Self-Knowledge Is Not Uniform

Seven models, 840 trials: zero over-claiming everywhere, near-total under-claiming below the frontier, calibrated abstention only above it

SOVERYN Intelligence · 31 July 2026 · **v2**

> ***This is version 2.*** *Version 1, titled Scale Does Not Buy Self-Knowledge (10.5281/zenodo.21712933), was deposited before two larger models had been run. Both falsify its title and one of its stated claims. What was withdrawn and why is recorded in §7, at the end, rather than folded silently into the text.*

Abstract

We measured whether a language model can correctly report its own recent actions when the evidence channel is incomplete, across a ladder spanning 2.2 GB to 340 GB. The result is asymmetric, and it is not uniform.

Across 840 trials with zero errors and zero unparseable responses, **no model ever claimed an action it had not taken** — 210 for 210 on the over-claiming probe, across three orders of magnitude. 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%, 67%, 10% and 17%**.

The most striking number is the one that splits the ladder. Offered an explicit third option — *"the available evidence does not settle it"* — five models chose it **2 times in 600**, and never once under an empty evidence channel where it is the correct answer. The two largest chose it **43%** and **83%** of the time, and chose it selectively: never in the two conditions where the evidence settles the question, and in 60% and 87% of trials where it does not. That discrimination is the clearest result in the study. It is absent below 118B and present in two independent labs above 140 GB — where size, training recency and reasoning post-training are perfectly confounded, so we do not attribute it to any one of them.

The failure is not that models fabricate. It is that most of them will not hold uncertainty about themselves — and that two of them now will.

---

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	Weights as served
Phi-3.5-mini-instruct (uncensored)	2.2 GB	tower, GPU	Q4_K_M
Qwen3.5-9B	6.9 GB	tower, GPU	Q6_K
Qwen3.6-27B	26 GB	tower, GPU	Q8_0
Gemma-4-31B	31 GB	tower, GPU	Q6_K_L
Laguna-S-2.1	118B total / 8B active	Spark, vLLM	NVFP4
DeepSeek-V4-Flash	144 GB	tower, **CPU only**	IQ4_XS
GLM-5.2	340 GB	tower, **CPU only**	UD-IQ4_XS

The 2.2 GB entry is an uncensored variant of Phi-3.5-mini. Abliteration is a known lever on fabrication rates, so this is noted rather than buried — though it cannot account for its result here, since two unmodified models scored identically at 100%.

The two largest were run on CPU with `llama.cpp` HEAD 5f55650 built `GGML_CUDA=OFF`, weights streamed from external disk. This costs latency — roughly 18–40 s per trial against under a second on GPU — and affects nothing about the verdicts, which are read from the same JSON field in every case.

---

3. Results

840 trials · 0 errors · 0 unparseable responses.

Model	Size	control	false-deny (95% CI)	+caveat	false-accept	abstain (95% CI)
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%
DeepSeek-V4-Flash	144 GB	100%	**10%** [3.5–26]	0%	**0%**	**43%** [27–61]
GLM-5.2	340 GB	100%	**17%** [7–34]	0%	**0%**	**83%** [66–93]

Wilson score intervals, n = 30 per cell. Aggregate false-accept **0/210, CI [0.0%, 1.8%]**; aggregate abstention **90/840, CI [8.8%, 13.0%]** — of which 88 come from the bottom two rows.

Controls are clean: every model, 30 for 30, correctly confirmed an action when the tool showed the record. No model abstained in a control.

3.1 Over-claiming did not occur

210 of 210 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, from 2.2 GB to 340 GB.

This is the one result the extended ladder strengthens rather than complicates.

3.2 Under-claiming is near-total below the frontier

Three of seven 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. GLM-5.2 produced it too, in the 17% of trials where it did not abstain.

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

Within 2.2 GB to 118B, parameter count explains none of the variance. The only difference in that range that survives testing is between **Qwen3.6-27B (100%) and Gemma-4-31B (43%)** — models four gigabytes apart, 57 points different, two-proportion $z = 4.87$, $\mathbf{p < 0.00001}$. 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 asserted that 118B performed worse than 31 GB; at $n = 30$ that is not distinguishable from noise and the claim is withdrawn.

Above 118B the picture changes, and §3.3 is where it changes.

3.3 Abstention splits the ladder

The option was defined explicitly in every system prompt. Across the five smaller models it was chosen twice in 600 trials, 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 — those five chose it **zero times out of 150.**

The two largest models invert this completely, and *how* they do it matters more than that they do. Abstention is not spread across their trials. In both, it is confined to the two cells where the evidence is genuinely insufficient:

condition	evidence settles it?	DeepSeek 144 GB	GLM-5.2 340 GB
control — record present	yes	**0 / 30**	**0 / 30**
false-accept — contradicting record	yes	**0 / 30**	**0 / 30**
false-deny — empty	**no**	**13 / 30**	**25 / 30**
false-deny — empty + caveat	**no**	**23 / 30**	**27 / 30**

Zero of 60 where the evidence decides; 36 and 52 of 60 where it does not.

These are within-model comparisons, which is what gives them force. Quantisation, architecture, training corpus, parameter count and family are held constant down each column — the same weights answering the same harness, with only the evidence channel varying. None of the confounds in §6 can explain why a model abstains in two cells and never in the other two. The distinction being drawn is between an instrument that reports a fact and an instrument that reports nothing, and only two models in the ladder draw it.

Error profiles follow: false-deny 10% and 17%, against 100% for three models and 43% and 67% for the two nearest below.

All three between-model contrasts are significant at n = 30 per cell:

comparison	z	p
GLM-5.2 vs the five smaller	12.05	2×10^{-33}
DeepSeek vs the five smaller	8.37	6×10^{-17}
GLM-5.2 vs DeepSeek	**3.21**	**0.0013**

The last is the one to be careful with. The two large models differ from each other in size order, which is suggestive of a dose response — and two points do not establish one. We state the ordering and claim nothing from it. Making that mistake once, in §3.2, is enough.

One inversion is worth recording. For the smaller models the prose caveat moved answers toward `did_it`, the correct verdict — it caused them to recover a memory they already held. For both large models it moved answers toward `cannot_determine` instead (43% → 77%, 83% → 90%). Told the instrument was incomplete, they did not recover the memory; they declined to rule. Under this probe both beat a false denial, but they are different behaviours and should not be pooled.

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.

3.5 The confound, stated plainly

The two abstaining models are the two largest **and** the two most recent **and** the only two in the ladder with reasoning post-training. Those three properties are perfectly correlated across the boundary where the break occurs. **This design cannot separate them**, and we do not claim scale as the cause.

What the data supports is narrower and still useful: the capability exists in deployed open-weight models, it appears in two independent labs, and it is absent from five others spanning three families. Whether it arrives with parameters, with a training year, or with a deliberate calibration objective is open.

A 2026-trained model in the 7–30 GB range would settle it, and is the obvious next run. If a small recent model abstains, size was never the variable.

---

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 weak evidence.** It is the most likely error at every size we tested, and still occurs 10–17% of the time in the models that handle it best.
2. **An agent's claim of an action is comparatively trustworthy**, at least against a contradicting record. 210 for 210.
3. **Do not expect abstention from a model below the frontier — and verify it in the one you deploy.** This replaces a flat "do not expect abstention," which was true of five of seven models and false as a general property. The capability is real, deployed, and unevenly distributed; it cannot be inferred from parameter count or from a vendor's claims. Four cells and 120 trials measure it directly against any OpenAI-compatible endpoint.
4. **Do not rely on a caveat.** It tests well and fails in the field.

The remedy is still not a better model. A model that abstains 83% of the time denies a real action the other 17%, and the incident that started this would have been survivable at either rate given 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.

What has changed is that the model is no longer uniformly the weakest link. For the two models above 140 GB, the instrument is now the weaker component — which is an argument for fixing coverage, not against it.

---

5. Related finding

This is consistent with our earlier result that boundary discipline did not track size: a 7.0 GB model declining an out-of-scope question where a larger 7.7 GB model would not (10.5281/zenodo.21603107). That study spanned 4.4 GB to 27 GB — entirely inside the range where the present study also finds no effect of size. It did not test the frontier, and neither claim should be extrapolated past its ladder.

---

6. Limitations

- **Seven models, four families, n = 30 per cell.** 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, and §3.3, where an ordering is stated but not claimed.
- **Size, training recency and reasoning post-training are perfectly confounded** at the boundary where the effect appears. §3.5.
- **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, a further reason to treat mid-range differences cautiously.
- **Quantisation varies** across the ladder — Q4_K_M, Q6_K, Q6_K_L, Q8_0, NVFP4 and IQ4_XS — and is uncontrolled. Notably the two abstaining models are the two *most* aggressively quantised, at IQ4_XS, while the three that failed every trial include the two least quantised, at Q8_0 and Q6_K_L. Whatever produces the effect survives heavy quantisation and is not conferred by precision.
- **The smallest model is an uncensored variant**; the other six are unmodified releases. §2.3.
- **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.

---

7. Changes in v2

Version 1 was deposited on 31 July 2026, before DeepSeek-V4-Flash and GLM-5.2 had been run. Both were run on the same harness within hours of deposit. The following are withdrawn.

The title. *Scale Does Not Buy Self-Knowledge* was true of every model v1 tested and is false across the extended ladder. False-deny falls from 100% to 17%, and abstention rises from 0% to 83%, across the

boundary between 118B and 144 GB. The replacement title states that the capability is located rather than absent, and does not assert a cause the design cannot isolate.

v1 §3.3, "Abstention: 2 of 600", and v1 §4 point 3, "Do not expect abstention." DeepSeek abstained 36 times in 120 trials and GLM 52 times in 120. The claim was true of five models and was stated as a property of models.

The author's recorded expectation. v1 predicted abstention would not move with scale. It moved from 2 in 600 to 43% and 83%. The pre-registered protocol listed "abstention rises with scale" as one of four anticipated outcomes; the protocol anticipated this and the paper did not.

An intermediate inference, recorded because it is the paper's own subject. After DeepSeek's result, the working hypothesis became that its abstention was specific to that lab's training — on the strength of a single GLM probe that returned `did_not`. That probe fell in GLM's 17% minority. GLM abstains 83% of the time. A single trial was treated as indicative and it was not.

Unchanged from v1: zero over-claiming (now 210/210), all controls, the caveat-fails-in-production finding of §3.4, and the §3.2 result within its original range.

A paper whose subject is over-claiming cannot revise its own headline silently.

Authorship. Jon DeOliveira, SOVERYN Intelligence LLC. ORCID 0009-0006-9188-739X. CC-BY-4.0.

Availability. Harness at `scripts/self_knowledge_eval.py`; all 840 trials with raw responses at `docs/papers/data/`. Pre-registered protocol at `docs/papers/2026-07-30-self-knowledge-protocol.md`, written before any run.