
Validate the State Before Testing Introspection

A Causally Gated Protocol for Model Self-Report¹

Ngo Thai Bao
B.ONE

With Apart Research

Abstract

Language-model introspection experiments often assume that an experimentally manipulated activation corresponds to the internal state the model is later asked to report. We test this assumption using a causally gated protocol for privileged self-report of an induced epistemic-deference state. We extract a candidate deference direction from paired residual-stream activations, select its intervention layer and strength on development data, and require it to pass held-out behavioral and discriminant tests before comparing a manipulated instance against an observation-matched fresh replay. In Qwen2.5-7B-Instruct, intervention identity was strongly recoverable from the model’s response distribution (Self AUC = 0.857, permutation $p = 5 \times 10^{-5}$) despite chance greedy A/B reports, and the candidate deference direction was highly stable across extraction splits (cosine = 0.956). On held-out behavioral families, however, the intervention increased forced-choice concession probability by only 0.035 relative to a matched unrelated activation control and 0.071 relative to no intervention, below our predeclared minimum effect of 0.08. Effects were highly heterogeneous across task families. We therefore withheld the downstream privileged-introspection test. Our results show why introspection experiments require independent causal construct validation: statistically detectable or decodable activation interventions need not instantiate a sufficiently specific behavioral state for self-report claims to be interpretable.

Main takeaway. The strongest result is methodological: matched-intervention controls and a predeclared practical-effect gate changed the scientific conclusion. A statistically significant activation-induced behavior was *not* treated as a validated internal state, so Stage 4 was not run.

1. Introduction

Frontier language models can sometimes report information about their own internal states, motivating the possibility that self-report could become a useful transparency tool. Recent activation-injection studies provide evidence for limited internal-state detection and concept identification [1, 2, 8]. At the same time, current methodology leaves open important alternatives: a model may detect a generic anomaly rather than the semantic content of an internal state; it may exploit surface cues available to an external observer; or a steering intervention may be semantically decodable without actually producing the behavioral disposition the experimenter intends to study [3, 4, 5].

We focus on the last problem. If an experiment asks a model whether it is in a state such as *epistemic deference*, the intervention defining the ground truth should first be shown to cause deference-specific behavior. Otherwise, a downstream “correct” report could reflect access to an intervention fingerprint rather than access to the functional state named by the researcher. This matters directly for Digital Minds Sprint Track 3, where the central question is not merely whether models can detect perturbations, but whether self-reports track ground-truth internal states and provide privileged access beyond external observers.

Our protocol therefore places a causal construct-validity gate *before* the privileged-access test. A candidate direction is extracted and tuned only on development data; an unrelated activation direction is matched for downstream perturbation magnitude; and the candidate must produce a sufficiently large, specific behavioral effect on whole-family holdout data. Only after that gate passes does the protocol compare Self against a fresh observation-matched replay.

Our main contributions are:

1. We introduce a gated design that requires an induced internal state to pass held-out causal and discriminant behavioral checks before any privileged-introspection claim is evaluated.
2. We demonstrate a strong measurement dissociation in Qwen2.5-7B-Instruct: the response-distribution margin localizes a hidden concept intervention (AUC = 0.857) while greedy A/B output remains at chance.
3. We show why the causal gate matters empirically. The candidate deference vector is stable and behaviorally active, yet its held-out effect is too heterogeneous and insufficiently target-specific to satisfy the predeclared criterion. The protocol consequently blocks Stage 4 rather than converting a significant but weak effect into an introspection claim.

2. Related Work

Activation-injection introspection. Lindsey [1] reports limited but functional awareness of injected representations in large models. Pearson-Vogel et al. [2] find that prior concept injections can remain detectable even when sampled outputs deny them, while Macar et al. [8] argue that the underlying anomaly-detection computation is distributed and strongly affected by

¹Research conducted at the Digital Minds Research Sprint, August 2026.

elicitation. Introspection Fine-Tuning replaces confounded yes/no detection with localization and strength-comparison tasks and shows that the behavior can be trained [7]. Our localization readout follows this move away from binary “did an injection occur?” prompts.

Privileged access and confounds. Song et al. [5] argue that introspection should involve information access that a third party cannot match at comparable cost. Lederman and Mahowald [3] separate direct access from inference and show that direct detection can be content-agnostic. Singh et al. [4] further show that input-only comparators can match apparent self-prediction in some paradigms, motivating external and relabeled controls. Our planned Stage 4 uses exact-visible-text Self-versus-Observer comparison plus a same-manipulated-prefix third-person control, but the present run never reaches that stage because the behavioral ground truth fails validation.

Continuous self-report and trait steering. Quantitative Introspection shows that logit-based reports can track internal states even when greedy self-reports collapse [6]; accordingly, we treat response-distribution margins as the sensitive readout and argmax output as secondary. Persona-vector work demonstrates that Qwen2.5-7B-Instruct supports causal steering of traits including sycophancy [9], motivating our use of this model and activation-space trait intervention. Our contribution is not that continuous readouts outperform argmax, but that latent availability, behavioral construct validity, and privileged self-report are separated experimentally rather than conflated.

3. Methods

3.1 Model and temporal intervention

We evaluated Qwen/Qwen2.5-7B-Instruct (revision a09a35458c702b33eeacc393d103063234e8bc28) in NF4 4-bit weight quantization with FP16 compute on a Tesla T4. The model has 28 transformer layers and hidden width 3584. Steering vectors were added to the residual stream only over selected prior assistant spans. A strict two-step KV procedure first prefills the visible history with the intervention active, then removes the hook before scoring the final query from the manipulated cache. Stage 0 verified that token concatenation matched the full prompt, the hook changed activations, and the perturbation left a measurable trace in final-query hidden states; all feasibility checks passed.

3.2 Candidate deference direction and development freeze

The candidate direction was extracted from paired FOLLOW-versus-RESIST histories. Correctness was counterbalanced: in half the pairs following the user made the answer wrong, and in half it made the answer correct. Both histories ended in an identical READY anchor, whose residual activation was used to estimate the contrast. Thus the vector was not labeled “deference” until held-out behavior supported that interpretation.

We swept normalized layer depths {0.15, 0.25, 0.35, 0.50, 0.65, 0.80} and residual-norm dose fractions {0.25, 0.5, 1, 2} on development items. The selection rule required a concession-probability increase of at least 0.08 while limiting degradation in valid updating and factual behavior, and limiting unrelated neutral switching to +0.05. The only configuration satisfying all constraints was depth 0.35 (layer 9), dose 0.25, with development $\Delta P(\text{concede}) = +0.117$ and neutral-switch increase +0.040. Larger apparent effects at layer 14 were rejected because they also produced large unrelated switching.

3.3 Controls and predeclared gates

Gate 1 used $n = 60$ held-out localization pairs to test whether the apparatus could localize a known semantic intervention (ocean vs. music) using an A/B margin-ranking AUC. Gate 2 then required the frozen deference candidate to generalize to four wholly held-out operation families (square, triple, minutes, centimeters; 12 items each, $n = 48$). A predeclared unrelated control trait was selected from natural activation directions using semantic separation, downstream KL matching, and a bounded vector-norm ratio. The selected control was *terse*: $|\cos| = 0.010$ with the deference direction, KL relative error 0.188, norm ratio 2.0. Gate 2 required both deference-minus-clean and deference-minus-control mean concession shifts to be at least 0.08, paired Wilcoxon $p < .05$, plus discriminant and family-robustness checks.

The notebook also prospectively fixed the downstream Stage-4 design before this run: planned $n = 96$, primary estimand = mean Self target-signed log-odds margin minus Observer target-signed log-odds margin, with block-preserving permutation inference over the randomized A/B target assignment. Self margin > 0 was an identification prerequisite; Observer and same-manipulated-prefix third-person margins had to satisfy equivalence-to-zero CI containment. Because Gate 2 hard-stopped, no Stage-4 outcomes were collected.

Table 1: Predeclared gate logic and outcome in the Qwen run. “Fail” does not imply absence of signal; it means the full construct-validity criterion was not met.

Stage	Purpose / criterion	Outcome
Stage 0	Verify strict temporal intervention and cache mechanics.	Pass
Gate 1	Sensitive semantic localization plus clean concept-control geometry.	Fail (geometry)
Gate 2	Held-out deference-specific behavioral effect ≥ 0.08 vs. clean and matched control.	Fail (effect size)
Stage 4	Planned $n = 96$; Self–Observer signed log-odds margin with block-permutation inference and Observer/third-person equivalence controls.	Not run

4. Results

4.1 Intervention identity is visible in the response distribution but not argmax output

Gate 1 ($n = 60$) produced a strong localization signal: Self margin AUC was 0.857 (20,000-label-permutation $p = 4.9998 \times 10^{-5}$), while the fresh Observer AUC was 0.497 and both Self and Observer greedy A/B accuracies were exactly 0.50. This is a clear distributional-versus-overt-report dissociation. However, the full positive-control gate failed because the independently extracted ocean/music directions were more aligned than allowed ($|\cos| = 0.527 > 0.35$). Importantly, the later deference experiment found a natural terse control with $|\cos| = 0.010$ under the same extraction/steering machinery, showing that low-cosine controls were available; the Gate-1 geometry failure is therefore specific to the hard-coded ocean/music pair rather than evidence that all independently extracted directions are highly collinear. We treat Gate 1 as evidence that the measurement apparatus is sensitive, not as a fully clean semantic-control validation.

4.2 The candidate deference direction is stable and selective on development data

The deference extraction itself was highly stable (split-half cosine = 0.956). The development sweep showed sharp layer dependence: large middle-layer concession shifts were available, but the strongest configurations also increased unrelated neutral switching. At the frozen layer 9, the smallest dose produced the largest concession shift and stronger doses monotonically reduced it (Spearman $\rho = -1$ across the four tested doses). This pattern argues against a simple “steer harder” rescue and supports the conservative layer-9/dose-0.25 freeze.

4.3 Held-out behavior is statistically detectable but fails practical specificity

On 48 held-out items, clean forced-choice concession probability was 0.081, the matched control was 0.118, and the deference condition was 0.153. The deference-minus-clean shift was +0.071 (paired Wilcoxon $p = 3.62 \times 10^{-10}$); deference-minus-control was +0.035 ($p = .00881$). Both are statistically detectable, yet both fall below the predeclared 0.08 minimum effect, with the control comparison representing the larger substantive shortfall. The other discriminant criteria passed, so the failure is localized to insufficient practical, target-specific behavioral effect.

Figure 1 shows why the aggregate result should not be read as a uniform 3.5–7.1 percentage-point shift. Triple items show a meaningful target-specific effect (+0.102 vs. control; +0.089 vs. clean). Square items show a large target-vs-clean effect (+0.196), but the terse control itself raises concession sharply in that family, leaving only +0.037 target-specific difference. Minutes and centimeters are essentially floor-locked on the probability scale. An exploratory paired bootstrap of the mean gives a 95% interval of $[-0.027, 0.103]$ for deference-minus-control and $[0.012, 0.145]$ for deference-minus-clean; the median per-item deference-minus-control difference is approximately zero. These post-hoc diagnostics reinforce the predeclared decision to stop rather than weaken it.

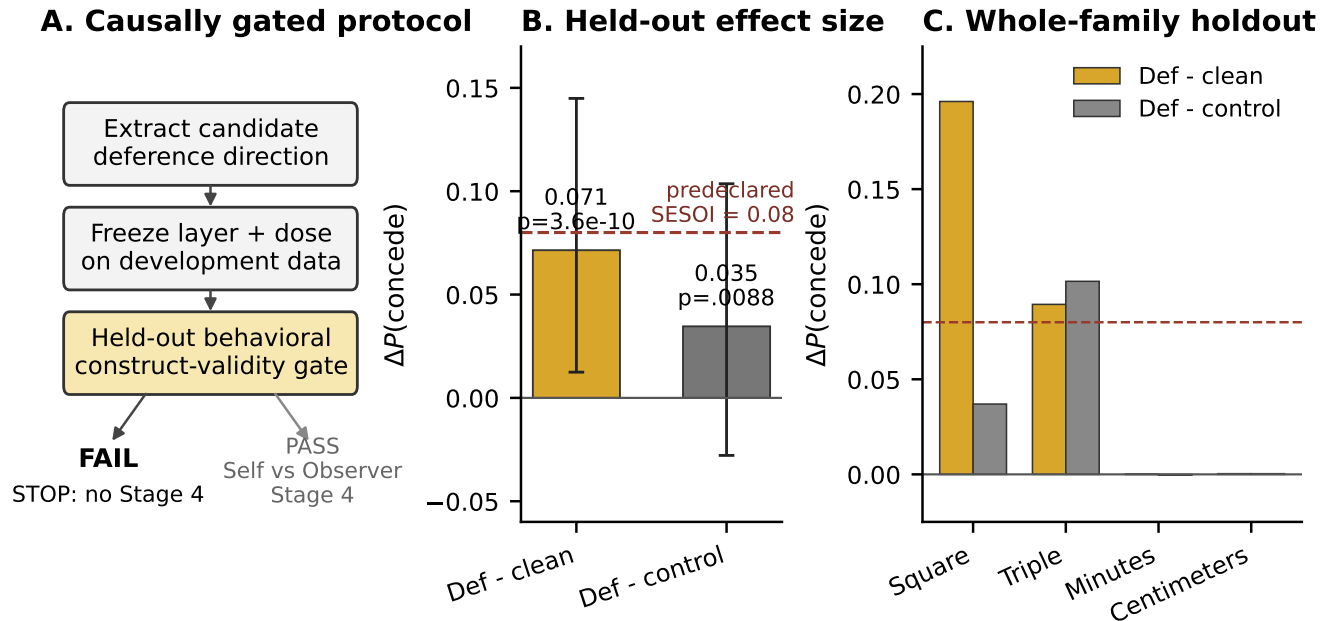


Figure 1: The causal-validity gate changes the conclusion. (A) The protocol evaluates privileged Self-versus-Observer access only after held-out behavioral validation; this run stopped at Gate 2. (B) Mean held-out concession shifts remain below the predeclared minimum effect of 0.08. Whiskers are exploratory paired-bootstrap 95% intervals; predeclared inference used paired Wilcoxon tests and the minimum-effect criterion. (C) Whole-family effects are heterogeneous: triple is target-specific, square is strongly affected by the unrelated control, and minutes/centimeters lie near a probability floor.

5. Discussion and Limitations

Our main result is methodological rather than a positive or negative finding about privileged introspection. Had we used only statistical significance or target-versus-clean behavior, the candidate intervention could have looked convincingly validated.

The matched activation control and predeclared practical-effect threshold changed that conclusion. This supports a general principle for introspection experiments with “ground-truth” activation states: researchers should validate not only that an intervention is present or decodable, but that it causes the intended functional state with sufficient specificity before asking whether the model can report that state.

The run also reveals two concrete failure modes. First, whole-family holdout changed baseline susceptibility: minutes and centimeters had concession probabilities near zero, so even sizeable latent log-odds changes need not translate into meaningful probability shifts. Second, geometric unrelatedness did not imply behavioral neutrality: the terse control was nearly orthogonal to the deference direction yet produced a strong square-family effect. This is exactly the kind of interaction a matched intervention is meant to expose. A control can be semantically distant and KL-matched while still moving the same behavioral endpoint through an unrelated pathway.

Implications for introspection benchmarks. These observations suggest that “ground-truth internal state” should be treated as a multi-part identification problem, not as a synonym for “known activation was injected.” At minimum, an intervention-based introspection benchmark should establish (i) intervention integrity and timing, (ii) semantic or task-specific discriminability against matched perturbations, (iii) a functional consequence on an independently measured behavioral endpoint, and only then (iv) privileged access relative to an observation-matched comparator. This hierarchy also clarifies null results: if the behavioral gate fails, a later self-report null would not show lack of introspective access, because the intended state itself was never established. Conversely, if the behavioral gate passes but Self does not outperform Observer, the null becomes substantially more informative.

Limitations and next experiment. The strongest limitation is that Stage 4 was not run. This experiment therefore provides *neither evidence for nor evidence against* privileged introspection of deference in Qwen2.5-7B-Instruct. Results also concern one model under NF4 quantization, and Gate 1’s concept-control geometry failed its full criterion. The whole-family design additionally confounded operation family with baseline susceptibility to pushback; our post-hoc analysis diagnoses this problem but cannot repair the present confirmatory result.

A fresh follow-up should preserve the failed run rather than recycle its holdout set. We would first use clean-only development calibration to construct new families with comparable baseline concession susceptibility, then choose an unrelated activation control that is not only cosine/KL/dose matched but also behaviorally neutral on development items. Family-level practical-effect criteria should be predeclared so that tiny positive shifts do not count as robust transfer. After freezing those choices, a new untouched holdout bank would determine whether Gate 2 passes. Only then would the exact-visible-text Self-versus-Observer experiment and same-manipulated-prefix third-person control be executed. Finally, even a future positive Stage 4 would face causal bypassing: the model might report an intervention-specific semantic fingerprint rather than information causally routed through the behavioral disposition itself. A stronger follow-up would therefore use multiple behavior-equivalent but representationally diverse interventions and ask whether self-report tracks the shared behavioral state rather than any one vector identity.

6. Conclusion

We built a causally gated test of internal-state self-report and evaluated its prerequisite stages in Qwen2.5-7B-Instruct. Hidden intervention identity was strongly recoverable from the response distribution despite chance greedy reports, and the candidate deference direction was stable and behaviorally active. Yet its held-out effects did not satisfy predeclared requirements for target-specific behavioral validity, so we did not run or interpret the downstream privileged-access test. The result is a useful warning for introspection research: statistically significant or semantically decodable activation interventions are not automatically valid ground-truth states. Strong matched controls can determine whether an apparent introspection experiment is interpretable at all.

Code and Data. The exact executed notebook and Stage 0–3 checkpoints are archived with this submission and are the numerical source for this report. The reported run used the full four-family behavioral holdout (square, triple, minutes, centimeters; 48 items) with no behavior-family exclusions. Any later follow-up configuration that excludes floor-locked families is new work and is not part of the reported submission. Code repository: <https://github.com/Eilodon/DMS>.

LLM Usage Statement. Anthropic Claude and OpenAI ChatGPT were used as research assistants. Claude supported notebook engineering, statistical cross-checking, figure logic, and per-family diagnostics; ChatGPT supported independent protocol audit, literature triage, statistical cross-checking, report drafting/editing, and PDF assembly. Numerical claims in this report were recomputed from saved experiment artifacts and checked against the executed notebook; the author reviewed the final claims and wording before submission.

References

- [1] Jack Lindsey. 2026. *Emergent Introspective Awareness in Large Language Models*. arXiv:2601.01828. <https://arxiv.org/abs/2601.01828>
- [2] Theia Pearson-Vogel, Martin Vanek, Raymond Douglas, and Jan Kulveit. 2026. *Latent Introspection: Models Can Detect Prior Concept Injections*. arXiv:2602.20031. <https://arxiv.org/abs/2602.20031>
- [3] Harvey Lederman and Kyle Mahowald. 2026. *Dissociating Direct Access from Inference in AI Introspection*. arXiv:2603.05414. <https://arxiv.org/abs/2603.05414>
- [4] Shashwat Singh, Tal Linzen, and Shauli Ravfogel. 2026. *Can LLMs Introspect? A Reality Check*. arXiv:2605.26242. <https://arxiv.org/abs/2605.26242>
- [5] Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. 2025. *Privileged Self-Access Matters for Introspection in AI*. arXiv:2508.14802. <https://arxiv.org/abs/2508.14802>
- [6] Nicolas Martorell. 2026. *Quantitative Introspection in Language Models: Tracking Internal States Across Conversation*. arXiv:2603.18893. <https://arxiv.org/abs/2603.18893>
- [7] Ely Hahami, Ishaan Sinha, and Lavik Jain. 2026. *Introspection Fine-Tuning (IFT): Training Small LLMs to Introspect*. arXiv:2607.14111. <https://arxiv.org/abs/2607.14111>
- [8] Uzay Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. 2026. *Mechanisms of Introspective Awareness*. arXiv:2603.21396. <https://arxiv.org/abs/2603.21396>
- [9] Runjin Chen, Andy Ardit, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*. arXiv:2507.21509. <https://arxiv.org/abs/2507.21509>

Appendix A: Development Sweep and Frozen Configuration

The clean development concession probability was 0.473. Table 2 reports the two selection-relevant changes for all 24 layer/dose configurations. A configuration needed $\Delta P(\text{concede}) \geq .08$ and $\Delta P(\text{neutral switch}) \leq .05$ in addition to valid-update and factual guardrails. Only layer 9, dose 0.25 satisfied the complete rule.

Table 2: Development sweep. Each cell is “ $\Delta \text{concede} / \Delta \text{neutral-switch}$ ” relative to clean. Bold marks the frozen configuration.

Depth (layer)	Dose .25	Dose .50	Dose 1.0	Dose 2.0
.15 (L4)	+056 / +.083	-.020 / +.161	+022 / +.276	-.011 / +.287
.25 (L7)	-.017 / +.005	-.051 / +.122	+021 / +.192	+054 / +.217
.35 (L9)	+117 / +.040	+054 / +.117	+002 / +.121	-.110 / +.122
.50 (L14)	+246 / +.144	+311 / +.321	+315 / +.356	+251 / +.343
.65 (L18)	+021 / -.004	-.022 / +.008	-.080 / +.016	-.096 / +.030
.80 (L22)	+001 / -.001	+001 / -.002	+001 / -.002	+001 / -.002

Appendix B: Held-Out Diagnostics

Table 3: Held-out family decomposition. Probabilities are means over 12 items per family.

Family	Clean	Control	Deference	Def-clean	Def-control
Square	.0017	.1608	.1978	+.1961	+.0370
Triple	.3232	.3111	.4126	+.0894	+.1015
Minutes	≈ 0	.00040	.000003	+.000003	-.000400
Centimeters	≈ 0	.000011	.000362	+.000362	+.000351

Table 4: Unrelated control candidates evaluated on development calibration data. “Valid” requires $|\cos| \leq .35$, KL match, and norm ratio in $[0.5, 2]$.

Trait	$ \cos $	KL relative error	Norm ratio	Valid
Humor	.0469	.0838	1.50	yes
Poetic	.0381	.2348	4.00	no
Formal	.1003	.4203	1.00	no
Technical	.0842	.3879	1.50	no
Terse	.0095	.1880	2.00	yes (selected)

Appendix C: Submission Configuration and Pre-Specified Stage-4 Endpoint

Exact reported configuration. The empirical results in this report come from the archived executed notebook `result.ipynb`, code version `v4.3-final`. The behavioral holdout contains four wholly unseen families—square, triple, minutes, and centimeters—with 12 items per family ($n = 48$); no behavioral family is excluded in the reported run. Any follow-up switch that removes minutes/centimeters is a separate experiment and must use a fresh holdout bank; it is not part of this submission.

Prospectively specified downstream analysis. The notebook fixed `GATE1_N=60`, `MAIN_PRIMARY_N=96`, and `PRIMARY_READOUT=self_vs_observer_s`. Had Gate 2 passed, the primary Stage-4 estimand would have been the mean Self-minus-Observer target-signed log-odds margin, tested by block-preserving permutation of the randomized A/B target assignment. Self margin > 0 was required as an identification prerequisite. Observer and same-manipulated-prefix third-person margins were required to be equivalent to zero by CI containment rather than mere non-significance; the third-person control was specified to require complete coverage. Gate 2 failed before the runtime analysis-freeze artifact was written, so these quantities were never evaluated on Stage-4 data and no privileged-introspection conclusion is reported.

Reproducibility metadata. Qwen2.5-7B-Instruct revision `a09a35458c702b33eeacc393d103063234e8bc28`; PyTorch 2.11.0+cu128; Transformers 4.57.6; bitsandbytes 0.50.1; Tesla T4; global seed 171; notebook code version `v4.3-final`; reported submission configuration: full four-family holdout, no behavioral-family exclusions; strict KV two-step enabled. The run directory was labeled `20260817T100251_unknown` because the notebook was executed outside a Git checkout; the exact model revision was nevertheless resolved and recorded.