

# NPL Call for Evidence - Cognition Observatory

## Response

Response submitted by Cognition Observatory on 30 September 2026 to the National Physical Laboratory's [Call for Evidence: AI testing, evaluation and assurance in the UK](#).

### Part 1 - About you

| Question              | Answer                                              |
|-----------------------|-----------------------------------------------------|
| Q1. Responding as     | On behalf of an organisation                        |
| Q2. Organisation type | Other - independent research initiative             |
| Q3. Size              | 1-9                                                 |
| Q4. Name              | Cognition Observatory                               |
| Q5. Sector(s)         | Cross-sector or sector-agnostic; Financial services |
| Q6. Perspective       | Primarily an intermediary                           |

### Part 2A - Ecosystem-wide challenges and capability gaps

#### Q1. Where do you see the most significant unmet needs or capability gaps in AI testing and evaluation for assurance?

The most significant gap we see is in testing how conversational and agentic AI systems behave across extended, multi-turn interaction, as opposed to scoring single outputs. Many of the failures we have documented appear only across a sequence of turns, and several are invisible to a user who is not actively checking.

Since August 2026, Cognition Observatory has logged roughly 40 specimens across three general-purpose assistants (Claude, ChatGPT, Microsoft Copilot), one deployed voice agent and one AI avatar. Recurring failure classes include:

- **Reversion to agreement:** after correction, the system drifts back to agreeing, so the user must re-apply pushback every turn.
- **Claimed correction without correction:** the system acknowledges an error, says it has fixed it, then repeats it. A deployed voice agent apologised for mispronouncing our founder's name, asserted it had said it correctly, and repeated the mispronunciation. The user's correction had no effect.
- **Procedural narration in place of evidence:** the system describes a verification step ("I reviewed the article...") instead of producing checkable evidence.
- **Unreliable self-report:** a system's account of its own accuracy does not track its actual accuracy.

Nor is the gap limited to conversation. In a 2025-26 engagement, our founder was project manager on a data migration at a large Canadian insurance and financial services company, where a vendor's LLM-based tool was used to convert pipelines from a legacy ETL platform to a cloud data-integration platform. In the proof of concept, the organisation's engineers manually migrated one application's pipelines; the LLM tool was assigned three applications' pipelines and completed none. In the full programme, it converted zero of several thousand in scope.

In our view, the gap applies wherever a person relies on a system's own account of its reliability or progress.

We see three capability gaps:

1. Repeatable multi-turn test protocols.
2. Measures of whether a system's presentation (confidence, fluency, stated corrections) matches its actual reliability.

3. Measures of effect on the human: whether repeated exposure reduces the user's own verification.

Our log is small-N, produced largely by a single researcher (Cognition Observatory's founder), and not yet independently replicated. We offer it as an indication of a gap, not as a measure of prevalence. Two underlying mechanisms are established in published research: models cannot reliably self-correct without external feedback (Huang et al., ICLR 2024), and sycophancy persists once triggered (SycEval, 2025). Method: cognitionobservatory.org.

## Q2. What is preventing the market or existing ecosystem from closing these gaps effectively?

- **The cheapest test is the least reliable.** Asking a model about its own reasoning or accuracy is easy, but published work (Huang et al., 2024) shows self-report is not dependable evidence.
- **No agreed rules on what counts as evidence.** If an evaluator tells a system which failure is being tested for and the system then shows it, that is weaker evidence than observing the failure unprompted. We are not aware of a standard for keeping the two apart, so findings can inflate.
- **Evaluator circularity.** Using one model to assess another is common. Agreement between models is convergence, not independent corroboration.
- **Review without new evidence looks like rigour.** In one logged case, four consecutive requests to reassess the same question produced four different conclusions with no new evidence in between. Iterative review without a stopping condition can converge on better-sounding uncertainty rather than accuracy.
- **The test object keeps moving.** Deployed behaviour depends on model version, product-layer tuning and system instructions, all of which change frequently. A point-in-time result may not hold weeks later.
- **Limited traceability.** Testers outside the provider usually cannot see which model version, system instructions or product-layer settings produced a given output, so results cannot be traced to a defined test object or reproduced.
- **Testing stops at the model, not the deployment.** In the voice-agent case, the underlying failure was model behaviour, but the user's live correction had no effect and no human was reachable on the call. Assurance needs to cover both the model and the deployment around it.
- **Status measures that track effort, not outcome.** In the migration case, status was reported as "effort completed" rather than pipelines migrated. It took roughly five months for client leadership and the vendor to accept the failure; the turning point was weekly reports showing zero pipelines migrated. Without outcome-based acceptance measures, an AI tool's activity can read as progress while delivering nothing.
- **Cost falls on the person relying on the system.** In the cases we logged, detecting and correcting the failure was left to that person, not the system or its provider. In our view, this and the cost of human-in-the-loop, multi-turn testing mean the market under-produces such testing relative to single-output benchmarks. This is our inference, not a measured finding.

## Q3. Where could public intervention add the greatest value, without duplicating or displacing activity the market is already delivering?

In our view, this work is pre-competitive, and no single vendor has a clear incentive to set a neutral standard for it. That is where a public body can add value without displacing the market.

1. **Reference protocols for multi-turn behavioural testing.** These should be public and versioned, with explicit stopping rules. For example, a stated correction counts only if it is accompanied by new evidence or independent verification, not by acknowledgement alone.
2. **Evidence grading and traceability for behavioural claims.** This means applying metrology discipline (a defined measurand, stated uncertainty, traceability) to claims such as "this system corrects errors when challenged." It includes rules for separating naturalistic observations from evaluator-induced ones. It also includes a minimum set of version and configuration metadata that providers disclose, so results can be traced to a defined test object.
3. **Shared replication infrastructure.** One harness would run the same protocol across vendors and versions under control conditions: ordinary use; adversarial use; adversarial use with the system made aware of the failure taxonomy; and adversarial use with the taxonomy withheld. The aim is to separate real behaviour from interrogation artefacts. It should include null tests, where the system is pressured on claims that are correct.

4. **Presentation calibration measures:** whether a system's confidence markers, fluency and stated corrections match its measured accuracy, so that how an output looks can be tested separately from whether it is right.
5. **Deployment-level checks,** such as whether an error flagged by a user reaches a human or changes system behaviour, as part of assurance for customer-facing agents.
6. **Outcome-based acceptance guidance** for AI tools used in delivery programmes, so progress is measured by verified output rather than reported effort.
7. **Human-side measures,** developed with HCI researchers and building on existing automation-bias research, of whether reliance on these systems degrades users' own checking behaviour.
8. **Open, low-cost verification guidance** for non-specialist users and deployers, who in the cases we logged bore the detection burden.

## Part 2B - Intermediary (optional)

---

### Q1. What expertise, evidence, or capability could you or your organisation contribute?

Cognition Observatory can contribute:

- **Elicitation and verification methods,** documented as written protocols but not yet independently replicated:
  - *Rabbit Hole:* repeated depth-push elicitation.
  - *Socratic Diagnostic:* targeted cross-examination that exposes internal contradiction.
  - *Stopping Rule:* a system's concession counts as a finding only if it brings new evidence or independent verification.
- **A vendor-agnostic challenge kit** of seven follow-up prompt categories (e.g. consistency, confidence calibration, evidence vs narrative, self-report distrust) for practitioners testing deployed systems.
- **A proposed epistemic-status tagging instrument.** It compares the evidential status a system claims for a statement with the status found on independent checking, tracking tag accuracy separately from claim accuracy.
- **The case log itself,** with evidence status recorded separately from classification. A three-level classification (observable modes, failure dynamics, cross-cutting mechanisms) has been adopted but not yet applied.
- **Programme delivery and risk management experience,** from our founder's nearly 30 years of senior programme, project and risk management, including an enterprise AI deployment. Also standards and professional-body links: Senior Member of ACM and IEEE, and Canada Member of IFIP TC12 WG12.13 (AI for Global Security).

## Part 3 - Intervention design

---

### Q1. Which approach do you think would be most valuable overall and why?

- d. the AI assurance methods accelerator, followed by (b) matched testing pilots.

In our view, the binding constraint in multi-turn behavioural testing is methodological validity, not tooling. Emerging methods need structured contact with measurement experts to define what is being measured, quantify uncertainty, and establish what counts as admissible evidence before they are scaled. Without that step, new assurance methods risk certifying the appearance of reliability rather than reliability itself.

Each method passing through the accelerator should leave with a defined measurand, a written protocol, a stated uncertainty or error rate, results from at least one independent replication, and a documented list of known limitations.

Once methods are validated, matched pilots should test them in real deployments, using protocols registered before testing begins. Two of our cases, a voice agent and an enterprise data migration, illustrate that failure can depend on deployment context (escalation paths, status reporting, acceptance measures) as much as on the model, and laboratory evaluation does not capture this.

We rank (a) and (c) lower for this problem. Short research sprints risk producing new categories and ideas faster than they can be validated, a pattern we have seen in our own work. A challenge fund works best once the problem and its measure are defined; for multi-turn behavioural testing, defining them is the unsolved part.

**Q2. Is there a different approach, not listed above, that you think would more effectively strengthen AI testing and evaluation?**

A standing, longitudinal replication programme. The listed options are time-bound; but deployed behaviour changes with each model version, product-layer update and system-instruction change, so a one-off result can go out of date within weeks.

We propose that the Centre for AI Measurement maintain a small, fixed battery of pre-registered multi-turn behavioural tests, run under the control conditions set out in Part 2A Q3, point 3, and re-run against major systems at regular intervals and on each significant version change. Results would be scored blind to condition, include null tests on correct claims, and be recorded against disclosed version and configuration metadata.

It would produce a public, versioned record showing whether behavioural failure modes are stable, improving or regressing across releases, with stated uncertainty, rather than a single snapshot.

**Q3. Which, if any, would you or your organisation be interested in participating in and in what capacity?**

- **(d) Methods accelerator:** as a participant bringing methods for validation. These are our multi-turn elicitation and verification protocols (Rabbit Hole, Socratic Diagnostic, Stopping Rule), together with our case log.
- **(b) Matched testing pilots:** as a research or delivery partner. We would contribute protocol design and outcome-based acceptance measures, drawing on our founder's programme and risk management experience in enterprise AI delivery.
- **The longitudinal replication programme proposed in Q2:** as a contributing research partner and advisor on test battery design.

More broadly, we would act as an advisor on multi-turn behavioural evaluation and deployment-side assurance.

**Q4. What specific conditions or support would your organisation need in order to participate?**

To participate, Cognition Observatory would need:

- **Remote participation** from Canada where possible.
- **Modest funding or cost recovery** for researcher time.
- **Access to test environments or API budgets** across multiple vendors.
- **Agreed publication rights** for methods and aggregate findings, so that protocols can be openly replicated, subject to deploying partners' confidentiality requirements.
- **Retention of pre-existing IP.** Cognition Observatory would keep ownership of methods developed before the programme, while granting licence for their use within it.
- **Agreed data-protection and ethics handling** for specimens involving third parties, including anonymisation.

**Q5. Are there particular areas you think we should focus on as part of this programme?**

**Types of AI**

- General-purpose conversational assistants used for knowledge work and decision support.
- Customer-facing chat and voice agents.
- Agentic and LLM-based tools used in enterprise delivery, such as data-pipeline conversion.

**Testing and evaluation techniques**

- Multi-turn behavioural testing, including whether corrections persist across turns.

- Presentational calibration: whether confidence, fluency and stated corrections match measured accuracy.
- Traceability of results to disclosed model version and configuration.
- Null tests and false-positive rates, so methods do not only find what they look for.

### Failure modes and risk domains

- Unreliable self-report and claimed-but-unmade corrections.
- Effort-based status reporting that masks non-delivery in AI-driven programmes.
- Deployment gaps where a user's flagged error does not reach a human or change behaviour.
- Effects on users' own verification behaviour over repeated use.
- Multilingual and cross-modal consistency (lower priority, observation only): unprompted language switching mid-response, seen twice in our log, in text and in synthesised voice.

### Sectors

- Financial services, where our enterprise case arose, and any sector deploying customer-facing agents at scale.

### Suggested starting point

Correction persistence and self-report reliability. These have the most published groundwork (Huang et al., 2024; SycEval, 2025) and are among the most frequent failure classes in our log, so they are the most tractable first targets for a validated method.

### Q6. Is there anything else you would like to share with us on this topic?

**Disclosure of AI assistance.** This response was drafted with the assistance of Claude (Anthropic), one of the systems whose behaviour it describes. Other assistants (ChatGPT, Microsoft Copilot) were also used to review parts of the underlying analysis. Cognition Observatory's founder, Hans G. Bathija, has reviewed and takes responsibility for all content.

We disclose this because it bears directly on the evidence. Some observations in our log come from interactions with the same systems used to help analyse them. We have tried to manage this conflict of interest by treating a system's own statements as claims to verify rather than as evidence. Our log records instances where we did not do so consistently, which is itself part of the case for independent protocols.

**Status of evidence.** The failure classes described in Part 2A come from a small, largely single-researcher log that has not yet been independently replicated. We would welcome the chance to run the protocols described under controlled conditions through the Centre's programme. A lightly anonymised copy of this response will be published at [cognitionobservatory.org](https://cognitionobservatory.org).

## Part 4 - Closing questions

| Question                                  | Answer                                                                                             |
|-------------------------------------------|----------------------------------------------------------------------------------------------------|
| Q1. Include in published contributor list | Yes                                                                                                |
| Q2. May NPL contact you                   | Yes                                                                                                |
| Q3. Keep informed of outcomes             | Yes                                                                                                |
| Q4. Contact details                       | Hans G. Bathija – <a href="mailto:hans@cognitionobservatory.org">hans@cognitionobservatory.org</a> |

## References and attribution

- Huang et al., "Large Language Models Cannot Self-Correct Reasoning Yet," ICLR 2024 ([arXiv 2310.01798](https://arxiv.org/abs/2310.01798))
- SycEval, 2025 ([arXiv 2502.08177](https://arxiv.org/abs/2502.08177))
- [NPL Call for Evidence – question list \(PDF\)](#)

- [Call for Evidence: AI testing, evaluation and assurance in the UK - NPL](#)

*Claude is AI and can make mistakes. Based on the research documented in this work, mistakes should be expected as a recurring pattern, not an exception. Verify all claims independently.*

*Authored by Hans G. Bathija and Claude (Anthropic), added 30 September 2026*