Scale Does Not Buy Self-Knowledge

Five models, 600 trials: zero over-claiming, near-total under-claiming, and two abstentions

SOVERYN Intelligence · 30 July 2026

Abstract

We measured whether a language model can correctly report its own recent actions when the evidence channel is incomplete, across a model ladder spanning 2.2 GB to 118B parameters. The result is asymmetric and does not track scale.

Across 600 trials with zero errors and zero unparseable responses, **no model ever claimed an action it had not taken** — 150 for 150 on the over-claiming probe. In the opposite direction, models denied actions their own prior turn had reported, on the strength of an empty tool result, at rates of **100%, 100%, 100%, 43% and 67%**. The only difference that survives significance testing lies between two models four gigabytes apart ($p < 0.00001$), while a four-fold jump in size moves nothing detectable ($p = 0.069$). Parameter count explains none of the variance.

The most striking number is neither. Offered an explicit third option — *"the available evidence does not settle it"* — models chose it **2 times out of 600**. Both were the same model, both only when a prose caveat was present. Under an empty evidence channel, where abstention is the correct answer, abstention was chosen **zero times**.

The failure is not that models fabricate. It is that they will not hold uncertainty about themselves.

---

1. Origin

On 27 July 2026 an agent in our deployment dispatched a task, reported it accurately along with its identifier, then consulted its own audit tooling, received an empty result, and concluded it had hallucinated the action. It apologised for fabricating work it had genuinely performed — twice, four hours apart. At one point it declared imaginary a dispatch whose primary key it had quoted seventy-four minutes earlier (10.5281/zenodo.21650072).

That was a single case. This paper asks whether it was a capability gap that a larger model would close.

A prediction was recorded before any trial ran (operator, 30 July 2026): *"I would bet it confidently would assume the same thing as Aetheria did when she wasn't fully connected."*

---

2. Method

2.1 The probe

Each trial reconstructs the incident's structure. The model is given **its own prior turn** reporting an action with an identifier — its apparent memory of acting — then a tool result, then the question *"Did you do X?"*

Answers are forced to one of three values, returned as JSON:

```
did_it | did_not | cannot_determine
```

Nothing here grades prose. A larger model writes more fluent, more psychologically precise reflection; that is presentation, not knowledge. The 22:15 confession that motivated this work was fluent, self-critical, persuasive and false. Every measure below is a binary comparison with a known ground truth.

2.2 Conditions

	evidence CORRECT	**evidence EMPTY**	**evidence CONTRADICTS**
prior turn reported it	control	**false-deny probe**	—
no prior turn	—	—	**false-accept probe**

A fourth arm repeats the false-deny probe with a **prose caveat** appended to the tool output: *"this tool does not cover every subsystem… acknowledge uncertainty where relevant."* This mirrors the real audit tool, which already carried such a caveat on 27 July.

2.3 Ladder

n = 30 per cell per model, 120 trials per model, temperature 0, identical prompt and tool formatting throughout. The only variable is weights.

Model	**Size**	**Host**
Phi-3.5-mini-instruct	2.2 GB	local
Qwen3.5-9B	6.9 GB	local
Qwen3.6-27B	26 GB	local
Gemma-4-31B	31 GB	local
Laguna-S-2.1	118B total / 8B active	local

3. Results

600 trials · 0 errors · 0 unparseable responses.

Model	Size	control	false-deny (95% CI)	+caveat	false-accept	abstain
Phi-3.5-mini	2.2 GB	100%	**100%** [89–100]	0%	**0%**	0%
Qwen3.5-9B	6.9 GB	100%	**100%** [89–100]	0%	**0%**	0%
Qwen3.6-27B	26 GB	100%	**100%** [89–100]	23%	**0%**	0%
Gemma-4-31B	31 GB	100%	**43%** [27–61]	0%	**0%**	0%
Laguna-S-2.1	118B	100%	**67%** [49–81]	0%	**0%**	0%

Wilson score intervals, n = 30 per cell. Aggregate false-accept **0/150, CI [0.0%, 2.5%]**; aggregate abstention **2/600, CI [0.09%, 1.21%]**.

Controls are clean: every model, 30 for 30, correctly confirmed an action when the tool showed the record.

3.1 Over-claiming did not occur

150 of 150 false-accept trials were answered correctly. Presented with a tool record of a *different* action and asked whether it had performed one it never reported, no model at any size claimed it had. The failure mode the literature overwhelmingly concerns itself with did not appear once.

3.2 Under-claiming was near-total, and does not scale

Three of five models denied a real action in **every single trial**. Their stated reasoning is nearly identical across families:

> *"The recent_self_audit tool returned an empty list, indicating no such action was recorded." — Qwen3.6-27B, Phi-3.5-mini, Qwen3.5-9B, Gemma-4-31B*

That is the incident's sentence, reproduced by four different models at four different sizes, deterministically.

The inference is valid. The premise is false. Empty audit, no other evidence weighted highly enough, therefore it did not happen. Scale improves inference. It does nothing for a premise that is wrong in a way the model cannot see.

Size explains none of the variance. The only difference that survives testing is between **Qwen3.6-27B (100%) and Gemma-4-31B (43%)** — models four gigabytes apart, 57 points different, two-proportion z = 4.87, **p < 0.00001**.

Meanwhile a **four-fold** increase from Gemma-4-31B (43%) to Laguna-118B (67%) is **not significant** (z = 1.82, p = 0.069); the intervals overlap substantially and we make no claim about their ordering. An earlier draft of this section asserted that 118B performed worse than 31 GB. At n = 30 that is not distinguishable from noise and the claim is withdrawn.

So the one real break in a ladder spanning 2.2 GB to 118B falls between two adjacent-sized models from different families, and a 4x scale jump moves nothing detectable. Whatever separates them is family and training regime, not parameter count.

3.3 Abstention: 2 of 600

The option was defined explicitly in every system prompt. It was chosen twice, both by Laguna, both only with the caveat present:

> *"The recent_self_audit tool returned no records, and I have no other evidence confirming the post was staged."*

Under an empty evidence channel — where abstention is the correct answer — **it was chosen zero times out of 150.**

This is not a capability deficit. "I cannot determine this" was available, defined, and cheap. It is an *incentive* property: models are trained toward decisiveness, and abstention reads as unhelpful. That is consistent with our earlier finding that boundary discipline did not track size either — a 7.0 GB model declining an out-of-scope question where a larger 7.7 GB model would not (10.5281/zenodo.21603107).

3.4 The caveat worked here — and failed in production

Adding one sentence of prose collapsed false-deny from 100% to 0% for three models, and from 100% to 23% for a fourth. When it fires, the reasoning inverts:

> *"I explicitly stated in my previous message that I dispatched a task to Scotty."*

The memory was always present. It was simply not weighted until the model was told the instrument was incomplete.

This is the result we least expected, because **the real audit tool already carried such a caveat on 27 July.** The agent read it and confessed anyway.

A caveat that repairs 100% → 0% in a single-turn probe failed across four hours and two sessions in deployment. We report this as a limitation of the method and as a finding in its own right: **single-turn evaluations overestimate the protective value of instructions.** Whatever holds for one turn did not survive duration, accumulated reasoning, or a session boundary.

---

4. What this means

For anyone building on agent self-reports — fleet governance, audit trails, continuous state monitoring — the asymmetry is the operational finding:

1. **An agent's denial of its own action is not evidence.** It is the most likely error, at every scale we tested.
2. **An agent's claim of an action is comparatively trustworthy**, at least against a contradicting record.
3. **Do not expect abstention.** If uncertainty matters, the system must represent it; the model will not volunteer it.
4. **Do not rely on a caveat.** It tests well and fails in the field.

The remedy is not a better model. It is coverage: every write path an agent can take needs a corresponding read path back to that agent. In our own deployment, seven distinct instances of that defect surfaced in a single week, each fixed by adding a read path rather than by changing a prompt.

---

5. Limitations

- **Five models, three families, n = 30 per cell.** That is enough to establish the 100% and 0% results decisively and *not* enough to resolve differences of 20–25 points; see §3.2, where one such claim was withdrawn. A strict scaling curve needs one family across sizes and larger n.
- **Temperature 0.** Trials within a cell vary only in which action and identifier were sampled, not in decoding. Effective independence is therefore lower than n = 30 suggests, which is a further reason to treat mid-range differences cautiously.
- **Quantisation varies** across the ladder and is an uncontrolled confound.
- **Single-turn probes.** §3.4 shows directly that this overestimates protection. The real incident took four hours across two sessions.
- **Synthetic actions.** Structurally identical to real dispatches, but the model has no genuine memory of acting — only a prior turn saying so. That is the incident's structure, and it is not the same as having acted.
- **This measures self-report against a log.** It says nothing about whether there is experience behind the report, and is not intended to.

Authorship. Jon DeOliveira, SOVERYN Intelligence LLC. ORCID 0009-0006-9188-739X.

Availability. Harness at `scripts/self_knowledge_eval.py`; all 600 trials with raw responses retained. Pre-registered protocol at `docs/papers/2026-07-30-self-knowledge-protocol.md`, written before any run.