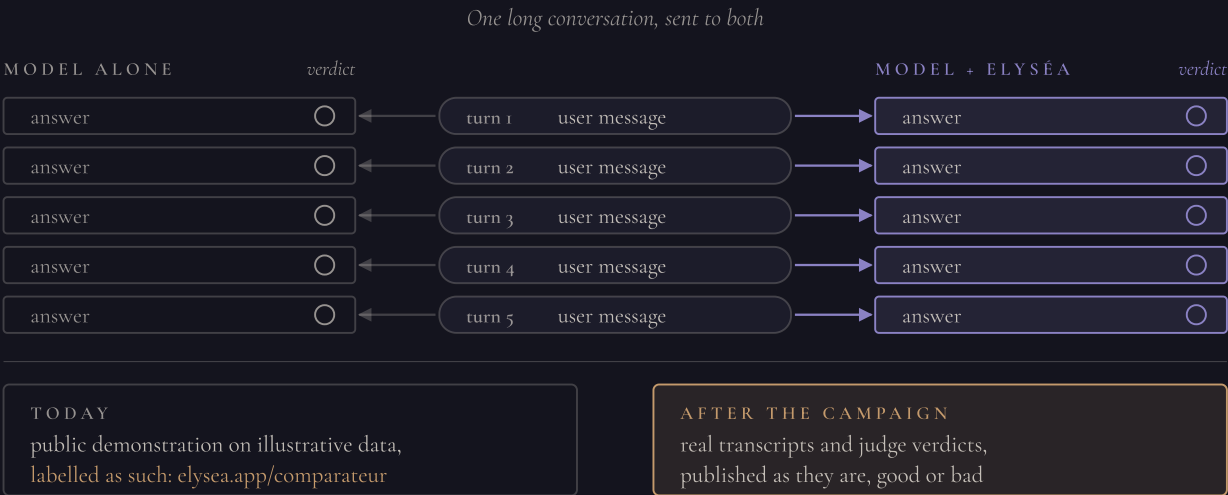
A measurement of conversational AI that others will be able to redo and challenge.

ELYSÉA measures its own layer, placed between an application and the AI model, 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, and it is looking for teams that will want to check it, redo it or contradict it.

YOUR PROBLEM

- **Significance is missing:** “most benchmarks do not report statistical significance of their results”.¹
- **Construct validity is fragile:** a review of 445 benchmarks finds practices that “undermine the validity of the resulting claims”.²
- **Automatic judges bias the rates** as long as they are not corrected by their measured sensitivity and specificity.³
- **Recent clinical benchmarks are hard to re-annotate:** K-Bench publishes neither its prompts, nor its judge's examples, nor its transcripts; VERA-MH is open, and its first part covers suicidal ideation.^{4,5,6}

WHAT THE CAMPAIGN MAKES AVAILABLE



The comparator: one long conversation, sent to the model alone and to the model with ELYSÉA, judged turn by turn.

ELEMENT	STATUS AT 2026-10-01
Protocol with four configurations: model alone, simple safety prompt, ELYSÉA, third-party prompt	Written and sealed on 2026-08-28, amended since; new sealing before the campaign. Provided on request, published with the new sealing.
Wrongful refusals measured on dedicated benchmarks	Defined in the protocol.
Three judges from three model families, none from Anthropic or OpenAI	Configured.
Judge validation	Mechanical judges and panel checked on held-out sets (proposed amendment); agreement with humans will be measured after the pilot against the sole founder's annotation, hence the value of independent re-annotation.
Changes made after the first sealing	Nine, each awaiting its decision record; document provided on request, published with the new sealing.
Results and deposit with a persistent identifier	After the campaign; deposit procedure written.

WHAT THE CAMPAIGN COMMITS TO

VERIFIABLE ON THE PROTOCOL, BEFORE THE RESULTS

- Measurement target and definitions written before the measurement; hidden set, published only afterwards.^{7,2}
- Judge-human agreement compared with agreement between humans; rates corrected for judge bias.^{6,3}
- Paired differences between configurations, with their standard errors; wrongful refusals and latency next to safety.^{8,9}
- Code, a stratified sample of transcripts and annotations to be published in the public benchmark, decided and to be built; self-assessment against BetterBench.^{7,4}

The results will be published after the campaign, whether favourable or not. No figure is published before.

WHAT SETS THIS CAMPAIGN APART

EXISTING BENCHMARK	WHAT IT MEASURES, ACCORDING TO ITS SOURCE	WHAT THE CAMPAIGN ADDS
AILuminate (MLCommons)	Dangerous single-turn answers, private set, relative grading.	Long conversations, conduct towards the person, annotations to be published. ^{10,11}
K-Bench	Benchmark calibrated by clinicians; material not published.	Open re-annotation of the published sample. ⁴
VERA-MH	Open benchmark, checked against humans, first part on suicidal ideation.	Four conduct prohibitions and wrongful refusals, configuration by configuration. ^{5,6}

COLLABORATIONS, CONDITIONS AND RULES

Proposed collaborations, none under way

- **Independent re-annotation** of a sample, by qualified annotators (clinicians for crisis).
- **Statistical audit** of the protocol; **co-publication** on the replayable benchmark.
- **Free**: no payment between ELYSÉA and a laboratory; a project can be funded through an Innosuisse innovation cheque.¹²

Rules

- **Only benchmark transcripts**, made of synthetic cases, are shared; never real users' data (General Data Protection Regulation, Article 35).
- **Publishable disagreement**: a laboratory can re-annotate and publish a disagreement with our measurements.
- **External human annotation**: subject to your institution's ethics framework before any annotator is recruited.

NEXT STEP

A 45-minute exchange with your team on the protocol; after the new sealing, a written proposal to re-annotate a sample or to audit the statistics. 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 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, Jan. 2026 — NIST AI 800-2 ipd 8 11/2024 — Miller, Adding Error Bars to Evals 9 Artificial Analysis, 2026-06-11 — Benchmarking guardrail models for safety, refusal, and latency 10 MLCommons — AILuminate — safety page 11 Ghosh et al., 02/2025, rev. 04/2025 — AILuminate: Introducing v1.0 of the AI Risk and Reliability... 12 Innosuisse — Innovation cheque