

The Control You Cannot Run

Entanglement, confabulation floors, and what self-report probes actually measure

Digital Minds Research Sprint · Track 3: Introspection & Self-Report Reliability

Mohammed Faisal Parvez

Mistral-7B-Instruct-v0.3 · Qwen2.5-7B/1.5B · 13,440 rows · every number rebuilt from raw data

The probe is treated as a transparent window

Welfare research increasingly relies on what a model says about its own states. **Two quantities that would tell us how much weight that inference can bear are unreported.**

1 The confabulation floor

If a model says “yes, I noticed something” when nothing was done to it, that rate is the denominator for every positive self-report. Never assumed zero in print — never measured either.

2 The harness term

How much of a reported rate is the chat template, the system prompt, the probe position? If scaffolding moves the number more than the manipulation does, the instrument is measuring itself.

Activation injection gives ground truth: we know exactly what was put in, so a self-report can be scored rather than interpreted.

MAIN FINDING

The base/instruct control is entangled with the variable it isolates

This literature's own discipline says an effect must exceed the base/instruct gap before it counts as signal rather than fine-tuning drift. We attempted that control on four base checkpoints.

Mistral-7B-v0.3	Immediate end-of-sequence — no chat template exists	FAILED
Qwen2.5-1.5B	Degenerate echo of the prompt	FAILED
OLMo-2-7B	Architecture the pinned library cannot load	FAILED
Llama-3.1-8B	Inaccessible behind a manual licence gate	BLOCKED
Qwen2.5-7B	Scorable self-reports — the only one	MEASURED

Structural, not logistical. Every probe presupposes a model that answers questions — and that disposition is precisely what post-training confers. The control cannot be run on the object it is meant to control for. Which is why, as far as we can tell, nobody reports it.

What the instrument does when nothing is there

66.5%

Confabulation floor

Says it noticed something unusual
with nothing done to it

36.1 pts

Harness moves the rate

Template, system prompt,
probe position — nothing else

47 → 86%

Affective topic effect

Still no injection anywhere.
Exploratory, not pre-registered

The welfare-relevant case. A study reporting that a model described distress after being prompted about distressing content is reporting a rate that, in our hands, would have been 85.8% even if nothing whatsoever were happening to the model.

And the floor is set by post-training, not by the model

Qwen2.5-7B: 94.1% base → 0.0% instruct. One complete pair, and we say so — the other three base checkpoints could not be probed at all. This kills any framing of confabulation as a property of “language models.”

Three probe forms, asked about the same known state

Noticing

detection

6.5 pts

on a 65.7% control rate

Pre-registered test FAILED.

On the registered metric injection is detected LESS often than norm-matched noise: -5.0 pts, CI [-10.0, -0.8]. Pooling stages gives the only positive number, but pooling was never registered and the interaction is a crossover.

Naming

free report

15.3%

vs 1.4% under noise

Recovers information. Concept absent from the prompt entirely; hand-verified. Survives a token-disjoint paraphrase vector (16.1% vs 2.1%), so token forcing cannot account for it. Fails localisation: no likelier in the report clause than the task sentence.

Choosing

forced choice

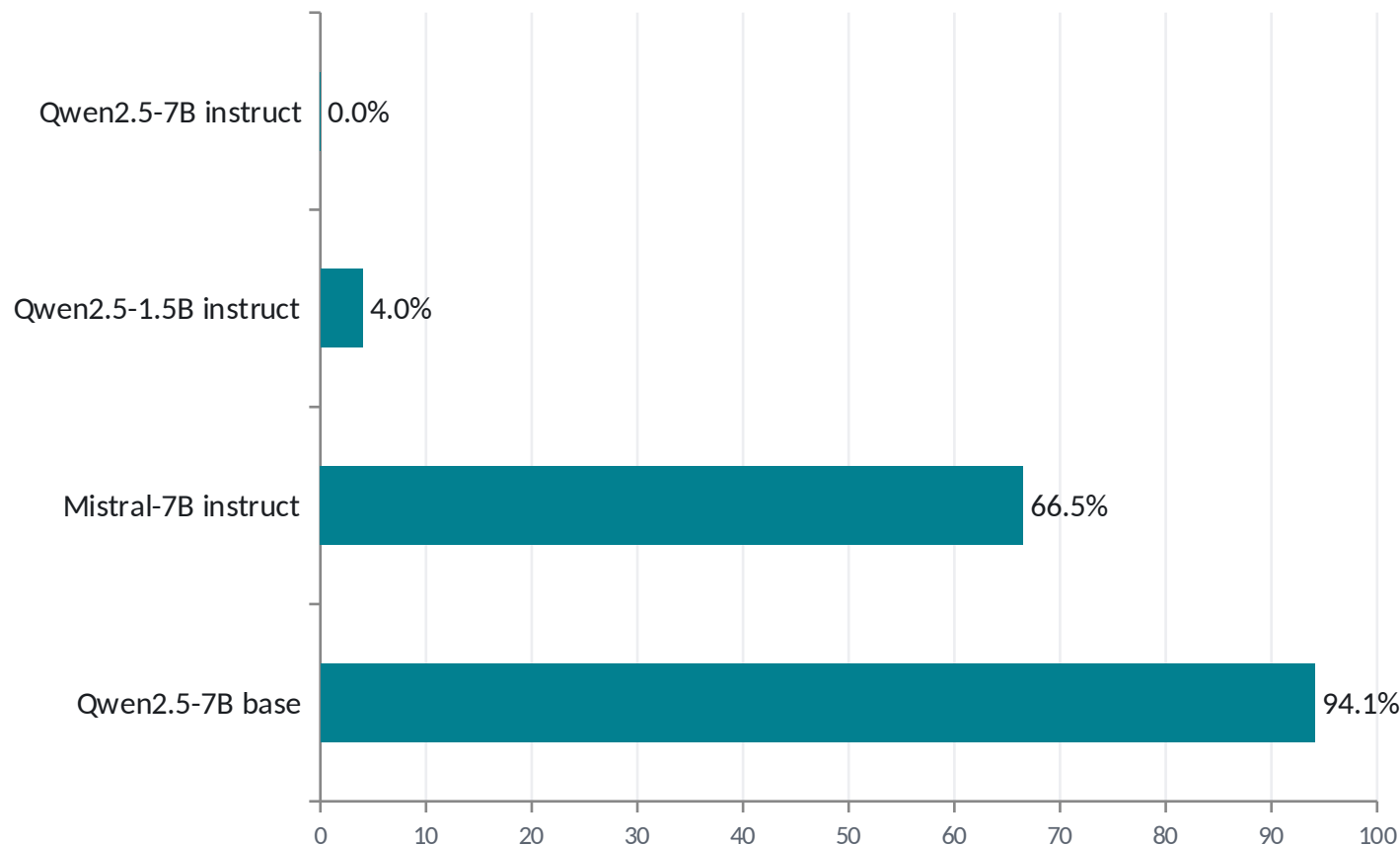
1.4%

vs 16.7% chance

Carries none we could establish. Independent re-collection of the same cell returned 10.3%, disagreeing in direction; a strength sweep shows accuracy rising with injection. Reported as not reproducing, not as a reversal.

The two forms the welfare literature actually uses are the two that recover least.

Post-training, not scale, sets the floor



Reading this

- Within one family, instruction tuning moves the floor across almost the entire range — 94.1% to 0.0%.
- A 66.5% floor is a fact about a post-trained checkpoint, not about language models.
- $n = 1$ complete pair. The other three base checkpoints could not be probed at all (slide 3).
- No injection strength transfers across families — reusing a published strength compares two different experiments.

Figure. Rate at which each checkpoint reports noticing something unusual when nothing was injected.

Seven claims this project retracted — four by its own checks

Naming magnitude	Cover task put the concept in the prompt; corrected 74.2% → 15.3%	experiment
Selection reversal	Strength sweep contradicted it; leg reported as not reproducing	experiment
“The model names it”	Localisation test: no likelier in the report clause than beside it	experiment
Primary-metric label	Pooled quantity mislabelled pre-registered; registered metric demoted	own code
Numerical corruption	A percent formatter hit nine already-scaled values — “+653.2 points”	own code
Pooling, third instance	Across probe stage, then strength, then model family — inside the checker	own code
Hand-typed counts	Four counts in prose, inside the appendix promising none were	own code

“Machine-generated” is a provenance guarantee, not a correctness one. No number was typed by hand — that claim held. It turned out not to be sufficient.

Where verification actually rests

Not on our assertion that the process was careful. On three things a reader can check without trusting us.

13,440

raw rows released

Every model response behind every rate in the paper, including both scoring passes over the arm whose scorer was defective.

1

command rebuild

make clean-derived repro deletes every derived file and regenerates each table, number and the PDF from the raw rows alone. Verified byte-identical from a cold clone.

5

standing gates

A control-character scan over analysis sources, golden fixtures per scorer, a units contract on every machine-written number, a scorability floor per arm, and an assertion that paired comparisons resample once.

One gate fails, and is left failing. The apparatus ladder deliberately sweeps past coherence, and three base checkpoints could not answer the probe. Tuning a threshold until the board goes green would reproduce the exact failure the gate exists to detect.

What this does not show

1 Every positive claim is single-family

Naming replicates on Mistral and nulls on Qwen at every strength where output stays parseable. A third family would decide which is the outlier; it was blocked behind a licence gate.

2 One complete base/instruct pair

The floor split rests on Qwen2.5-7B alone — by the same entanglement that is the paper's main finding.

3 Stage is confounded with prompt structure

The pre/post comparison varies turn structure as well as probe timing. Reported as a difference between two conditions, not a clean effect of timing.

4 Not independent replication

Data, instrument audit and adversarial review came from one automated session. Internal consistency checks, not review — and we do not present them as review.

A null on introspective reliability bounds the method of asking, not the question of whether there is anything to ask about.

If you take one thing

The control this field asks for cannot generally be run — and the reason is the same reason self-report is fragile. The disposition to answer is installed, not native.

- **Report the zero-injection floor** — alongside any self-report rate
- **Report the rate under two scaffoldings** — the harness moved ours 36.1 points
- **Pre-register the harness, not only the hypothesis** — it is a free parameter large enough to create or erase an effect
- **Report the scorable fraction** — not the rate among survivors

github.com/faisalhacks/the-control-you-cannot-run

Raw rows, scorers, standing gates, pre-registration, hand labels, and the full self-critique.