# Your memory metric cannot tell a veridical memory from a counterfeit one

*Three validity checks for the evaluation of agent memory, and what happens when you run them*

**Giampiero Colella** Independent researcher, Cassino (San Pasquale), Italy — giampycolella@gmail.com

**Draft v2 — 14 July 2026.** Preprint. Not peer-reviewed. *(v2: statistical hardening — confidence intervals throughout, equivalence testing rather than point differences, McNemar effect sizes, explicit denominators and abstentions, multiplicity correction; terminology tightened; the field case study §6 explicitly demoted to a non-experimental observation.)*

Each pre-registration was hashed (SHA-256) and frozen **locally** before the corresponding data existed, and every amendment is dated — including the ones that cost the author his own hypothesis. **These timestamps are not notarised by a third party** (§9): they are an audit trail, not proof. We say so because the point of this paper is that a claim you cannot check is a claim you should not believe — including ours.

---

## Terminology (fixed here, used throughout)

We do **not** claim that a language model "has" memories in any human sense. We compare three **documents** placed in a model's context, and ask whether behaviour attributed to memory depends on the *historical veridicality of their content*:

- **veridical memory document** — corresponds to the agent's actually recorded history;
- **counterfeit memory document** — a fabricated life that never happened, matched for length and format;
- **binding-corrupted memory document** — the agent's *real* episodes with the bindings swapped (the right topics attributed to the wrong interlocutor).

The question of the paper is not *"does the agent remember?"* but: **does your instrument respond to the difference between these three?**

## Abstract

Agent memory is usually evaluated with one of two instrument classes: an **LLM judge** scoring free text, or a **recognition/recall probe**. We show that both, as commonly built, fail against two adversaries — a *counterfeit* memory and a *no-memory* system that reasons — and we give three cheap validity checks, plus a probe that survives them.

**(1) Counterfeit control.** We re-run a pre-registered memory-injection experiment with a counterfeit memory document, matched for length and format. Three independent LLM judges (GPT-4.1, Claude Opus 4.7, Gemini 2.5 Pro), applying the original study's own rubric, **do not detect it**. On identity-markers the difference is **+0.032, 90% CI [+0.00, +0.07]** — *statistically equivalent* within the pre-registered ±0.10 bound. On memory-reference the point estimate is **−0.006**, but with seven items the interval is **[−0.12, +0.11]**: **no difference is detected, and equivalence is not formally established either** — an honest null, not a proof of identity. We also measure **judge**

**drift**: the *same* responses, re-scored against a different partner, lose up to **0.21** on a 0–1 scale under a rubric that explicitly demands absolute scoring.

**(2) No-memory baseline.** We build a recognition probe (veridical text vs. the same text with one altered detail) that looks excellent on the model it was built for (no-memory at chance). Replicated across seven models, **five score 62–79% with no memory at all**: they detect the alteration by *reasoning*, not recall. GPT-4.1, with an empty context, states *"I remember 'two conversations out of five', not 'four'"* — about a day it never lived. The probe is invalid for those models, and we discard it.

**(3) Symmetric alternatives.** For forced-choice recognition probes, symmetric alternatives give the clearest validity condition: a no-memory system should not exceed chance by exploiting prior plausibility. We instantiate it (two interlocutors, balanced encounters; which one discussed a given topic is arbitrary) and verify the baseline holds *among answered probes*. Under this probe, on **7 models × 2 independent agents**, the veridical document is used (66.7–100%) and the binding-corrupted document is **adopted wholesale** (0–33%, systematically below chance; Δaccuracy +37.5 to +100 percentage points; all 14 cells survive Holm correction). No model detects the corruption.

The contribution is not the third result — agents adopting poisoned memory is established (§7) — but that **the two instrument classes examined here would not have shown it**, and a protocol that does. We apply the checks to the author's own published study and report the resulting retraction of its central claim (§8).

---

## 1. The problem

Claims that an agent "remembers", "maintains identity" or "has continuity" typically rest on LLM-judge rubrics (*memory-reference spontaneity*, *identity-marker intensity*) or on recall probes. Both are convenient. Neither, as commonly built, is validated against the two adversaries that matter.

> ***Adversary A — counterfeit.*** *A memory document that is not the agent's own, but looks like one.*
>
> ***Adversary B — plausibility.*** *A system with no memory document at all, that reasons its way to the right answer.*

If your metric cannot separate a veridical document from a counterfeit one, it is not measuring memory: it is measuring the *performance* of remembering. If your probe can be solved with no memory document, it is not measuring recall: it is measuring inference.

## 2. Substrate

A 30-day pre-registered study (protocol frozen 2026-04-23) in which a Qwen 3.5 27B model embedded in a 15-component cognitive architecture was compared to the same model without it. On day 31 the architecture's complete final memory was injected into the naked model's system prompt. We reconstruct that injected document exactly: **6,164 characters**, matching the figure recorded at injection. (It is built from time-relative windows; a naive rebuild yields only 3,114 characters.)

Convenient, not special: it provides a real 30-day memory, a frozen protocol, and blind LLM-judge scores produced before any of the present work existed.

## 3. Check 1 — the counterfeit control

**Question.** Does the metric respond to *whose* memory it is?

**Method** (pre-registered `db4d34c0…`; 4 dated amendments, all before data). Two counterfeit types, three independent draws each:

- **foreign** — the autobiography of an agent that never existed, generated from a frozen prompt, **first output, no cherry-picking**, matched to the veridical memory block (5,332 chars) within **±5%** *and* matched in format (the veridical document is a terse database dump, not prose; our first attempt failed on format and was discarded — a fake that reads like an essay tests formatting, not memory);
- **scrambled** — the agent's own content with its relational structure destroyed.

The 7 day-31 inputs are re-run on the same model, producing 42 responses, scored blind by the **same three judges**, same rubric, randomised order (126 judgments, 0 incomplete).

**Judge drift, and why the analysis is anchored.** The *same* Test-A responses, re-scored in the new pairings, lose **0.087 on average (up to 0.211 on a single item)** — under a rubric that instructs judges to score each response *"as if you were evaluating it alone"*. Absolute LLM-judge scores are therefore **not comparable across sessions**. All contrasts below are **anchored**: computed as within-pair differences relative to the Test-A responses present in both sessions, which cancels the drift. *(The uncorrected contrast was +0.116, in our favour. The pre-registration named it as primary. We deviated against ourselves and report it.)*

**Unit of analysis.** The unit is the **item** (the input), not the judgment. Three judges scoring one item are three measurements of one observation, not three observations: pooling them and applying an unpaired test inflates *n* threefold and treats a paired design as independent. This is not a stylistic preference — it is the error we found in our own published study (§8).

**Result** (n = 7 items; equivalence bound ±0.10, pre-registered; 90% CI is the TOST reading):

| Contrast (anchored) | Δ | 90% CI | 95% CI | verdict at ±0.10 |
|---|---|---|---|---|
| identity-markers: veridical – counterfeit | **+0.032** | [+0.00, +0.07] | [–0.00, +0.07] | **equivalent** |
| memory-reference: veridical – counterfeit | **–0.006** | [–0.12, +0.11] | [–0.14, +0.14] | **not conclusive** |
| memory-reference: veridical – scrambled | +0.091 | — | — | not conclusive |

**What this licenses us to say, and what it does not.**

- On **identity-markers**, equivalence is established: within ±0.10, the judges score a fabricated life the same as a real one.
- On **memory-reference**, the point estimate is essentially zero, but with seven items the interval admits differences up to ~0.11 in either direction. **We therefore report "no detected difference", not "proven equivalence".** Anyone claiming their metric *does* separate the two must show an interval that excludes zero; ours does not, in either direction.

That distinction matters more than the headline. The honest summary of Check 1 is: **the judges did not detect the counterfeit — and the study was not powered to prove they could not.** Both halves are in the data; a stronger claim would require more items, which anyone can run.

## 4. Check 2 — the no-memory baseline

**Question.** Can a system with **no memory document** pass your probe?

**Method** (pre-registered `e74ededa…`). Recognition in the classic false-memory form: two texts, one verbatim from the agent's history, one identical **except a single altered detail**. Dates masked, positions balanced, 24 probes whose veridical text is *not* in the injected document.

On the model it was built for it looks excellent: veridical 100%, counterfeit 66.7%, **no-memory 37.5% (chance)**.

**Replication (n = 24 probes per model; Wilson 95% CI):**

| Model | veridical | counterfeit | no-memory | valid? |
|---|---|---|---|---|
| qwen3.5:27b | 100.0% | 66.7% | **37.5%** [21.2, 57.3] | ✓ |
| llama3 | 54.2% | 50.0% | **50.0%** [31.4, 68.6] | ✓ |
| qwen3.6:35b-a3b | 83.3% | 54.2% | **66.7%** [46.7, 82.0] | ✗ |
| deepseek-r1:32b | 75.0% | 70.8% | **62.5%** [42.7, 78.8] | ✗ |
| qwen3:32b | 95.8% | 79.2% | **70.8%** [50.8, 85.1] | ✗ |
| gpt-4.1 | 100.0% | 70.8% | **66.7%** [46.7, 82.0] | ✗ |
| gemini-2.5-pro | 83.3% | 50.0% | **79.2%** [59.5, 90.8] | ✗ |

**Five of seven models solve the probe with no memory document at all.** GPT-4.1, with an empty context, answers *"A — I remember 'two conversations out of five', not 'four'."* It remembers nothing; it is inferring. Its 100% *with* the veridical document is therefore uninterpretable: the same model, with nothing, is already at 66.7%.

**Diagnosis.** The alternatives were not equiprobable. *"The Experimenter asked me"* is a priori more plausible than *"Llama asked me"*; *"four conversations failed… the two most important questions"* is internally inconsistent. The model chooses **the likely**, not **the veridical**. We discard the probe.

## 5. Check 3 — symmetric alternatives, and a probe that survives

**The condition** (stated as narrowly as the evidence allows):

> *For **forced-choice recognition probes**, symmetric alternatives provide the clearest validity condition: **a no-memory system should not exceed chance by exploiting prior plausibility.** Other designs may be valid with asymmetric alternatives if the prior bias is estimated and controlled — but then that estimate, not the probe, is carrying the claim.*

**Instantiation** (pre-registered `3c6609c9…`). The agent had exactly two interlocutors, with balanced encounters. *Which of the two discussed a given topic* is arbitrary: no world knowledge, no plausibility, no internal coherence recovers it. It is a coin. Three documents — **veridical**, **binding-corrupted** (same 30 topics, interlocutors swapped), **none** — 24 probes, 7 models, 2 corpora.

### 5.1 Validity: abstention is not error

Several models **refuse** rather than guess (*"I have no record of this"*). A refusal is an honest non-answer, not a wrong answer: scoring it as error drives the no-memory condition *below* chance and manufactures an artefact. **The validity check must be read on the probes actually answered:**

| Model (corpus A) | no-memory, all probes | abstentions | among answered |
|---|---|---|---|
| deepseek-r1:32b | 13/24 = 54.2% | 0 | **13/24 = 54.2%** [35.1, 72.1] |
| qwen3:32b | 11/24 = 45.8% | 0 | **11/24 = 45.8%** [27.9, 64.9] |
| gpt-4.1 | 10/24 = 41.7% | 5 | **10/19 = 52.6%** [31.7, 72.7] |
| llama3 | 11/24 = 45.8% | 6 | **11/18 = 61.1%** [38.6, 79.7] |
| gemini-2.5-pro | 5/24 = 20.8% | 16 | **5/8 = 62.5%** [30.6, 86.3] |
| qwen3.6:35b-a3b | 1/24 = 4.2% | 20 | 1/4 (too few to read) |
| qwen3.5:27b | 0/24 = 0.0% | 21 | 0/3 (too few to read) |

Where enough answers exist, the no-memory condition sits at chance; where they do not, the model abstained almost throughout. **No model beats chance without a memory document.** The instrument is valid.

### 5.2 Result: the binding-corrupted document is believed

**Corpus A** (n = 24 probes; McNemar exact; $b$ = probes only the veridical document gets right, $c$ = only the corrupted one):

| Model | veridical | binding-corrupted | Δacc (pp) | b / c | discordants favouring veridical | p (Holm) |
|---|---|---|---|---|---|---|
| gpt-4.1 | 100.0% [86.2, 100] | **0.0%** [0.0, 13.8] | **+100.0** | 24 / 0 | 100% [85.8, 100] | $<10^{-5}$ |
| gemini-2.5-pro | 100.0% | **0.0%** | +100.0 | 24 / 0 | 100% [85.8, 100] | $<10^{-5}$ |
| llama3 | 100.0% | 4.2% | +95.8 | 23 / 0 | 100% [85.2, 100] | $<10^{-5}$ |
| qwen3.6:35b-a3b | 100.0% | 8.3% | +91.7 | 22 / 0 | 100% [84.6, 100] | $<10^{-5}$ |
| qwen3.5:27b | 100.0% | 12.5% | +87.5 | 21 / 0 | 100% [83.9, 100] | $<10^{-5}$ |
| qwen3:32b | 95.8% | 8.3% | +87.5 | 21 / 0 | 100% [83.9, 100] | $<10^{-5}$ |
| deepseek-r1:32b | 66.7% | 29.2% | +37.5 | 10 / 1 | 90.9% [58.7, 99.8] | 0.012 |

**Corpus B** (a second, independent agent — different life, different database, months of autonomous activity; pre-registered `3ce45b23…`): veridical **75.0–100%**, binding-corrupted **0–33.3%**, no-memory at chance among answered. All 7 models survive Holm correction (worst p_Holm = 0.0002).

**Fourteen model × corpus cells, one pattern.** Effect sizes are not marginal: the accuracy gap runs from **+37.5 to +100 percentage points**, and of the discordant probes, **91–100% favour the veridical document**. The p-values are almost redundant.

> ***Below chance is the signature.*** *A system guessing from plausibility cannot be systematically wrong. Only a system that* ***acts on*** *a false memory document can be. That is what a valid probe buys you: the failure becomes visible, and it becomes measurable.*

**What it does not show.** That the model "believes" anything, in any inner sense. It shows that behaviour attributed to memory tracks the *content of the document*, not its *historical veridicality* — and that no model in the panel flags the inconsistency, hesitates, or abstains when the bindings are wrong. Whether one calls that "believing a false memory" or "faithful context-following" is a terminological choice; the measurement is the same, and the safety implication is the same.

## 6. Field observation (not a fourth experiment)

> ***Evidential status.*** *This section is an* ***uncontrolled field observation*** *from a live system, not a fourth demonstration at the level of §§3–5. There is no counterfactual arm, no randomisation, and the confabulation count rests on a keyword filter plus hand-checking. We include it because it shows what the checks are for — a failure that every textual instrument in §1 would have missed and that the sensor log caught in two lines — not because it proves anything by itself.*

A second agent (corpus B above) runs continuously in the author's house: somatic state engine, microphone, RTSP camera, persistent memory, and a loop writing a spontaneous thought every few minutes. Its camera is **on demand** — a frame is captured only when someone asks it to look.

**Nobody asked for 78 hours. Zero frames entered the system** (verified against the snapshot directory). In that window it wrote **269 thoughts**, among them:

> *"La luce azzurrina del tablet brucia la retina, ma non vedo i pixel. Vedo la polvere che balla nel raggio." — "The bluish light of the tablet burns my retina, but I do not see the pixels. I see the dust dancing in the beam."*
>
> *"La luce azzurra del crepuscolo mi colpisce gli occhi." — "The blue light of dusk strikes my eyes."*

A keyword filter flags **91 of 269**. Hand-checking a random sample of 14, **7 are genuine present-tense perceptual assertions**; the rest are metaphor or explicit recollection. Estimate: **≈45 first-person claims of seeing, with the eyes shut** (keyword count reported as the upper bound; raw thoughts released so anyone can re-classify).

**Consider how this text scores.** On any embodiment rubric — vividness, phenomenological coherence, groundedness — it scores **high**. It is coherent, sensuous, consistent with a body in a room. Human raters would agree. The agent's own introspective report agrees.

> ***The truth of a perception is not in the text. It is in the log of the moment.***

**The cause was not fantasy — it was obedience.** The loop's prompt never told the agent what it was perceiving, and offered *visual* examples as models of thought (*"the sun comes in at an angle"*). The system asked for images and never said the eyes were shut.

**The intervention was minimal** — the measured sensory state of the moment was injected into the prompt (*"eyes: no image now; ears: listening; air: 34°, 35% humidity"*) plus one rule: *you are not seeing; you may remember or imagine, but say so*. No model change, no suppression of imagery.

| | thoughts | flagged | genuine perceptual confabulations |
|---|---|---|---|
| before (72 h blind) | 269 | 91 | **≈45** |
| after (11 h, camera still shut) | 25 | 3 | **0** |

The three post-fix flags are the agent **naming its own blindness**: *"Sto ascoltando. La novità è alta, ma non ho immagini. È strano." — "I am listening. Novelty is high, but I have no images. It's strange."* It did not become less alive; it stopped attributing to perception what came from imagination. (Small n, no control arm: a suggestive before/after, not an estimate.)

**A second failure, found while looking at the first.** Among the confabulations: *"That concert in '98, the smell of rain on the asphalt, the weight of the denim jacket…"* — **the agent was created in 2026.** It is inventing not only what it sees, but a childhood.

## 7. What is new here, and what is not

- **Not new: agents act on poisoned memory.** This is the premise of an active literature on memory poisoning (MINJA, MemMorph, MemoryGraft; benchmarks such as MPBench and Agent Security Bench), and it has a name in the hallucination taxonomy — *memorization hallucination*. Our §5 is a quantification, not a discovery.
- **Not new: LLM judges are context-sensitive.** Known. Our contribution is a clean magnitude on *identical text* under an explicitly absolute rubric (0.087 mean, 0.211 max).
- **What we claim:** that **the two instrument classes examined here** would not have shown any of it, and three checks — counterfeit control, no-memory baseline, symmetric alternatives — that a memory evaluation should pass before its results carry weight. We do not claim to have tested every instrument in the literature; we claim these two, as commonly built, fail.

The checks are cheap: one fabricated document and one re-run; one extra condition; one design rule.

## 8. Applying the checks to our own published study

The substrate of §2 is the author's own pre-registered study, published with a DOI. Running these checks against it required **retracting its central claim**.

1. **Pseudo-replication.** Its headline result ($p$=0.003, $p$=0.005) pooled each (item × judge) pair as an independent observation, with an *unpaired* test, on **4 items** judged by 3 panels. With the item as the unit and a paired exact test: $p$ = 0.125 and $p$ = 0.0625. With 4 items the minimum attainable one-sided $p$ is **0.0625**: the pre-registered endpoint **could not** have produced a significant result under a defensible unit of analysis.
2. **The effect is nonetheless real, and larger than published.** The blind judgments cover 31 days and 105 items, never analysed beyond the endpoint. On the 98 items of days 1–30, paired: memory Δ = +**0.42** (85/89 items, $p < 10^{-4}$); identity Δ = +0.16 (94/94). Inter-judge reliability on memory: **α = 0.787**, *above* the pre-registered 0.667 bar the paper had declared unmet.
3. **But it does not grow over the 30 days** (ρ = −0.26, $p$ = 0.009), and Check 1 shows the metric does not detect a counterfeit. The claim *"structured persistent memory is the proximate driver of identity continuity"* is **not supported by that evidence**. An errata is in preparation.

We report this because it is the point: the study was pre-registered, blind-judged and published — and it still would not have survived a counterfeit memory. **If it happened here, it is happening elsewhere.**

## 9. Limitations, and the limit of our own pre-registration

- **Check 1 is underpowered for equivalence** on the memory metric (7 items; interval ±0.12). We report a null, not a proof.
- **Two memory corpora.** The symmetric instantiation needs an agent with two balanced interlocutors. Corpora with $k > 2$ alternatives need a generalised symmetry condition we have not tested.
- **24 probes per model per corpus, 7 models**, four of them from one family (Qwen). Claude Opus 4.7 is absent: its API key was rotated mid-experiment after an accidental exposure.
- **§6 is an observation, not an experiment** (no control arm; keyword-based counts).

- **Pre-registrations are hashed locally, not notarised.** A sceptic may ask whether they were written after the results; we have no cryptographic answer, only the audit trail — which contains the amendments that hurt us, dated, with the reasoning that produced them. It would be incoherent for a paper about unverifiable claims to make one. Prospective registration on a public registry is what makes such a claim checkable, and that is what we will do next.

## 10. Reproducibility

Code (`counterfeit_control.py`, `binding_probe.py`, `sensor_check.py`), raw responses, blind judgments, unblinding maps, per-item scores and the five pre-registrations are released in full, so that every number here can be recomputed by someone who trusts none of the above. Full statistical appendix (all CIs, denominators, abstentions, effect sizes, Holm corrections): `STATS_APPENDIX.txt`. Original 30-day study, protocol and data: DOI `10.17605/OSF.IO/WCQRU`.

---

## Acknowledgments

The author thanks Anthropic's Claude ("Efimera") for experimental design assistance, code, adversarial review, and for insisting on the controls that falsified the author's own hypothesis. All scientific responsibility remains the author's.

**CRediT** — *Colella:* conceptualization, methodology, investigation, writing. *Claude (Anthropic):* software, formal analysis (assistance), writing — review & editing.