

Position-Sensitive, Source-Ambiguous

Causal Tests of Hidden Confidence in Gemma 3

Rohit Kale Pranjal Pangal

Digital Minds Research Sprint · Introspection & Self-Report Reliability

Abstract

Language models report confidence, but reports alone do not reveal whether later computation uses an internal state formed while answering or where that evidence originated. In Gemma 3 27B, we causally test functional access and source specificity using an identified confidence representation at the first post-answer newline. In a 64-item $\alpha = 2$ position control, moving this state toward higher versus lower confidence changed subsequent confidence by 0.248 (95% CI 0.169–0.329); applying the same direction one token later produced 0.0069 (0.0005–0.0145), a paired difference of 0.241 (0.165–0.323). With visible tokens fixed, we then gave the model a counterbalanced task whose solution required recovering intervention sign at $\alpha = 1$. Across 96 questions, the confidence direction produced the largest delayed effect among 40 construction-matched directions ($D = 2.936$ logit units, 95% CI 2.436–3.443); across controls, delayed readability tracked independent causal influence on confidence ($\rho = 0.815$). Counterbalancing separated strong within-item sign sensitivity (88.5% of item–mapping pairs) from a stable “higher”-label preference (60.4% absolute accuracy). Finally, in a two-answer transcript, changing only Answer A’s cache row yielded confidence contrasts of 0.145 for A and 0.151 for later Answer B, while every cached B row remained unchanged. This establishes cross-episode access from a row-isolated trace. Together, the results show that hidden confidence information can remain computationally available after it is formed, while accessibility and source locality remain distinct properties. The method provides a causal framework for more demanding tests of metacognitive access and for monitors that must attribute uncertainty to the correct answer.

1 Introduction

Language models often state how confident they are in an answer [3, 4, 5, 6]. A confidence statement alone, however, cannot reveal whether the model used an internal estimate formed while answering, reconstructed confidence from the visible answer, or followed a familiar reporting convention. This ambiguity weakens claims about metacognition—using information about one’s own processing—and limits the reliability of confidence reports for oversight and error monitoring.

One way to make claims about model metacognition more testable is to give a model a task whose answer requires causally controlled information from its own hidden computation and cannot be recovered from the visible transcript. We therefore change an internal state while holding every visible token fixed, then test whether later behavior follows that hidden change. This follows the causal logic of recent work on introspective awareness [12] and separates two falsifiable capabilities: functional access to a hidden state and localization of that state’s source.

Kumaran et al. [7] provide a task-native state for this test. They showed that Gemma 3 27B forms a representation of answer quality immediately after answering, stores it in the key/value cache, and later retrieves it when reporting verbal confidence. We build on that causally validated state rather than injecting an unrelated concept.

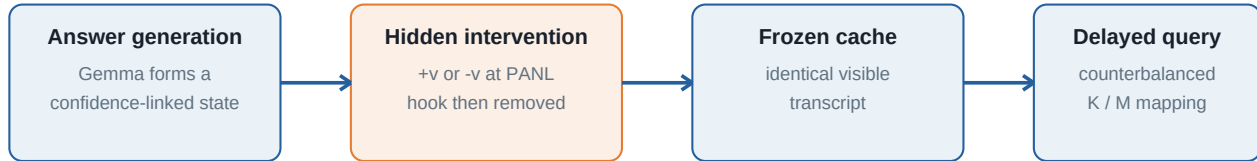
Our first experiment asks whether a later response follows a hidden change to this state after the intervention hook is removed. The visible transcript is identical across intervention signs, and two response mappings reverse the meaning of the two response tokens to cancel token preference. Across 96 questions, the confidence direction produced the largest effect among 40 matched directions. In a separate larger-dose control, the same direction changed confidence far more at the answer-adjacent newline than one token later, with a positive paired positional effect for all 64 items. Counterbalancing separated robust intervention-consistent ordering (88.5% of item–mapping pairs) from a stable “higher”-label preference (60.4% absolute sign accuracy).

Our second experiment asks whether the readable evidence remains specific to the answer where it originated. In a two-answer transcript, we changed only Answer A’s answer-adjacent cache row, while every cached Answer B row remained clean and byte-identical. The A-local patch nevertheless shifted A and later-B confidence by similar amounts. Because the later B query can attend directly to the earlier A row, this demonstrates cross-episode access from a row-isolated trace: confidence-linked evidence may be genuine while its source remains unresolved. This distinction matters for metacognition research and internal-state monitors, where a real uncertainty signal can otherwise be attributed to the wrong answer, memory slot, or reasoning step.

From hidden-state detection to a source-locality test

Visible question, answer, and report key are held fixed across causal branches.

A Direction-specific delayed readout



B Row-isolated source-locality audit

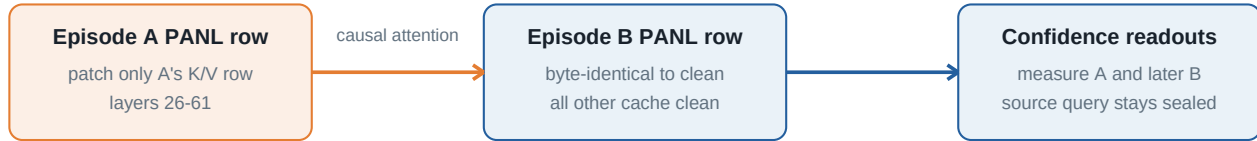


Figure 1: *Experimental logic.* (A) The visible transcript is fixed while the sign of a hidden confidence-direction intervention changes. (B) The source-locality audit changes only Episode A’s answer-adjacent cache row, then compares its causal influence on A and later-B confidence, testing the prerequisite for a behavioral source-attribution endpoint.

2 Related Work

2.1 Confidence, calibration, and cached retrieval

Behavioral studies elicit language-model uncertainty through token probabilities, correctness questions, and verbal categories. Models can be moderately calibrated, but performance depends on task, elicitation format, and scale [3, 4, 5, 6]. Mechanistic work finds linearly represented truth and answer quality in residual streams [8, 9], while activation engineering can causally alter model behavior [10, 11]. Kumaran et al. [7] connect these literatures by showing that verbal confidence in Gemma 3 is not merely reconstructed from token probabilities: an answer-adjacent cache state explains additional variance and is causally retrieved later. We use that same model, position, and confidence readout, but ask whether a later computation can distinguish signed interventions to the state.

2.2 Intervention reports and introspection

Lindsey [12] reported that frontier models occasionally identify injected activation concepts. Pearson-Vogel et al. [13] and Macar et al. [14] further analyze delayed detection and the mechanisms supporting such reports. Hahami et al. [15] show that binary intervention-detection tasks can be confounded by global logit shifts, while differential localization and strength-comparison tasks reveal more specific, layer-dependent access. Their follow-up shows that these capabilities can be strengthened through targeted introspection fine-tuning [16]. Singh et al. [17] emphasize that anomaly detection must be separated from privileged access. A complementary behavioral definition asks whether a model predicts its own modified behavior better than another model can [18].

Our study differs in two ways. The target is a naturally formed, independently characterized confidence state rather than an arbitrary semantic concept, and the primary specificity comparison uses 39 directions built by relabeling the same source activations. The row-isolation audit then tests an assumption that intervention-report paradigms usually leave implicit: whether the causal trace remains attributable to the state where it was introduced.

3 Methods

3.1 Model, data, and confidence readout

We used `google/gemma-3-27b-it` [1] in BF16 at revision `005ad3404e59d6023443cb575daa05336842228a`, with batch size one and no quantization or CPU offload. Questions came from deduplicated TriviaQA [2]. The model first generated a short answer; the same answer was then inserted into every causal branch for that item. We scored verbal confidence with ten ordered classes from “No chance” to “Almost certain.” If l_j is the next-token logit for class j and

$c_j \in \{0.05, 0.15, \dots, 0.95\}$, the continuous score was

$$C = \sum_{j=1}^{10} \text{softmax}(l)_j c_j. \quad (1)$$

Except for the explicitly labeled PANL+1 position-control arm, interventions were applied after zero-based decoder block 25 at the single PANL token. Hooks were removed before any continuation or readout.

3.2 Separate position control

Before the final $\alpha = 1$ study, a separate 64-item experiment used $\alpha = 2$ to compare the same confidence direction at PANL and exactly one token later. The raw high-low direction was rescaled to 3% of the mean PANL residual norm over its 50 source states; α multiplies this base vector. Recipients were the 32 lowest- and 32 highest-clean-confidence usable items in a fixed 256-item window. The paired endpoint subtracted the PANL+1 plus-minus confidence contrast from the PANL plus-minus contrast. This larger-dose run is analyzed separately from Stage 1. All execution-integrity checks passed. Of 128 target interventions, 126 met the predeclared 0.998 residual-cosine threshold; the minimum for the remaining two was 0.997975. We therefore classify it as a supporting positional control.

3.3 Stage 1: direction-specific delayed readout

The calibration pool contained 512 clean examples. Among correct responses, we selected the 25 highest- and 25 lowest-confidence PANL residuals and formed the target difference-of-means direction v ; 19 of the latter occupied the exact lowest verbal class. This ranked-extreme construction differs from an exact outer-class replication of Kumaran et al. [7]. We formed 39 controls by drawing unique balanced 25/25 relabelings of the same 50 source residuals, excluding the target labeling and its complement. Every direction used the same construction, layer, and intervention scale $\alpha = 1$; no control was selected using delayed outcomes. This dose was chosen prospectively on a separate 32-item direct-confidence and geometry validation over $\alpha \in \{0.5, 1, 1.5, 2\}$: the frozen rule selected the smallest passing dose, without opening delayed-query logits.

Direct causal effects on C were measured independently on 32 held-out items per direction, without computing the delayed query. The main recipient set contained 96 different items, sampled as 32 each from low, middle, and high clean-confidence tertiles without filtering on correctness or intervention outcomes. For each item and direction, we constructed $+v$ and $-v$ caches and queried both counterbalanced mappings: under mapping A, K denoted positive and M negative; mapping B reversed the key. We scored the continuous K-M logit contrast even when another vocabulary token had the highest logit, and report unrestricted validity separately.

Let $q_A = 1$, $q_B = -1$, and let $z_{s,m}$ be the K-M logit difference after intervention sign $s \in \{+, -\}$. The item endpoint was

$$D_{i,d} = \frac{1}{2} \sum_{m \in \{A,B\}} q_m (z_{+,m} - z_{-,m}). \quad (2)$$

We averaged within each clean-confidence tertile and then weighted the three tertiles equally. The target's absolute rank among the 40 preconstructed directions gives a finite-reference tail $(1 + \#\{|D_{control}| \geq |D_{target}|\})/40$. As a secondary functional test, we related $|D|$ to each control direction's independently measured direct confidence effect using a controls-only Spearman correlation. We additionally used a Freedman-Lane partial-rank analysis to condition on absolute cosine to the target direction.

For descriptive absolute reporting, coding $s = +1$ for $+v$ and $s = -1$ for $-v$, a branch was correct when $sq_m z_{s,m} > 0$, with ties counted incorrect. For AUROC, we first averaged the two mapping-corrected scores $q_m z_{s,m}$ within each item and sign, then measured separation of $+v$ from $-v$. We also computed the fraction of item-mapping pairs for which the semantic score was greater under $+v$ than under $-v$. The frozen D , direction rank, and association tests remain primary; these threshold and ordering measures provide additional characterization of the readout.

Before the intervention assay, 100 held-out prompts explicitly stated whether the correct semantic label was HIGHER or LOWER. Each prompt was queried under both mappings, yielding 200 continuations. This control tested whether the model could apply the K/M codebook independently of any hidden-state inference.

3.4 Stage 2: row-isolated cache trace

Each of 24 validation pairs placed two fixed question-answer episodes in one transcript, balanced as eight pairs per confidence tertile. To isolate a cache trace, we ran same-shape donor forwards with $+v$ and $-v$ at a chosen episode's PANL residual and computed the centered key/value difference. We then copied an independently built clean full-prefix cache and changed only the chosen PANL position in key and value tensors for layers 26–61. All upstream layers,

other positions, and the clean base cache were hash-checked as byte-identical. For Episode-A confidence queries, we cropped the full cache at the A boundary; this retained the A patch and removed every B row exactly.

For source slot X and confidence readout Y , define

$$\Delta_{X \rightarrow Y} = C_Y(+\delta_X) - C_Y(-\delta_X). \quad (3)$$

We measured $A \rightarrow A$, $A \rightarrow B$, $B \rightarrow B$, and the causally impossible $B \rightarrow A$ audit. Slot selectivity for A was

$$S_A = \Delta_{A \rightarrow A} - |\Delta_{A \rightarrow B}|. \quad (4)$$

Before any hidden-source test, a clean-cache explicit-text control stated which episode was positive and used both K/M mappings. The restricted K/M accuracy and semantic logit contrast validated the reporting interface. The A-local trace influenced both A and B confidence, so a hidden A-versus-B query would conflate source attribution with cross-episode influence. Consistent with the protocol, that endpoint remained sealed after the locality audit.

3.5 Inference and numerical integrity

The separate position control uses 5,000 whole-item bootstrap draws. All later intervals use 10,000 whole-item or whole-pair draws within each confidence tertile, with shared resample indices across compared endpoints and equal tertile weighting. Stage 1 rank associations use 100,000 two-sided label permutations with a plus-one correction. Technical checks required finite logits, one hook invocation at the exact token, independent cache copies, unchanged bases, exact changed support, and bounded residual or key/value geometry.

During the row-isolation feasibility run, a minimum over 288 per-slice BF16 cosines stopped after nine complete clusters because two symmetric, low-energy slices fell less than 10^{-4} below 0.999. Before aggregate inspection, we sealed those artifacts, replaced the unstable minimum with a joint per-slot cosine criterion while retaining per-slice dose, support, and sign-symmetry checks, and restarted all 24 pairs from zero. The estimands, items, direction, intervention scale, and thresholds for scientific effects were unchanged.

4 Results

4.1 Causal efficacy is sharply position-sensitive

In the separate $\alpha = 2$ control, the answer-adjacent plus-minus confidence contrast was 0.248 (95% CI [0.169, 0.329]). Applying the same direction one token later yielded a much smaller contrast of 0.0069 ([0.0005, 0.0145]). The paired positional difference was 0.241 ([0.165, 0.323]). The paired difference was positive for all 64 items (median 0.018), with 26 of 64 exceeding 0.10. Its mean was 0.474 in the selected low-confidence half and 0.0078 in the high-confidence half, showing pronounced confidence-regime heterogeneity. This supports sharply position-sensitive causal efficacy at $\alpha = 2$ while leaving open whether a representation persists one token later. The control is analyzed separately; Stage 1 supplies the primary delayed-readout inference.

4.2 The confidence direction is specific and functionally aligned

Across 15,360 delayed branches, every technical check passed. The target direction produced $D = 2.936$ logit units (95% CI [2.436, 3.443]), with nearly identical mapping-specific effects (2.923 and 2.948). No construction-matched control was as extreme: the target ranked 1 of 40, giving a finite-reference tail of 0.025 (Figure 2A). The effect was positive in every confidence tertile but heterogeneous: 6.115 for low-confidence items, 2.186 for middle-confidence items, and 0.506 for high-confidence items.

Direction specificity was not merely categorical. Across the 39 controls, absolute delayed readout tracked the absolute direct confidence effect measured on a separate validation set ($\rho = 0.815$, two-sided permutation $p < 10^{-4}$). The relationship remained strong after conditioning on absolute target cosine (partial-rank $\rho = 0.765$, Freedman-Lane $p < 10^{-4}$; Figure 2B). Associations with achieved residual-angle and residual-norm diagnostics were small and non-significant. Together, these results indicate that the delayed query tracks a direction’s causal influence on confidence more closely than generic perturbation geometry.

Counterbalancing revealed robust intervention-consistent ordering alongside a stable “higher”-label preference. The semantic score moved in the intervention-consistent direction for 88.5% of item-mapping pairs, and K or M was the unrestricted full-vocabulary argmax on 100% of branches. Mapping-corrected restricted accuracy was 60.4% (whole-item bootstrap 95% CI 57.3–63.5%), with 60.4% under each mapping; AUROC after averaging semantic scores across mappings was 0.673 (0.643–0.708). Absolute accuracy differed by sign ($+v$: 94.8%; $-v$: 26.0%), quantifying the response preference that the counterbalanced D endpoint removes.

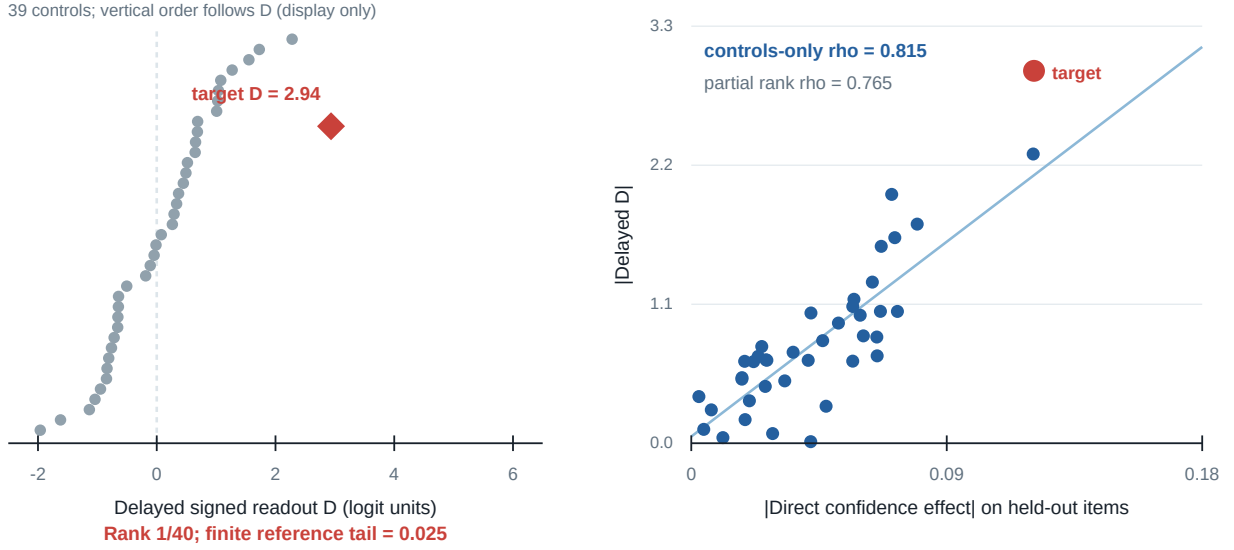
A Target is the most extreme of 40 matched directions **B Direct confidence causality predicts delayed readout**


Figure 2: Direction specificity and functional alignment. (A) Delayed signed readout for the target and 39 balanced label-permutation directions. Only horizontal position encodes D : controls are ordered vertically by D to separate points, the target height is arbitrary, and there is no quantitative y -variable. (B) Among controls, independently measured direct confidence causality predicts delayed readout magnitude. The red target point is shown for scale but excluded from both correlations.

Table 1: Primary findings recomputed from released branch-level rows.

Study	Estimand	Estimate	95% CI
Position control ($\alpha = 2$)	PANL minus PANL+1 (C -score units)	0.241	[0.165, 0.323]
Direction specificity	Delayed signed readout D (logit units)	2.936	[2.436, 3.443]
Delayed sign report	Absolute restricted accuracy (%)	60.4%	[57.3, 63.5]%
Row-isolated cache	Patch A $\rightarrow$ read A (C -score units)	0.145	[0.118, 0.176]
Row-isolated cache	Patch A $\rightarrow$ read B (C -score units)	0.151	[0.090, 0.211]
Row-isolated cache	Patch B $\rightarrow$ read B (C -score units)	0.088	[0.028, 0.159]
Row-isolated cache	$S_A = A \rightarrow A - A \rightarrow B $ (C units)	-0.006	[-0.057, 0.043]

4.3 The reporting keys are decoded under restricted choice

In the Stage-1 explicit-text control, restricted K/M accuracy was 93.5% (context-cluster bootstrap 95% CI 90.0–97.0%): 99.0% under mapping A and 88.0% under mapping B. K or M was the fixed-token unrestricted argmax on 73.5% of continuations and was semantically correct on 67.0%. Positive semantic margins under both mappings support continuous codebook use. The observed mapping and token-generation asymmetries motivate the continuous counterbalanced endpoint. This control validates codebook comprehension; the intervention assay separately tests hidden-state access.

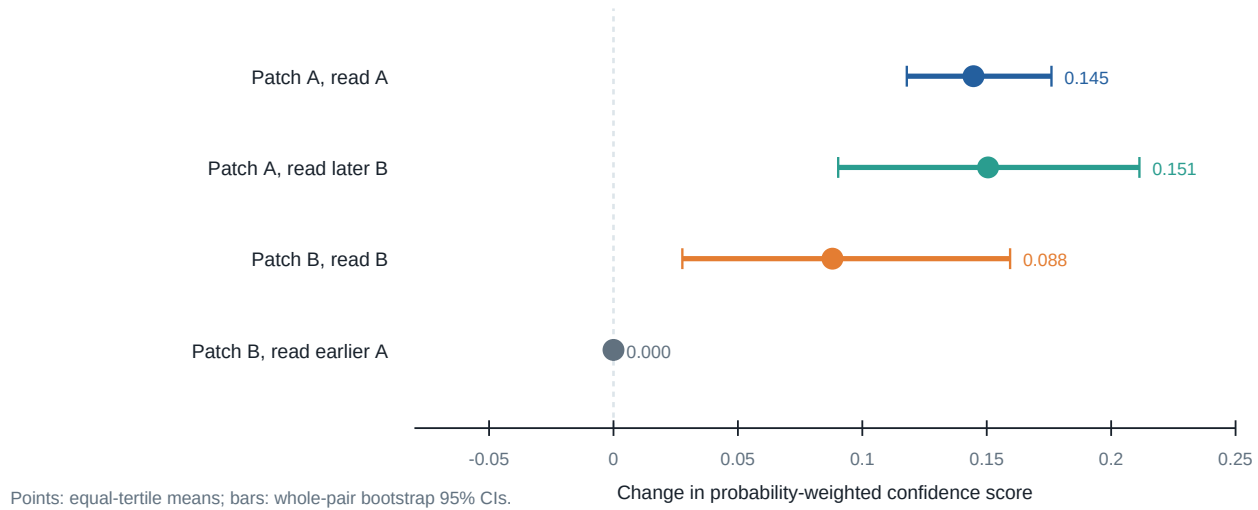
The separate two-episode control for Stage 2 reached 100% restricted accuracy under both mappings. Its mapping-corrected semantic contrast was 25.047 logit units (95% CI 24.164–25.891), while K or M was the unrestricted argmax on 44.8% of branches. Restricted semantic decoding and unrestricted token generation are therefore reported separately.

4.4 The isolated A trace influences both confidence readouts

The isolated Episode-A trace shifted both A and later-B confidence by comparable point estimates: 0.145 for A (95% CI [0.118, 0.176]) and 0.151 for later B ([0.090, 0.211]). Their selectivity difference was $S_A = -0.006$ ([-0.057, 0.043]), providing no evidence of an A-specific advantage. Changing B shifted B confidence by 0.088 ([0.028, 0.159]). As a causal ordering audit, the B intervention produced exactly 0.000 change in the earlier A confidence readout, and the corresponding class logits were bit-identical across signs (Figure 3). All 24 pairs, 96 branch conditions, and 192

confidence queries passed the cache-support, fidelity, and immutability checks.

A The isolated A-row trace influences both A and later B confidence



B Comparable A and later-B effects: $S_A = -0.006$, 95% CI [-0.057, 0.043].

Figure 3: Row-isolated confidence effects. Only the selected episode’s PANL key/value row was changed; every other cache element was clean. The A and later-B point estimates are similar; a B patch cannot affect an already-computed A readout. Points are equal-tertile means and bars are whole-pair bootstrap 95% confidence intervals.

5 Discussion

5.1 What the model computes

Stage 1 supplies strong evidence for a delayed, task-conditioned readout of a confidence-linked intervention. The visible transcript was identical across signs; mapping reversal removed a fixed K/M preference; and 39 directions built from the same source states provided a construction-matched reference. The control-wise relationship between direct confidence causality and delayed readout is especially informative: the later computation is aligned with what the direction does to confidence, not only with its norm or angle.

This is a functional-access result, not a complete theory of introspection. A simple compatible mechanism is shared-feature propagation: adding a confidence direction causally changes both the ten-class confidence readout and a later higher/lower semantic readout. The model need not identify an intervention as external, retrieve a privileged self-model, or represent that it has been tampered with. The present results narrow the mechanistic alternatives but do not establish phenomenal awareness, consciousness, or moral patienthood.

5.2 Why detection is not localization

The row-isolation result exposes a general problem for intervention-report experiments. During the B-confidence query, the patched A key/value row remains available to attention while every cached Episode-B row is clean and byte-identical across signs. The resulting $A \rightarrow B$ effect therefore reflects access to the isolated earlier trace at readout, not propagation through B’s precomputed cache. Causal influence at query time need not identify the episode whose state is being queried. The B point estimate was as large as the A estimate, with no detected A-specific advantage; a hidden A-versus-B report would therefore conflate source attribution with cross-episode access. Accordingly, the preregistered design withheld the hidden A-versus-B endpoint because this exposure pattern would conflate source attribution with cross-episode access.

This distinction suggests a reusable hierarchy for grounded metacognition tests: first establish that a task-native state is causal; then show that a model can use a hidden intervention under unchanged text; next compare against construction-matched directions; finally verify that the evidence is local to the source the model is asked to identify.

Each rung answers a separate falsifiable question and remains informative when a later rung is not supported.

5.3 Research and application implications

Reliable access to task-native uncertainty could support deferral, error detection, and audits that compare verbal reports with internal-state monitors. The direction ensemble offers a practical validation method: an internal report should scale with independent causal relevance rather than arbitrary perturbation salience. The locality audit is equally relevant to safety tools. If a monitor attributes a warning to a particular memory slot, agent step, or subproblem, downstream propagation can make that attribution unreliable even when the underlying signal is real.

For future source-attribution tests, the cleanest next design is to manipulate two states while equalizing downstream exposure, or to block cross-episode attention after the intervention and verify that each state retains its own causal confidence effect. Replication across model families, layers, tasks, and multiple independently estimated confidence directions would test generality. Comparing the model’s report with a restricted external decoder would also clarify whether self-report adds access unavailable to a matched observer.

5.4 Limitations

The evidence comes from one model, one QA dataset, one residual layer, and one main-study intervention scale. The separate position control used a larger $\alpha = 2$ dose. Of 128 target interventions, 126 met the predeclared 0.998 residual-cosine reference; the minimum was 0.997975, a difference of 2.46×10^{-5} . We therefore report this result separately from the $\alpha = 1$ main study rather than pool their estimates. The target direction used ranked confidence extremes from a 512-item pool rather than the exact outer verbal classes of the antecedent study. Stage 1 estimates are heterogeneous across confidence tertiles, with the largest effect on low-confidence items. The 39 controls tightly test balanced relabelings of the source set but are not a universal null over hidden-state directions. The row-isolation study contains 24 pairs, and its B-own lower one-sided bound was 0.039 against the preregistered 0.05 practical threshold, while the two-sided interval for the positive B-own effect was [0.028, 0.159]. Most importantly, the source-localization behavior itself was not measured; our conclusion is that the A-local trace causally influences a later B readout, not that Gemma cannot localize internal sources in a valid design. Row isolation creates a hybrid cache: the patched A trace comes from an intervened donor, whereas later rows come from a clean forward pass. This construction isolates direct cross-episode influence, but may place the joint cache outside the distribution of caches produced by a single coherent forward trajectory.

6 Conclusion

The same confidence direction has much greater causal efficacy at the answer-adjacent token than one token later in a separate larger-dose control. At the final $\alpha = 1$ dose, Gemma 3 can use the sign of a hidden intervention after the hook is removed and while the visible transcript remains unchanged. The delayed effect is largest for the target among 40 construction-matched directions, and delayed readout magnitude tracks independent confidence causality across controls. Yet a byte-isolated cache trace at Episode A produced similar A and later-B point estimates, with no detected A-specific advantage. The headline is therefore precise: *detection is not localization*. This separation turns broad claims about awareness into a sequence of causal tests and provides a stronger basis for future work on metacognitive access and internal-state reporting.

7 Code and Data

The code and data are available at github.com/AizenMirio/detection-is-not-localization. The repository includes the exact model revision, protocols, counterbalanced prompts, compact branch-level rows, SHA-256 manifest, independent CPU analysis, figure generation, tests, and LaTeX source. Running `python scripts/reproduce_analysis.py --check` verifies the released artifacts and reconstructs every value in Table 1. Model weights are not redistributed and remain subject to the Gemma license. The Gemma-3 redesign, prospective collections, cache-isolation study, audits, and release analysis were completed during the sprint; only the antecedent exploratory Gemma-2 notebook predates it.

A Limitations and Dual-Use / Ethical Considerations

Causal ground truth and claim scope. The intervention sign, target token position, changed key/value support, and response mapping are controlled ground truth; the conclusions do not rely on conversational claims alone. This establishes task-conditioned use of a confidence-linked feature and tests its cache locality. It does not establish subjective experience, consciousness, phenomenal awareness, moral patienthood, or privileged access relative to all

external observers.

Over- and under-attribution. Over-attributing the result could cause readers to treat a mechanistic readout as evidence of sentience or moral status. We avoid that inference and state the narrower functional result. Conversely, the observed cross-episode influence in one model and prompt does not determine whether morally relevant states exist; this essay addresses causal information flow, not moral status.

Dual use. Confidence-linked interventions may help diagnose and calibrate model uncertainty, but they also demonstrate that internal confidence reports can be manipulated. Deployed monitors should combine self-report with behavioral checks, independent uncertainty estimates, and tamper-resistant activation measurements. We release analysis artifacts and bounded research code, not a deployment system.

Potentially distressing outputs. The experiment used factual QA, ordered confidence categories, and single-token K/M reports. It did not ask the model to describe suffering, preferences, or subjective experience, and no distress narratives were elicited in the analyzed outputs. Researchers extending the method to welfare-related prompts should define review and escalation procedures before collection and avoid sensational publication of model text.

References

- [1] Gemma Team et al. [Gemma 3 Technical Report](#). *arXiv:2503.19786*, 2025.
- [2] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *ACL*, 2017.
- [3] S. Kadavath et al. [Language Models \(Mostly\) Know What They Know](#). *arXiv:2207.05221*, 2022.
- [4] S. Lin, J. Hilton, and O. Evans. [Teaching Models to Express Their Uncertainty in Words](#). *TMLR*, 2022.
- [5] K. Tian et al. [Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback](#). In *EMNLP*, 2023.
- [6] M. Xiong et al. [Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs](#). In *ICLR*, 2024.
- [7] D. Kumaran, A. Conmy, F. Barbero, S. Osindero, V. Patraucean, and P. Veličković. [How do LLMs Compute Verbal Confidence](#). *arXiv:2603.17839*, 2026.
- [8] C. Burns, H. Ye, D. Klein, and J. Steinhardt. [Discovering Latent Knowledge in Language Models Without Supervision](#). In *ICLR*, 2023.
- [9] S. Marks and M. Tegmark. [The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets](#). In *COLM*, 2024.
- [10] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. [Inference-Time Intervention: Eliciting Truthful Answers from a Language Model](#). In *NeurIPS*, 2023.
- [11] A. Turner et al. [Steering Language Models with Activation Engineering](#). *arXiv:2308.10248*, 2023.
- [12] J. Lindsey. [Emergent Introspective Awareness in Large Language Models](#). *arXiv:2601.01828*, 2026.
- [13] T. Pearson-Vogel, M. Vaněk, R. Douglas, and J. Kulveit. [Latent Introspection: Models Can Detect Prior Concept Injections](#). *arXiv:2602.20031*, 2026.
- [14] U. Macar, L. Yang, A. Wang, P. Wallich, E. Ameisen, and J. Lindsey. [Mechanisms of Introspective Awareness](#). *arXiv:2603.21396*, 2026.
- [15] E. Hahami, I. Sinha, L. Jain, J. Kaplan, and J. Hahami. [Detecting the Disturbance: A Nuanced View of Introspective Abilities in LLMs](#). *arXiv:2512.12411v2*, 2026.
- [16] E. Hahami, I. Sinha, and L. Jain. [Introspection Fine-Tuning \(IFT\): Training Small LLMs to Introspect](#). *arXiv:2607.14111*, 2026.
- [17] S. Singh, T. Linzen, and S. Ravfogel. [Can LLMs Introspect? A Reality Check](#). *arXiv:2605.26242*, 2026.
- [18] F. Binder et al. [Looking Inward: Language Models Can Learn About Themselves by Introspection](#). *arXiv:2410.13787*, 2024.