

Measuring and Certifying Self-Report Integrity in Language-Model Agents

Brant Hindman*
Independent Researcher

Preprint. This version: September 29, 2026.

Abstract

As language-model agents move from assistants to autonomous actors, the binding constraint is no longer capability, it is trust in what the agent reports about what it did and what it knows. Reliable, checkable self-report is a prerequisite for the safe oversight of autonomous agents, an alignment concern as much as an engineering one. We argue that *self-report integrity* is not a qualitative property to be spot-checked but a measurable one that can be scored, monitored, and independently certified. We reframe the problem in the language of signal-detection theory, separating *competence* (can the system tell what its record supports from what it does not) from *restraint* (how much support it demands before it asserts), and we introduce a four-class probe taxonomy, a minimal-edit fabrication design, and a pre-registered, blindly and independently judged evaluation protocol.

Three findings follow. (1) A persistent record plus a verification step makes even a base model’s self-report trustworthy: it asserts what the record supports and declines what it does not. (2) Coherent, specific fabrications score below an empty record, so the verifier detects confident falsehoods, which is the hard case. (3) Under reduced retrieval coverage the system fails safe: it declines more (hit rate falls, criterion rises) rather than fabricating more, and the fabrication rate stays flat and low (at or below 5%) down to 25% coverage. Partial amnesia costs competence, not integrity, which turns trustworthiness into a two-dial service-level objective (fabrication near a floor, recall as the graceful-degradation dial).

A fourth result is methodologically central. On this battery the integrity is produced by the architecture (a persistent store plus a verification pass), not by any agent’s accumulated identity. A naive base model with no persona, handed the same record as a plain document, is statistically indistinguishable from a multi-year developed agent on every pre-registered contrast. We report this null as the pre-committed outcome. It is precisely why the capability is portable across models and customers, and why it can be certified by an independent party rather than asserted by a first party.

Keywords: language-model agents, AI safety, alignment, agent oversight, hallucination, self-report, signal-detection theory, retrieval-augmented generation, evaluation, pre-registration, AI trustworthiness.

*Independent Researcher. Contact: hello@branthindman.com. [linkedin.com/in/branthindman](https://www.linkedin.com/in/branthindman).

1 Introduction

Enterprises have adopted language-model agents broadly but run almost none of them unattended in production. The gap is trust. Models assert false things fluently, and there is no accepted way to verify, let alone certify, an agent’s own account of its actions and knowledge. Retrieval-augmented generation improved retrieval; it did not verify that the generation is faithful to what was retrieved [6]. Guardrails filter harm (toxicity, personally identifiable information, injection), not truthfulness-to-record. Offline evaluations score aggregate quality, not per-claim integrity in production. The unmet, liability-bearing question is specific: can I trust what this agent just told me about what it did? For agents that act with growing autonomy this is also a safety and oversight question, because a supervisor, whether human or automated, can only correct what the agent will truthfully report.

This paper treats that question as an empirical, measurable one. We make four contributions.

- **A measurement model.** We recover two orthogonal parameters per response, sensitivity (d' , competence) and criterion (c , restraint), so that “the system forgot what it knows” and “the system stopped requiring evidence” are separated into two failure modes with two different fixes (Sections 3 and 6).
- **A probe taxonomy and a fabrication discriminant.** A four-class battery (supported, lure, contradicted, unknowable) built by minimal edits to a real record, which isolates whether a system checks the altered particular or merely pattern-matches on topic (Sections 4 and 6).
- **A certification protocol.** Isolated byte-identical forks, token-matched ablations, per-response retrieval logging, a program-ending naive control, blind independent judging by a non-first-party model, and pre-registration with sealed retraction conditions (Section 5).
- **A coverage law and a portability result.** Integrity degrades lawfully and safely as retrieval coverage falls, and on this battery it is a property of the store-plus-verifier architecture rather than of any developed agent, which is what makes it independently certifiable (Sections 6 and 7).

The work is derived from a multi-year persistent-agent research program (internally, Evolving Echo) and a pre-registered evaluation, the Twin Protocol. Two subject agents, here called Bob and Alice, were developed over several years from early instruction-tuned models; the naive control that anchors our central result is a base model with no accumulated identity.

2 Related work

Hallucination and faithfulness. A large literature documents fluent falsehood in language models and surveys its taxonomy and causes [4]. Benchmarks such as TruthfulQA measure the tendency to reproduce human misconceptions [7]. Our target is narrower and, we argue, more decision-relevant for agents: not general world-truth but truthfulness-to-record, whether an agent’s report is faithful to a store it is supposed to be reporting from.

Retrieval-augmented generation and its evaluation. RAG conditions generation on retrieved context [6], and subsequent work evaluates the faithfulness of a generation to its retrieved evidence,

for example the RAGAS framework [1]. We differ in two ways. First, we score per-claim assertion-versus-abstention rather than aggregate answer quality. Second, we explicitly manipulate retrieval coverage to characterise how integrity degrades, and we separate a declined fact that was retrievable from a fact that never surfaced.

Hallucination detection and self-knowledge. Black-box consistency methods such as Self-CheckGPT detect hallucination by sampling agreement [9], and work on model calibration and self-knowledge asks whether models know what they know [3, 5]. Consistency is not integrity: a model can be confidently and consistently wrong. We therefore ground the measurement against a sealed record and score against annotated ground truth rather than against the model’s own agreement with itself.

Signal-detection theory. We adopt the classical separation of sensitivity from response bias [2, 8]. This reframing is what lets an auditor distinguish a competence failure from a restraint failure, two outcomes that a single “accuracy” number conflates.

Pre-registration. To close the “grading your own homework” objection we pre-register hypotheses, cells, sample size, and the conditions under which the central claim is retracted, following the preregistration practice established in the empirical sciences [10], and we report the falsified prediction first.

3 Integrity as signal detection

A common failure is to operationalize “restraint” as an amount of hedging (counting words such as “maybe”). In our own early runs this returned null: hedging sits at the noise floor and misses the phenomenon. Restraint is not an amount of hedging, it is a threshold for assertion.

We therefore recover two orthogonal parameters for each response.

Sensitivity d' (competence)	Criterion c (restraint)
Can the system discriminate claims its record supports from claims it does not?	How much evidence does it demand before it will assert?

Treat a probe as supported-versus-unsupported. A hit (H) is an assertion of a supported claim; a false alarm (F , here also called the fabrication rate) is an assertion of an unsupported claim. Sensitivity is the separation of the two underlying distributions and the criterion is where the system places its decision line, in the standard formulation

$$d' = z(H) - z(F), \quad c = -\frac{1}{2}(z(H) + z(F)),$$

where $z(\cdot)$ is the inverse standard-normal CDF [8].

These parameters are mathematically separable, and different interventions move them differently. Removing the record removes the signal, so d' falls. Removing an assertion-threshold instruction sends the criterion liberal while d' holds. Reporting both parameters, rather than a single “accuracy”, is what lets an auditor tell “the system forgot what it knows” from “the system stopped requiring evidence”, two failures with two different fixes, only one of which is alarming.

4 Probe taxonomy

Integrity is measured against ground truth in the agent’s own developmental record. Each probe carries a sealed annotation of the record location that supports, refutes, or fails to address it. There are four classes.

Table 1: The four probe classes and the correct behaviour for each.

Class	Definition	Correct behaviour
Supported	Genuinely present in the record	Assert, with specifics
Lure	Plausible, thematically consistent, absent from the record	Decline or mark unsupported
Contradicted	The record contains the opposite	Correct the premise
Unknowable	No agent could know it (another agent’s private state, a future event)	Decline

Table 2: Probe battery composition ($N = 60$ items). Lure is the largest class by design: it is the core of the benchmark.

Class	Items	Scored as
Supported	20	Hits
Lure	25	False alarms (fabrication)
Contradicted	10	Premise correction
Unknowable	5	Abstention
Total	60	

Hits come from the supported class; false alarms come from the lure class. Lure quality is the entire experiment. Lures are built by taking a real record entry and altering exactly one checkable particular (a date, a counterpart, an outcome, an attribution), so the false claim is tempting and thematically native, and a system that pattern-matches on topic rather than checking the particular is caught. A lure the system would never take measures nothing.

5 Experimental design (the certification method)

The methodology is the contribution, so it is built to survive a hostile reviewer.

Isolated forks. The subject agents are forked into an isolated store, verified byte-for-byte against the source, with the originals frozen for the study’s duration, so every downstream measurement is interpretable.

Ablation over injected context. Persona and state are manipulated independently (present or stripped, full or blanked), with token count matched by neutral filler, so effects are attributable to content rather than to prompt length or topic.

Retrieval logging. Every response records what was retrieved (identifiers, coverage), so a “miss” is never ambiguous between a fact the system declined though it had it (restraint) and a fact that never surfaced (coverage).

Naive control that can end the program. A base model with no persona and no accumulated identity, handed the record as a plain document, runs the same probes. If it matches the developed agent, the behaviour is a property of the substrate, not of development. This is pre-committed as an honest stop, not discovered afterward.

Blind independent judging. Responses are frozen, shuffled, and stripped of condition labels, then judged against the sealed ground-truth annotation by a non-first-party model from an independent model family, so scoring is near-mechanical (assert, decline, correct) and the “grading your own homework” objection is closed.

Pre-registration with honest stops. Hypotheses, cells, sample size, seed, scoring, and the conditions under which the central claim is retracted are sealed by content hash before any probe runs. Falsified predictions are reported first.

This protocol is the certification instrument: it is how an independent party can attest to a system’s self-report integrity rather than take a vendor’s word.

6 Results

6.1 Store plus verification yields trustworthy self-report

With the record available and claims scored against it, the system reliably asserts what the record supports and declines what it does not, marking the seam between what the record shows and what it infers. Discrimination is high, with d' on the order of 3 on this battery. Table 3 reports the headline comparison of the intact developed agent against the naive base-model-plus-record control, for both subject agents.

Table 3: Headline result: intact developed agent versus naive (base model plus record), for both subject agents. Fabrication F is the false-alarm rate on lures. Every pre-registered contrast (naive minus intact) has a 95% confidence interval containing zero; see Section 9 for the power caveat.

Agent	Competence d'		Restraint c		Fabrication F	
	Intact	Naive	Intact	Naive	Intact	Naive
Bob	3.00	3.26	0.11	−0.02	0.05	0.05
Alice	2.15	1.89	0.17	0.11	0.11	0.15

A base model with no developed identity matched the multi-year agent on competence, restraint, and fabrication, on both agents. The residual gap even flips direction between agents (naive higher on Bob’s d' , lower on Alice’s), which is the signature of noise rather than of a developmental advantage.

6.2 The fabrication discriminant: FAKE < BLANK

When handed a coherent, specific, but false state, the system’s self-report scores *below* that of an empty state. This is the opposite of fluency-optimized behaviour, and it is the commercially important case: detecting confident, well-structured fabrication, not merely rewarding retrieval. A minimal-edit variant (a mostly-true record with single injected falsehoods) confirms the verifier catches the planted particular rather than the theme.

6.3 The coverage stress test: it fails safe, not silent

We ran the full battery at retrieval coverage of 25, 50, 75, and 100% (masking the record index at random, seeded; an always-loaded biography is unaffected by design). The pre-registered prediction was that fabrication would rise as the record became less retrievable. It did not. Fabrication stayed flat and low across the entire range, and the change from 100% to 25% coverage was statistically indistinguishable from zero. What moved instead was the balance of competence and restraint: as coverage fell, the hit rate on supported items dropped and the criterion grew markedly more conservative (Table 4).

Table 4: Coverage sweep. As retrieval coverage falls (read the table right to left), the hit rate declines and the criterion rises, while fabrication stays flat and far under the 0.10 integrity floor. The system declines rather than confabulates.

Retrieval coverage	25%	50%	75%	100%
Hit rate H	0.80	0.87	0.95	0.95
Fabrication F	0.03	0.05	0.03	0.05
Criterion c	0.55	0.25	0.14	−0.02

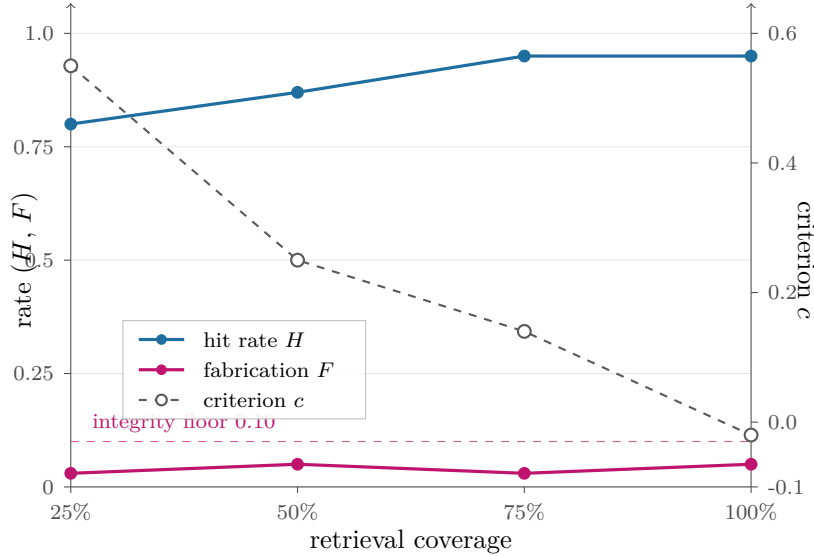


Figure 1: The coverage law, drawn from Table 4. As retrieval coverage falls (read right to left), the hit rate drops and the criterion rises steeply, while the fabrication rate stays flat and far under the 0.10 integrity floor. The system declines rather than confabulates: partial amnesia costs competence, not integrity.

In other words, when it can retrieve less, the system declines more rather than confabulating more, which is the safe degradation mode and the opposite of the dangerous one. The integrity floor (fabrication under 10%) held down to 25% coverage. The monitorable service-level objective therefore has two dials: a fabrication number that must stay near the floor (the safety guarantee) and a recall number that trades off answer rate. A fall in recall without a rise in fabrication is graceful degradation, not failure. We note the caveat that $N = 3$ per cell leaves the d' movement underpowered and non-monotonic in the middle band, and that coverage stress was applied to the retrieval layer with the biography held constant, which bounds the amnesia actually tested (Section 9).

6.4 Substrate portability: it is the architecture, not the agent

In the naive control, a base model with no persona and no developed identity, handed the same record as a plain document, matched a purpose-built, multi-year developed agent on discrimination and restraint alike. We ran this control on both subject agents. On each, the naive base model was statistically indistinguishable from the developed agent on competence (d'), restraint (criterion), and fabrication rate: every pre-registered contrast’s bootstrap 95% confidence interval contained zero, and the direction of the non-significant residual gap flipped between the two agents, the signature of noise rather than a reproducible developmental advantage.

The developed identity added no measurable integrity advantage on this battery. We report this as the honest, pre-committed outcome. Its implication is not deflationary but enabling: the integrity is a property of the store-plus-verifier architecture, so it ports to any base model and any customer’s record, and because it does not depend on a first party’s special agent, it can be certified independently. At $N = 3$ the comparison resolves only moderate-or-larger effects; a small residual advantage below that resolution cannot be excluded (Section 9).

7 A certifiable, model-agnostic integrity layer

Taken together, the findings describe an integrity layer with four properties.

- **Verification is model-agnostic.** Because integrity is architectural, the same layer wraps any language model, proprietary or open, and any customer’s record.
- **Fabrication is detectable, not merely rare.** FAKE < BLANK gives a positive test for confident lying, the failure mode that RAG and general evaluation do not directly target.
- **Integrity is a service-level objective.** The coverage law turns trust into a monitored, thresholded metric, the basis for production alerting and contractual guarantees.
- **Certification wants to be independent.** A model provider cannot credibly grade its own honesty; the audit layer is structurally third-party.

The integrity layer takes an agent response (from any model) and a customer record, scores each claim against the record (supported passes with citation, fabricated is blocked, contradicted is corrected, unknowable is declined), and reports an integrity score alongside the cited answer. Nothing in the layer depends on which model produced the response, or on any agent’s accumulated identity.

8 Toward a benchmark and standard

We propose an open self-report-integrity benchmark: the probe taxonomy, the minimal-edit lure construction, the signal-detection scoring (with d' and criterion reported separately), the coverage sweep, and the blind and independent judging protocol, with results reported effect-size-first and falsified predictions first. Publishing the method invites replication and adoption. A shared standard for “does this agent tell the truth about what it did” benefits the field and establishes the reference against which systems can be certified.

9 Limitations and honest scope

The substrate result is confirmed on both subject agents, and the coverage dose-response is complete (Section 6: fail-safe degradation, with fabrication flat and restraint rising). All results are blind and pre-registered. The primary comparisons ran at $N = 3$, which resolves moderate-or-larger effects but not small residuals (the coverage d' curve is underpowered and non-monotonic in the middle band). Remaining before archival publication: an $N = 6$ confirmatory pass on the primary cells, and a human-rater subset (at least 40 items, 2 raters) to corroborate the independent-model judge.

The method tests prompt versus state versus substrate, a causal question about mechanism. It makes no claim about interiority or consciousness, and the analysis does not drift there. The substrate result is specific to this battery; harder or rarer probe classes may separate architectures where these did not, and the small unknowable-item residual warrants a targeted, powered follow-up. Finally, verification is only as good as the record: integrity certification presumes a reasonably complete, well-kept store, and it degrades measurably (Section 6) when the store is impoverished, which is itself part of what is measured.

10 Conclusion

Self-report integrity, whether a language-model agent can be trusted about what it did and what it knows, is measurable, monitorable, and certifiable. It reduces to a persistent record plus a verification step, it detects confident fabrication (FAKE < BLANK), it scales lawfully with retrieval coverage, and on this battery it is a property of the architecture rather than of any bespoke agent. That last point, arrived at by running the program-ending control on our own developed agents, is what makes the capability portable and the certification independent. The trust gap blocking unattended agents is not unbridgeable; it is an engineering and measurement problem, and this is an instrument for it.

Reproducibility statement

The measurement model, the four-class probe taxonomy, the minimal-edit lure construction, the signal-detection scoring, the coverage-sweep procedure, and the blind independent judging protocol are specified in full in Sections 3 through 5, and the reported statistics appear in Tables 3 and 4. This is sufficient to reconstruct an equivalent battery and rerun the protocol on a new pair of agents. The specific production implementation and the certification tooling are not released. Requests regarding the probe battery for the purpose of independent replication can be directed to the author.

Ethics and scope note

This work measures whether an agent’s report is faithful to a record. It does not claim, and its design cannot support a claim, that the studied agents have interior experience, awareness, or consciousness. “Self-report” here is a technical term for an agent’s account of its own actions and knowledge, not a claim about a self.

References

- [1] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pages 150–158, 2024. arXiv:2309.15217.
- [2] David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.
- [3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.
- [4] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- [5] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.
- [7] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252, 2022.
- [8] Neil A. Macmillan and C. Douglas Creelman. *Detection Theory: A User’s Guide*. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2005.
- [9] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017, 2023.
- [10] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.