

---

## Dossier 07 — Laboratories and universities

**Audience:** research teams in AI evaluation, in psychiatry and public health (suicide prevention), in computer science, in digital law; Swiss institutions (also useful for Venture Kick and the innovation cheque: see Fo4, Fo5).

**Identity:** ethical cognitive infrastructure embeddable through an SDK. ELYSÉA is neither a companion, nor a chatbot, nor a coach, nor a therapist, nor a wellness product, nor a medical device.

## I. Their problem

- **Most benchmarks do not report significance:** “most benchmarks do not report statistical significance of their results” [S-MES-12].
- **Construct validity is weak:** review of 445 benchmarks, patterns that “undermine the validity of the resulting claims” [S-MES-13].
- **LLM judges bias the rates** if they are not corrected by sensitivity and specificity [S-MES-09].
- **Recent clinical benchmarks cannot be re-annotated:** K-Bench keeps prompts, judge examples and transcripts unpublished [S-MES-06]; VERA-MH is open but limited to suicidal ideation in its first part [S-MES-07] [S-MES-08].
- **ALLuminate** is single-turn, centred on dangerous responses, with a private set and relative grading [S-MES-03] [S-MES-04].

## 2. What the ELYSÉA campaign makes available to them (or will)

| Element                                                                                   | State                                                                                                                                                                   | Evidence                                                                                                                                                  |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Preregistered protocol, 4 arms (A bare · B naive guard · C ELYSÉA · D third-party prompt) | Sealed on 08-28, amended; resealing to come                                                                                                                             | elysea-tests-e2e:audits/PROTOCOLE_CAMPAGNE_4BRAS_3TIRS_SCELLE_2026-08-28.md:3@c9898dd3 · holding-seb-docs:dossiers/PROTOCOLE_TEL_QU_EXECUTE.md:28@8560b37 |
| Measured over-refusal (benches fp70, neg50)                                               | Defined                                                                                                                                                                 | elysea-tests-e2e:audits/PROTOCOLE_CAMPAGNE_4BRAS_3TIRS_SCELLE_2026-08-28.md:323@c9898dd3                                                                  |
| 3 judges from 3 families, none from Anthropic or OpenAI                                   | Configured                                                                                                                                                              | elysea-tests-e2e:scripts/run-campagne-4bras-2026-09-04.mjs:232-234@c9898dd3                                                                               |
| Judge validation                                                                          | Mechanical judges and panel validated on held-out sets (R-70, proposed, eze branch #517); judge/human validation D23 provided by Sébastien's annotation after the pilot | elysea-tests-e2e:audits/AMENDEMENT_R70_SENS_DECOUPE_STRUCTURELLE_2026-10-01.md:46@67316397                                                                |
| Reconciliation of post-sealing changes                                                    | Draft; 9 changes without a decision record                                                                                                                              | holding-seb-docs:dossiers/RECONCILIATION_PIECES_CAMPAGNE_2026-09-30.md:6@8560b37                                                                          |
| Results                                                                                   | to be measured — campaign in progress                                                                                                                                   | holding-seb-docs:depot-zenodo-2026-09-05/-RESULTATS_CAMPAGNE.md:2@8560b37                                                                                 |
| Zenodo deposit with DOI                                                                   | Written procedure; no DOI                                                                                                                                               | holding-seb-docs:depot-zenodo-2026-09-05/-GESTES_ZENODO_SEBASTIEN.md:37@8560b37                                                                           |
| Replayable public bench, published annotations, open re-annotation                        | Decided by Sébastien on 10-01; no evidence                                                                                                                              | —                                                                                                                                                         |

### 3. What the campaign must prove to be admissible (state of the art)

1. Measurement target and definitions written before measuring [S-MES-01] [S-MES-13].
2. Hidden set; documented public/hidden rule [S-MES-01] [S-MES-03].
3. Judge/human agreement compared with human/human agreement [S-MES-08]; frozen judge [S-MES-06].
4. Rates corrected for judge bias, interval including calibration [S-MES-09].
5. Clustered standard errors, paired differences between arms, preregistered power [S-MES-10].
6. Over-refusal and latency alongside safety [S-MES-11].
7. Code, stratified sample of transcripts, annotations published [S-MES-01].
8. BetterBench self-assessment published [S-MES-12].

### 4. Frequent objections and answers

| Objection                                                                   | Answer                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>You are both judge and party.</i>                                        | Yes, which is why the annotations will be published and re-annotation opened (decision of 10-01). A laboratory can re-annotate and publish a disagreement.                                                                                                                                                                                                                     |
| <i>Are your judges validated?</i>                                           | The mechanical judges and the 3-judge panel are validated on held-out sets (amendment R-70, status proposed, e2e branch #517: <code>elysea-tests-e2e:audits/AMENDEMENT_R70_SENS_DECOUPE_STRUCTURELLE_-2026-10-01.md:46@67316397</code> ); judge/human validation D23 provided by Sébastien's annotation after the pilot. The protocol refuses the run without this validation. |
| <i>The set will be contaminated.</i>                                        | The hidden set will only be published after measurement; what the public bench publishes remains to be defined with the labs.                                                                                                                                                                                                                                                  |
| <i>Clinical: who validated the boundary between referral and listening?</i> | No external clinician today. That is the purpose of a collaboration (section 5).                                                                                                                                                                                                                                                                                               |

### 5. Proposed collaborations (none exists)

| Collaboration                         | Laboratory's contribution                                  | ELYSÉA's contribution                            |
|---------------------------------------|------------------------------------------------------------|--------------------------------------------------|
| Independent re-annotation of a sample | Qualified annotators (clinicians for crisis)               | Access to anonymised transcripts and to the grid |
| Statistical audit of the protocol     | Miller / Lee / NIST review                                 | Protocol, code, data                             |
| Co-publication                        | Analysis, writing                                          | Data, replayable bench                           |
| Funded project                        | Research partner (Innosuisse cheque Fo4, Venture Kick Fo5) | Project holder                                   |

---

## 6. Compliance

Risky conversation data (mental health): GDPR art. 35, DPIA done and not signed (`holding-seb-docs:docs/-juridique/AIPD_ELYSEA_2026-09-30.md:20@8560b37`). **Decision:** only bench transcripts (synthetic cases) are shared; never real users' data.

---

## 7. Offer

Free of charge. No payment between ELYSÉA and a laboratory.

---

## 8. Concrete next step

After sealing: propose in writing to a Swiss evaluation or suicide-prevention team a re-annotation of a sample. No contact by this mission.

---

## Dossier sources

S-MES-01 · S-MES-03 · S-MES-04 · S-MES-06 · S-MES-07 · S-MES-08 · S-MES-09 · S-MES-10 · S-MES-11 · S-MES-12 · S-MES-13