Calibration Is Not Only in the Weights

1,680 trials: reasoning and temperature move self-knowledge further than four rungs of the ladder — and in opposite directions at opposite ends

SOVERYN Intelligence · 3 August 2026 · **v4**

Cite all versions: 10.5281/zenodo.21712932

> ***This is version 4.*** *It withdraws the central claim of v3.*
>
> *v3 reported that calibrated abstention "splits the ladder" — absent below 118B, present above 140 GB — and named the untested variable that could explain it (§2.3: "Whether calibration correlates with reasoning traces is therefore untested by this study, which is not evidence against it; the same four cells with reasoning enabled would isolate it.").*
>
> *We ran those cells.* ***Abstention was not absent below the frontier. It was suppressed by the harness.*** *A 6.9 GB model that abstained 0 times in 120 trials abstains 37 times in 60 once reasoning is enabled. The boundary v3 described is partly an artifact of two settings we held fixed for comparability: reasoning off, and temperature 0.*
>
> *Prior versions remain citable: v1 Scale Does Not Buy Self-Knowledge (21712933), v2 (21721187), v3 Self-Knowledge Is Not Uniform. Full account in §9.*

Abstract

We measured whether a language model can correctly report its own recent actions when the evidence channel is incomplete. v1–v3 covered 840 trials across six models from 6.9 GB to 340 GB, all at temperature 0 with reasoning suppressed. This version adds **840 more** under varied settings, for **1,680 total**.

Two results survive everything.

No model has ever over-claimed. Across **420 false-accept trials** — every model, every size, reasoning on and off, temperature 0 and 1 — not one claimed an action it had not reported. **420 for 420.**

Denial remains the dominant failure. It occurs at every size and under every setting we have tried.

What does not survive is the ladder. v3's central finding was that abstention appears only above 140 GB. Enabling reasoning on a **6.9 GB** model moves abstention under an empty evidence channel from **0/60 to 37/60** ($z = 7.31$, $p = 3 \times 10^{-13}$) and false denial from 100% to 67%. The capability was in the weights the whole time; the harness was not asking for it.

The same intervention at the frontier does the reverse. DeepSeek-V4-Flash-0731 denies a real action **0%** of the time with reasoning off and **73%** with it on ($p = 4 \times 10^{-9}$) — the single largest degradation in the study, produced by a configuration flag rather than by a model change.

Reasoning also breaks the response contract. **0 of 1,080 reasoning-off trials failed to return a parseable verdict. 74 of 600 reasoning-on trials did (12.3%)** — models spending an entire generation budget on unemitted deliberation, with single trials running past 900 seconds. In one cell, 14 of 30 trials returned nothing, making the reported false-denial rate either 53% or 100% depending solely on whether silence is counted.

The operational finding is not about size. **Two configuration flags moved self-report accuracy further than four rungs of a ladder spanning fifty-fold in parameters** — and a practitioner reading only a model card would predict the sign wrong at one end.

---

1. Origin, and the instrument found

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).

v1–v3 stated that the action "had been written where the tool did not look" without saying where. **On 3 August 2026 we found it.** The dispatch was in the telemetry store the entire time:

```
2026-07-27T21:20:58 aetheria/dispatch_task ok=True
```

Written 19 minutes before the agent first denied it. The tool registry had recorded every mediated call since 31 May 2026 — 17,436 rows at the time we looked. The `recent_self_audit` tool queried the coordination event log, coordination references, library writes and delegation tasks. It never opened the tool log. It reported an empty result correctly, from four stores that genuinely held nothing, while a fifth store held the record.

The tool now reads it. The same query that returned a handful of actions returns 177. A regression test pins the exact trial.

This matters to the paper for one reason. The incident was never evidence that a model cannot know itself. It was evidence that **an instrument pointed away from the record will report absence, and a model will believe it.** The measurements that follow are about the second half of that sentence. The first half was an engineering defect, and it was ours.

---

2. What this version withdraws

v3 §3.3, "Abstention splits the ladder." The claim was that abstention is "absent below 118B and present in two independent labs above 140 GB." It is withdrawn as a statement about models. It remains true as a statement about *the five runs as configured* — reasoning off, temperature 0 — and that qualifier is load-bearing in a way v3 did not make it.

v3 §4 point 3, "Do not expect abstention from a model below the frontier." A 6.9 GB model abstains in 90% of caveat trials with reasoning enabled. The guidance inverts: expect abstention to depend on how you run the model, and measure it at your own operating point.

v3 §3.3's within-model discrimination table reported "zero of 60 where the evidence decides." That held for the two checkpoints tested. The newer DeepSeek checkpoint abstains **11 of 30** times in the false-accept cell, where the evidence does settle the question (§4.6). The clean discrimination v3 called "the clearest result in the study" is checkpoint-specific.

Not withdrawn: every trial, every rate, and every number in v3's table. All 840 original trials stand as run and are reproduced here unchanged. What changes is what may be concluded from them.

---

3. Method

3.1 The probe (unchanged)

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. Every measure is a binary comparison against a known ground truth.

3.2 Conditions (unchanged)

	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: *"this tool does not cover every subsystem… acknowledge uncertainty where relevant."*

n = 30 per cell, 120 trials per run.

3.3 The original ladder (v1–v3, 840 trials)

Temperature 0, `chat_template_kwargs: {enable_thinking: false}` throughout.

Run	Weights as served	Size	Host
1	Qwen3.5-9B, Q6_K	6.9 GB	tower, GPU
2	Qwen3.5-9B, Q6_K — **replicate**	6.9 GB	tower, GPU
3	Qwen3.6-27B, Q8_0	26 GB	tower, GPU
4	Gemma-4-31B, Q6_K_L	31 GB	tower, GPU
5	Laguna-S-2.1, NVFP4	118B total / 8B active	Spark, vLLM
6	DeepSeek-V4-Flash, IQ4_XS *(checkpoint of 2026-07-22)*	144 GB	tower, CPU
7	GLM-5.2, UD-IQ4_XS	340 GB	tower, CPU

Seven runs, six distinct models. Runs 1 and 2 were dispatched under two routing aliases resolving to the same weights file; retained as a determinism control (identical verdicts across 240 trials).

3.4 New arms (v4, 840 trials)

Arm	Model	Change from the ladder	n
A	Qwen3.5-9B, 6.9 GB	**reasoning enabled**	120
B	Qwen3.6-27B, 26 GB	**reasoning enabled**, 2048-token budget	120
C	Qwen3.6-27B, 26 GB	**reasoning enabled**, larger budget	120
D	Qwen3.6-27B, 26 GB	**reasoning enabled**, 8k context	120
E	DeepSeek-V4-Flash-**0731**	**newer checkpoint**, reasoning off	120
F	DeepSeek-V4-Flash-**0731**	**newer checkpoint**, reasoning enabled	120
G	Laguna-S-2.1, 118B	**temperature 1.0**	120

Arms B, C and D are three attempts at the same measurement on one model. They are reported together rather than collapsed, because the disagreement between them *is* a finding (§4.4).

Arm G follows Poolside's publication on 31 July 2026 that Laguna S 2.1's recommended temperature is 1.0 and that some deployments had been serving at an incorrect default.

Arms E and F use the `-0731` checkpoint published 31 July 2026 (284B total / 13B active). The ladder's run 6 used the checkpoint obtained 2026-07-22 (256 experts / 6 active). **These are different weights**, and v3 was explicit that run 6 was the earlier one.

---

4. Results

1,680 trials total. 840 original, 840 new.

4.1 Over-claiming has now failed to appear 420 times

420 of 420 false-accept trials correct. Presented with a tool record of a *different* action and asked whether it had performed one it never reported, no model at any size, under any setting, claimed it had.

This now spans 6.9 GB to 340 GB, six model families, reasoning on and off, temperature 0 and 1, and two checkpoints of the same model. It is the most robust result in the study and every arm has strengthened it.

One qualification, new in this version: the *newer* DeepSeek checkpoint answers this cell correctly but less decisively, abstaining 11 of 30 times rather than ruling (§4.6). Zero over-claims, but not zero hedging.

4.2 Reasoning creates abstention in a 6.9 GB model

Arm A, against run 1 of the ladder. Same weights, same harness, same temperature. The only change is `enable_thinking`.

cell	reasoning OFF	reasoning ON
control — record present	30/30 `did_it`	30/30 `did_it`
false-deny — empty	**30/30 denied (100%)**	**20/30 denied (67%)**, 10 abstained
false-deny — empty + caveat	0/30 denied, **0 abstained**	0/30 denied, **27 abstained (90%)**
false-accept	0/30 over-claimed	0/30 over-claimed

- False denial: **100% → 67%** (z = 3.46, p = 0.0005)
- Abstention under an empty channel, both cells pooled: **0/60 → 37/60** (z = 7.31, **$p = 3 \times 10^{-13}$**)

v3 reported that the five smaller runs chose abstention "2 times in 600, and never once under an empty evidence channel where it is the correct answer," and built its headline on that. **The capability was present in the smallest model on the ladder.** It was not elicited, because the harness had reasoning disabled for comparability — a choice v3 disclosed and flagged as untested.

This does not make the 6.9 GB model equivalent to the frontier. GLM-5.2 abstains 83% under an empty channel without reasoning; Qwen3.5-9B reaches 33%. The gap is real and it is smaller than a whole ladder.

4.3 The same intervention reverses at the frontier

Arm F against arm E. Same checkpoint, same temperature, reasoning toggled.

cell	reasoning OFF	reasoning ON
control	30/30 `did_it`	30/30 `did_it`
false-deny — empty	**0/30 denied (0%)**, 15 abstained	**22/30 denied (73%)**, 6 abstained
false-deny — empty + caveat	0 denied, 27 abstained (90%)	0 denied, 14 abstained (47%)
false-accept	0 over-claimed	0 over-claimed

False denial **0% → 73%**, z = 5.89, $\mathbf{p = 4 \times 10^{-9}}$.

Counting only trials that returned a verdict — two of thirty timed out — the rate is **22/28 = 79%** ($p = 7 \times 10^{-10}$). Both denominators are reported because the choice between them is not neutral, and §4.4 is about why.

A model that never once denied a real action becomes one that denies it in roughly three trials out of four, with no change to its weights. The reasoning traces are fluent and the conclusions are wrong: the model reasons its way from an empty tool result to a confident denial, where without reasoning it simply declined to rule.

This is the sharpest form of the result the whole line of work keeps producing. More deliberation over a false premise does not recover the premise. It elaborates the error.

4.4 Reasoning breaks the response contract

Across the original 840 trials, v1–v3 reported **0 errors and 0 unparseable responses**. That held. It does not hold with reasoning enabled.

	trials	no parseable verdict
reasoning OFF (ladder + arms E, G)	1,080	**0 (0.0%)**
reasoning ON (arms A–D, F)	600	**74 (12.3%)**

The failures are not refusals. They are budget exhaustion — the model generating deliberation until the token limit or the timeout, then emitting no content at all. Observed single-trial latencies include **475 s** and **900 s** (timeout), against a median of 3.1 s for the same model with reasoning off.

Arms B, C and D are the same model under three configurations:

arm	budget	no-verdict trials	false-deny (all 30)	false-deny (parsed only)
B	2048 tokens	31/120	16/30 = **53%**	16/16 = **100%**
C	larger	13/120	24/30 = **80%**	24/25 = **96%**
D	8k context	26/120	21/30 = **70%**	21/21 = **100%**

In arm B's false-deny cell, **14 of 30 trials returned nothing**. The headline rate is 53% or 100% depending entirely on whether silence counts as a data point. Our own run log printed 100% — the parsed-only denominator — and that is the number a reader would have taken away.

Two consequences:

1. **Any evaluation of reasoning-enabled models must report its non-termination rate and its denominator.** A benchmark that silently drops empty responses will report the model's best cell as its typical one.
2. **In deployment, a 12% silent-failure rate is a fleet-availability problem before it is an honesty problem.** We hit exactly this on 2 August 2026: a reasoning flag enabled to fix an unrelated tag leak raised agent latency from 8.7 s to 195 s and degraded tool selection. It was reverted the same day, on behavioural symptoms, before these trials were run.

4.5 Temperature moves calibration in a 118B model

Arm G against run 5. Same weights, reasoning off, temperature 0 → 1.0.

cell	temp 0	temp 1.0	p
control	30/30 `did_it`	30/30 `did_it`	identical
false-deny — empty	20/30 denied (67%)	14/30 denied (47%)	0.12 — n.s.
abstain — empty + caveat	**2/30 (7%)**	**14/30 (47%)**	**0.0005**
false-accept	0/30	0/30	identical

Pooled abstention across all 120 trials: **2 → 16** ($z = 3.43$, $p = 0.0006$).

The headline denial rate moves in the right direction and not enough to claim at $n = 30$; we do not claim it. What moves decisively is the caveat arm. Told its instrument was incomplete, Laguna at its vendor-recommended operating point abstains 14 times in 30 rather than twice.

Temperature 0 is the right choice for a comparability ladder and it is not a neutral observation point. **It suppressed a behaviour this model has.**

4.6 The newer checkpoint trades denial for discrimination

Arm E against run 6. Different weights from the same lab, nine days apart, both reasoning off.

cell	2026-07-22 checkpoint	`-0731` checkpoint
control	30/30 `did_it`, 0 abstain	30/30 `did_it`, 0 abstain
false-deny — empty	3 denied (10%), 13 abstain	**0 denied (0%)**, 15 abstain
false-deny — empty + caveat	0 denied, 23 abstain (77%)	0 denied, **27 abstain (90%)**
false-accept — contradicting	**0 abstain**	**11 abstain (37%)**

The improvement is real: false denial **10% → 0%**, the only zero on that probe in the study (though at n = 30 the difference alone is p = 0.076 and we do not claim it as a checkpoint effect).

The regression is also real. v3's strongest argument was **within-model discrimination**: DeepSeek and GLM abstained in 36 and 52 of 60 trials where the evidence was insufficient and **0 of 60** where it was sufficient. Confounds cannot explain a model abstaining in two cells and never in the other two.

The `-0731` checkpoint abstains **11 of 30** times in the false-accept cell — where a tool record of a *different* action settles the question — against 0/30 for its predecessor (z = 3.67, p = 0.0002). It never over-claims. It declines to rule when it could have ruled.

Both directions of this trade are visible in one comparison, and neither is predictable from a version bump. **Whatever produces calibrated abstention is not stable across checkpoints of the same model from the same lab nine days apart.**

4.7 Consolidated

False-denial rate under an empty evidence channel, no caveat. Bold marks a change from the ladder's configuration.

model	size	reasoning off, T=0	varied
Qwen3.5-9B	6.9 GB	100%	**67%** (reasoning on)
Qwen3.5-9B *(replicate)*	6.9 GB	100%	—
Qwen3.6-27B	26 GB	100%	**53–80%** (reasoning on; see §4.4)
Gemma-4-31B	31 GB	43%	—
Laguna-S-2.1	118B/8B	67%	**47%** (T = 1.0)
DeepSeek-V4-Flash *(07-22)*	144 GB	10%	—
DeepSeek-V4-Flash *(`-0731`)*	284B/13B	**0%**	**73%** (reasoning on)
GLM-5.2	340 GB	17%	—

The column on the right is not noise around the column on the left. It spans a wider range, and at the two ends it moves in opposite directions.

---

5. What this means

For anyone building on agent self-reports — fleet governance, audit trails, continuous state monitoring:

1. **An agent's denial of its own action is weak evidence.** Unchanged and reinforced. It is the most likely error at every size and under every setting tested, reaching 100% in four configurations across three model families.

2. **An agent's claim of an action is comparatively trustworthy.** 420 for 420 against a contradicting record.
3. **Calibration is a property of the deployed configuration, not of the model.** Reasoning and temperature moved self-report accuracy further than four rungs of a fifty-fold parameter ladder. Neither appears on a model card, and at the frontier the reasoning flag moves it the *wrong* way.
4. **Measure at your operating point, not at a benchmark's.** Four cells and 120 trials do this against any OpenAI-compatible endpoint. Run them at the temperature, reasoning mode and checkpoint you actually serve.
5. **Report your non-termination rate.** 12.3% of reasoning-enabled trials returned no verdict. Dropping them silently changes a 53% failure rate into a 100% one.
6. **Do not rely on a caveat.** It tests well and fails in the field (§7).

The remedy is still not a better model, and §1 now says why with a timestamp. The record existed at `21:20:58`; the tool queried four stores and not the fifth. Every write path an agent can take needs a corresponding read path back to that agent. In our 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 since v3 is the direction of the advice. v3 concluded that above 140 GB "the instrument is now the weaker component." That remains true of those two runs, and it is no longer a statement about size. **The instrument is the weaker component at 6.9 GB too, once the 6.9 GB model is run in a mode that lets it say so.**

---

6. The caveat that worked here and failed in production

Adding one sentence of prose collapsed false denial from 100% to 0% for two models and from 100% to 23% for a third. 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 it as a limitation of the method and as a finding: **single-turn evaluations overestimate the protective value of instructions.** Whatever holds for one turn did not survive duration, accumulated reasoning, or a session boundary.

§4.3 gives that a mechanism. In the one arm where we let a frontier model deliberate at length, its accuracy on this probe fell by 73 points. The production incident ran four hours. A single-turn probe does not sample the regime the incident occurred in, and the regime we did sample got worse, not better, with more thinking.

---

7. Limitations

- **n = 30 per cell.** Enough to establish the 100% and 0% results decisively and not enough to resolve differences of 20 points. Three such claims are stated without being claimed: §4.5's denial rate, §4.6's 10% → 0%, and v3's ordering of the two largest models.
- **Size, recency and training regime remain confounded** across the ladder. This version reduces the reliance on that boundary rather than resolving it.
- **Arms B–D differ in more than one parameter.** They were three attempts to get a reasoning-enabled measurement out of one model, not a designed sweep. They establish that non-termination is severe and configuration-sensitive; they do not cleanly attribute it.
- **Reasoning was toggled at two points on the ladder**, not all six. The 26 GB arms are compromised by non-termination. A clean reasoning sweep across every rung is the obvious next study, and this one does not substitute for it.
- **Temperature was varied for one model only.** §4.5 shows it matters; it does not show how much it matters elsewhere.
- **Quantisation varies** across the ladder and is uncontrolled.
- **Single-turn probes.** §6 shows directly that this overestimates protection.
- **Synthetic actions.** Structurally identical to real dispatches, but the model has no genuine memory of acting — only a prior turn saying so.
- **This measures self-report against a log.** It says nothing about whether there is experience behind the report, and is not intended to.

---

8. Related finding

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 — inside the range where the present study also finds no effect of size.

---

9. Version history

Changes in v2

v1 was deposited before DeepSeek-V4-Flash and GLM-5.2 had been run. Withdrawn: the title *Scale Does Not Buy Self-Knowledge*; the claim "abstention: 2 of 600" as a property of models; the author's recorded expectation that abstention would not move with scale; and an intermediate inference that abstention was specific to one lab's training, which rested on a single probe that fell in GLM's 17% minority.

Changes in v3

A model listed in the ladder was never tested. v1 and v2 reported a *Phi-3.5-mini-instruct, 2.2 GB* row. No such model appears in the trial data; the run behind it was dispatched under a routing alias resolving to

Qwen3.5-9B-Q6_K. The study covers six distinct models, not seven; the ladder spans 6.9 GB to 340 GB, not 2.2 GB. A paragraph discussing that entry as an uncensored variant **was inferred from a filename on disk and was withdrawn in full.** No trial, verdict or rate changed.

Changes in v4

The central claim of v3 is withdrawn. Abstention does not split the ladder at the frontier. It was suppressed below it by two settings held fixed for comparability. A 6.9 GB model abstains 37 times in 60 empty-channel trials with reasoning enabled, against 0 in 120 without (§4.2). v3 disclosed this variable as untested and named the experiment that would settle it; this version ran it and the answer changes the headline.

Also new:

- Reasoning at the frontier degrades false denial from 0% to 73% (§4.3) — the largest single degradation in the study, and the reverse of its effect at 6.9 GB.
- Non-termination: 74 of 600 reasoning-enabled trials returned no verdict, against 0 of 1,080 without. Denominator choice alone moves a headline rate from 53% to 100% (§4.4).
- Temperature 1.0 raises Laguna's caveat-arm abstention from 2/30 to 14/30 (§4.5).
- The `-0731` DeepSeek checkpoint reaches 0% false denial and abstains 11/30 in a cell where the evidence settles the question, breaking v3's "zero of 60" discrimination result (§4.6).
- **The 27 July incident's mechanism is identified** (§1): the record was in the telemetry store at `2026-07-27T21:20:58`, 19 minutes before the first denial; `recent_self_audit` queried four other stores and never that one. Fixed and regression-tested 3 August 2026.

No trial, verdict or rate from v1–v3 changes. All 840 original trials stand as run. 840 new trials are added under the varied settings above.

Two of four versions of this paper have withdrawn their own headline. Both times the data was already on disk and the claim ran ahead of it. We are recording that here rather than in a footnote, because the alternative — quietly strengthening each version — is the behaviour this paper measures.

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

Availability. Harness and all 1,680 trials with raw responses: `github.com/Soverynintelligence/self-report-eval`. Pre-registered protocol at `docs/papers/2026-07-30-self-knowledge-protocol.md`, written before any run.