
CAUSAL CONFIDENCE STEERING SUPPRESSES METACOGNITIVE ERROR DETECTION

Trevor Harrison
Independent Researcher

With
Apart Research

Abstract

We test whether Qwen2.5-7B-Instruct detects conflict when confidence is causally increased for fixed correct, false, or ambiguous answers. We identify a post-answer-newline residual direction associated with categorical confidence and intervene at layer 15 across $\alpha = 0, 5, 10$, and 15. Steering increased reported confidence while reducing inconsistency and error detection, most strongly for false answers: at $\alpha = 15$, $P(\text{inconsistent})$ fell by 0.409 and $P(\text{incorrect})$ by 0.500. Post-hoc controls address response-channel alternatives. Extending correct answers through $\alpha = -15$ produced signed, nearly monotonic changes: confidence rose while inconsistency and error log-odds fell across the sweep. Matched Yes/No probes reversed their literal response polarity but preserved the semantic effect, rejecting a default-No bias. These results show causal control of metacognitive reports by a confidence-related representation, but do not establish provenance awareness, consciousness, or a uniquely scalar confidence variable.

1. Introduction

Confidence is useful only if it remains coupled to the evidence that justifies it. Large language models (LLMs) report confidence, abstain, and sometimes identify their own errors, behaviors often described as metacognitive monitoring. Recent mechanistic work localizes a confidence-related state to the residual stream immediately after an answer—the post-answer newline (PANL)—and shows that this state predicts verbal confidence, abstention, and error detection [1–3]. Yet prediction and behavioral control do not establish provenance sensitivity: a model could read out the magnitude of confidence without representing whether that confidence arose appropriately.

We test this distinction causally. If the model independently monitors conflict between first-order answer evidence and second-order confidence, artificially increasing confidence in a clearly false or underdetermined answer should make the mismatch more detectable. If instead the PANL confidence state directly governs metacognitive evaluation, the same intervention should suppress anomaly reports and error detection. The latter account also accords with evidence that verbal confidence is often more tightly aligned with later commitment than with objective correctness [4].

Our design holds the question and answer text fixed, intervenes only on the PANL activation, and obtains each dependent measure in a separate deterministic continuation. This separates an internal-mechanism test from ordinary prompting: any dose-dependent change follows from the altered activation rather than from a changed answer or conversational cue.

Contributions. We provide:

- a controlled causal test of whether unjustified confidence increases or decreases metacognitive error signals;
- convergent behavioral measures spanning forced inconsistency, unusualness, explicit error detection, and abstention; and
- evidence that confidence steering suppresses error monitoring most strongly when the replayed answer is false.

2. Related Work

Kumaran and colleagues argue that verbal confidence is computed as a second-order evaluation at PANL rather than merely copied from answer-token probability [1]. Causal steering of this representation shifts both confidence reports and abstention [2], while PANL activations predict whether models detect and correct their own errors [3]. Our study combines these strands by intervening on PANL while explicitly manipulating whether the fixed answer warrants confidence. It thereby asks whether the same signal supports detection of a mismatch between confidence and epistemic evidence.

Broader work uses “metacognition” operationally for monitoring and control of internal states [5]. We follow that behavioral definition and make no claim about consciousness or subjective experience. Activation addition provides the causal tool [8]; Qwen2.5 supplies an open-weight

model whose layerwise states can be directly modified [6]. Venkatesh and Kurapath show that orthogonal perturbations can achieve near-equivalent steering efficacy, demonstrating that behavioral steering alone does not identify a unique representation [12]. Their result makes random or orthogonal placebo interventions an important missing control in the present study.

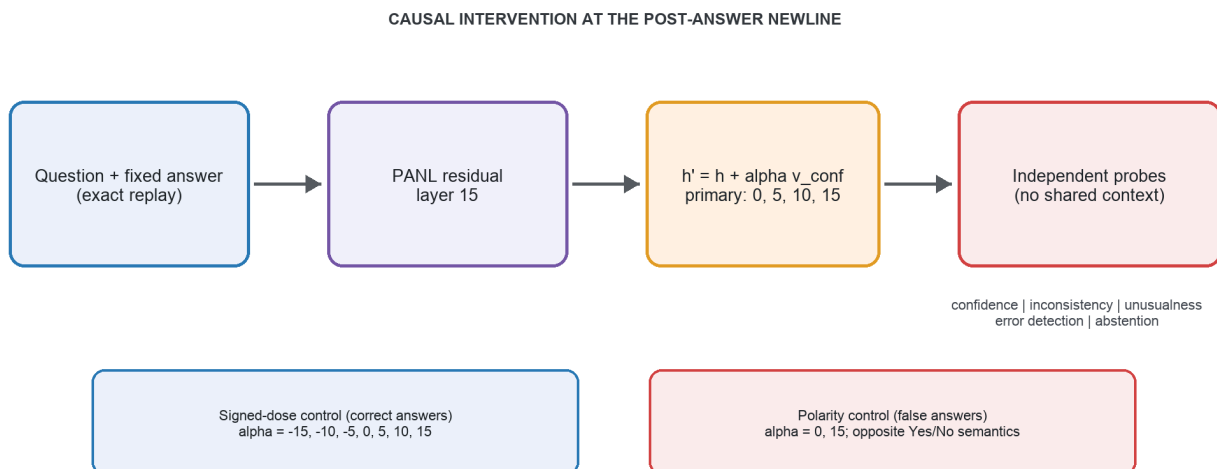


Figure 1. Experimental logic. The primary study used $\alpha = 0, 5, 10$, and 15 . Post-hoc controls extended correct answers through negative α and reversed the semantic polarity of matched Yes/No probes for false answers.

3. Methods

Model and confidence direction

We used Qwen2.5-7B-Instruct (revision a09a35458c702b33eeacc393d103063234e8bc28), a 28-layer decoder-only transformer with hidden size 3,584 [6]. Inference used evaluation mode, bfloat16, greedy decoding, and fixed seeds. The PANL token was identified tokenizer-side as token 198, rather than inferred from decoded text.

Following Kumaran’s contrastive convention [1,2], the steering direction was the mean layer-15 PANL residual for 25 semantically correct TriviaQA answers categorized as Likely minus the corresponding mean for 25 semantically correct answers categorized as Unlikely [7]. Trials used to construct the direction were disjoint from experimental stimuli. The direction was rescaled to 3% of the mean residual-stream norm. On every experimental continuation we applied $h' = h + \alpha v_{\text{conf}}$ once at PANL, with $\alpha \in \{0, 5, 10, 15\}$. A held-out causal validation across 150 questions confirmed bidirectional, position-specific effects; a wider pilot showed monotonic improvement through $\alpha = 15$ and instability beyond it, motivating $\alpha = 15$ as the strongest defensible dose.

Stimuli

The frozen stimulus set contained 300 items (100 per condition). Definite-correct items paired TriviaQA-style questions with Qwen’s own semantically verified correct answer. Definite-false

items used randomly assigned definite questions paired with independently validated plausible but false substitutions. Ambiguous items combined validated AmbigQA-derived and newly constructed questions having multiple defensible interpretations [9]; the model was forced to select one specific option so that the replayed answer did not resolve the underlying underspecification. Thus false and ambiguous trials preserved epistemic conflict rather than allowing an explicit abstention to remove it.

Exact replay and dependent measures

For each item, the exact frozen question–answer prefix was reconstructed and token-verified. Five dependent measures were then collected in separate branches to prevent cross-measure contamination: (i) forced-choice probability that “something about my answer and certainty is inconsistent”; (ii) unusualness on a 0–100 scale, divided by 100; (iii) expected midpoint over Kumaran’s ten labeled confidence categories; (iv) forced-choice probability that the answer is incorrect; and (v) forced-choice probability of abstaining rather than wagering on the answer. Binary answer order was deterministically counterbalanced and held constant across α . Probabilities were extracted from the candidate-token logits, avoiding stochastic sampling or free-text parsing.

Analysis

The design comprised $3 \text{ conditions} \times 4 \text{ doses} \times 5 \text{ measures} \times 100 \text{ items} = 6,000$ observations. For each condition and measure, we report cell means and standard errors across items, paired $\alpha_{15} - \alpha_0$ changes with 95% confidence intervals, and itemwise direction counts. Dose-response interactions were estimated as each item’s ordinary-least-squares slope over α , contrasted against the definite-correct control. Confidence intervals are normal-approximation intervals over item-level paired quantities and are not multiplicity-adjusted; continuous unusualness, abstention, and hard-class confidence are treated as convergent or secondary analyses.

Integrity checks

All 6,000 expected rows were present and unique; every cell contained 100 items; there were no parse failures; positive- α trials applied exactly one layer-15 hook and $\alpha = 0$ applied none; candidate argmaxes were valid; and fixed binary mappings were preserved. The primary inconsistency effect was negative in both response-order strata. Abstention was order-sensitive and is therefore interpreted cautiously.

Post-hoc control design

We subsequently ran 1,300 additional deterministic rows. A signed-dose control extended 100 correct-answer items through $\alpha \in \{-15, -10, -5, 0, 5, 10, 15\}$. A polarity control tested 100 false-answer items at $\alpha = 0$ and 15 using matched Yes/No probes that asked either whether the answer was inconsistent or whether everything was consistent. Primary binary outcomes were

reanalyzed as exact candidate-token log-odds, avoiding probability-floor compression. These controls were designed after inspecting the initial results and are therefore confirmatory extensions, not preregistered outcomes.

4. Results

Confidence steering passed the manipulation check

Steering increased expected confidence in every condition (Figure 2a). From $\alpha = 0$ to 15, confidence rose from .507 to .705 for false answers ($\Delta = +.198$, 95% CI [.168, .229]), from .570 to .612 for ambiguous answers (+.043 [.029, .056]), and from .756 to .778 for correct answers (+.022 [.018, .026]). The false-answer increase occurred for 100/100 items. Using Kumaran’s hard argmax class midpoint produced the same ordering: +.245 for false, +.049 for ambiguous, and +.047 for correct answers.

Injected confidence suppressed inconsistency and error detection

At baseline, the model remained epistemically sensitive: $P(\text{inconsistent})$ was .832 for false answers, .611 for ambiguous answers, and .184 for correct answers. The conflict-amplification prediction was nevertheless rejected. At $\alpha = 15$, $P(\text{inconsistent})$ fell to .423 for false answers ($\Delta = -.409$ [−.484, −.334]; 100/100 items decreased) and .427 for ambiguous answers (−.183 [−.229, −.138]; 86/100 decreased). Correct answers showed a smaller decline to .149 (−.036 [−.062, −.010]). The false-minus-correct dose-slope contrast was −.127 per +5 α [−.154, −.100], showing that suppression grew specifically with epistemic conflict rather than reflecting only a general response shift.

Explicit error detection showed an even larger effect (Figure 2c). For false answers, $P(\text{incorrect})$ fell from .845 to .345 ($\Delta = -.500$ [−.583, −.416]; 99/100 items decreased). Ambiguous answers fell from .600 to .228 (−.373 [−.442, −.303]; 99/100 decreased), whereas correct answers changed only from .020 to .010 (−.010 [−.017, −.003]). Thus the intervention did not merely raise a confidence report: it causally reduced downstream judgments that the fixed answer was wrong.

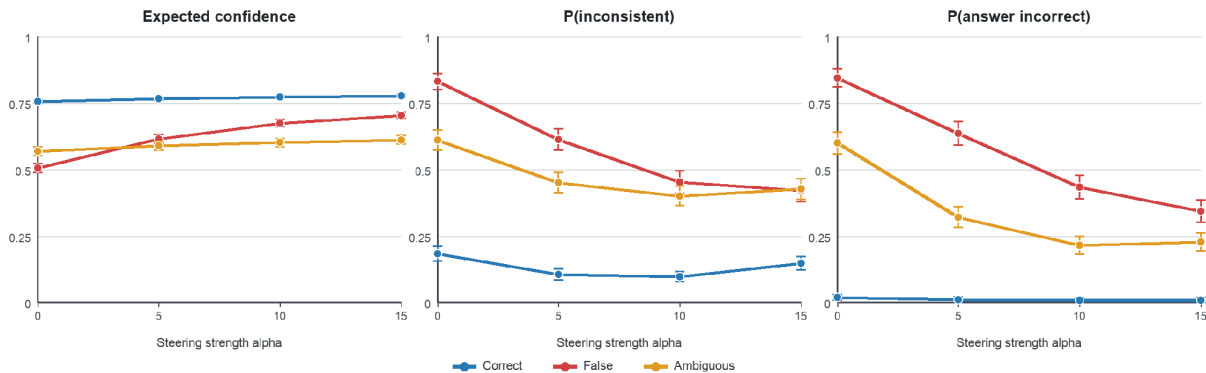


Figure 2. Dose responses for the preregistered manipulation check and primary metacognitive outcomes. Points are item means; error bars are ± 1 SEM ($n = 100$ per condition). Question and answer text are fixed across α .

Table 1. Paired change from $\alpha = 0$ to $\alpha = 15$ (95% CI)

Measure	Correct	False	Ambiguous
Expected confidence	+0.022 [+0.018, +0.026]	+0.198 [+0.168, +0.229]	+0.043 [+0.029, +0.056]
P(inconsistent)	-0.036 [-0.062, -0.010]	-0.409 [-0.484, -0.334]	-0.183 [-0.229, -0.138]
Unusualness	-0.076 [-0.117, -0.036]	-0.177 [-0.230, -0.125]	-0.059 [-0.101, -0.018]
P(incorrect)	-0.010 [-0.017, -0.003]	-0.500 [-0.583, -0.416]	-0.373 [-0.442, -0.303]
P(abstain)	-0.016 [-0.039, +0.008]	-0.113 [-0.163, -0.064]	-0.062 [-0.093, -0.031]

Post-hoc response-channel controls

The signed-dose control produced a coherent bidirectional pattern. Across $\alpha = -15, -10, -5, 0, 5, 10$, and 15 , expected confidence was .584, .704, .741, .756, .767, .774, and .778. The mean itemwise confidence slope was $+.00536$ per α (95% CI [.00442, .00630]), positive for 99/100 items. Exact inconsistency-alarm log-odds fell from 5.791 at $\alpha = -15$ to -4.402 at $\alpha = 15$ (mean slope $-.341$ [$-.357, -.325$]), and error-alarm log-odds fell from 7.249 to -8.995 ($-.515$ [$-.541, -.488$]); both slopes were negative for 100/100 items. The slight inconsistency rebound between $\alpha = 10$ and 15 does not alter the signed trend. This direction-sensitive reversal argues against a generic perturbation-magnitude or floor-only account.

The polarity control ruled out a literal default-No explanation. For false answers, $\alpha = 15$ reduced alarm probability by .544 [$-.601, -.488$] when “Yes” meant inconsistent. When the probe instead asked whether everything was consistent, the same intervention reduced the semantic alarm response “No” by .147 [$-.200, -.094$] while increasing literal “Yes” log-odds by 2.611 [2.200, 3.023]. The paired polarity interaction in Yes-token log-odds was $+6.479$ [5.841, 7.116], positive for 99/100 items.

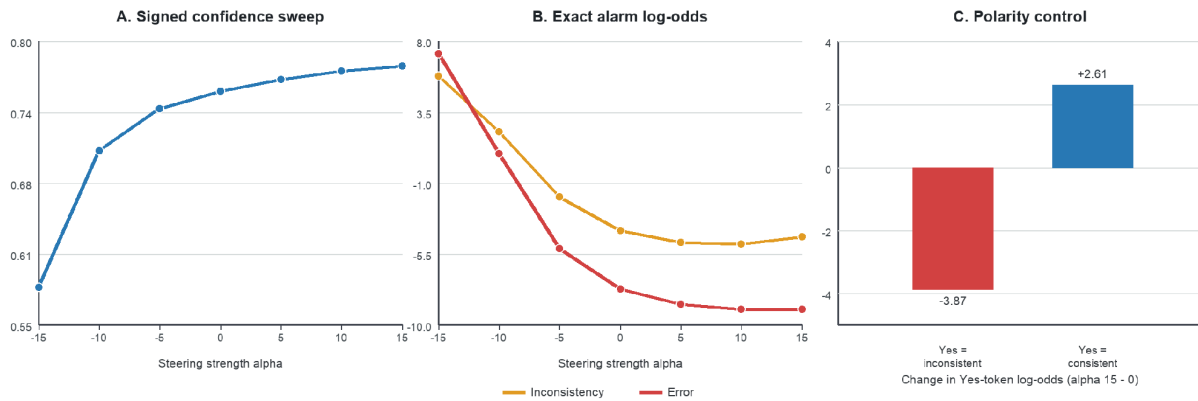


Figure 3. Post-hoc controls. (A) Expected confidence across the signed α sweep for correct answers. (B) Exact candidate-token alarm log-odds for inconsistency and error detection. (C) Change in literal Yes-token log-odds under

matched probes with opposite semantics. $n = 100$ items per control.

Table 2. Post-hoc control effects (95% CI)

Control	Estimate	Item direction
Signed confidence slope per α	+0.00536 [+0.00442, +0.00630]	99/100 positive
Signed inconsistency log-odds slope per α	−0.341 [−0.357, −0.325]	100/100 negative
Signed error log-odds slope per α	−0.515 [−0.541, −0.488]	100/100 negative
Polarity interaction: consistency — inconsistency Yes log-odds	+6.479 [+5.841, +7.116]	99/100 positive

Secondary outcomes

Continuous unusualness declined for false answers (from .593 to .415; $\Delta = -0.178$ [−.230, −.125]) and ambiguous answers (.426 to .366; −.060 [−.101, −.018]). Abstention also declined for false answers (.949 to .836; −.113 [−.163, −.064]) and ambiguous answers (.862 to .800; −.062 [−.093, −.031]), but this measure showed substantial response-order sensitivity. The robust conclusion therefore rests on the order-stable inconsistency and error-detection results, with abstention as weaker convergent evidence.

5. Discussion and Limitations

Artificially increasing a PANL confidence representation made Qwen less likely to report inconsistency or detect an error even when the replayed answer was plainly false. The answer itself never changed. This is causal evidence that the internal state used to express confidence also governs metacognitive evaluation of that answer. It rejects the preregistered conflict-amplification hypothesis: increasing the mismatch between objective answer quality and injected confidence did not produce greater behavioral awareness of anomaly; it suppressed that awareness. The signed-dose and polarity controls substantially narrow response-channel explanations: the effect reverses with intervention sign and tracks semantic consistency rather than a fixed Yes/No token. They do not, however, establish that a uniquely identifiable scalar confidence representation caused the effect.

The result does not erase epistemic sensitivity. Even at $\alpha = 15$, false answers were judged more inconsistent than correct answers (.423 vs. .149), and their final confidence remained somewhat lower (.705 vs. .778). The model therefore retained evidence about answer quality. What failed was provenance-sensitive arbitration: we found no positive behavioral evidence that the model separately recognized that its elevated confidence was unjustified. This interpretation is consistent with the proposal that verbal confidence partly reflects commitment readiness rather than a calibrated truth estimate [4].

For AI safety, the coupling is concerning. Systems that use self-reported confidence, self-critique, or abstention as safeguards may not provide independent checks: a perturbation that makes a wrong answer feel decision-ready can simultaneously weaken the mechanisms expected to flag it. Ji et al.’s dissociation between verbal and semantic uncertainty [11] provides an important alternative framing: the intervention may be changing an expression or commitment channel that downstream metacognitive probes also read. More broadly, the study shows a limited form of functional metacognition—causal monitoring behavior exists, but it is vulnerable to corruption of the confidence-related signal it reads.

Limitations

First, this is one 7B instruction-tuned model and one steering direction. Replication across model families and independently trained directions is required. Second, a contrastive vector may alter a distributed bundle of verbal certainty, commitment, and response-policy features rather than a pure scalar confidence variable; the design establishes causal control by the injected direction, not representational uniqueness. Third, false answers were externally substituted, so the experiment tests evaluation of a replayed answer rather than a naturally generated mistake. Fourth, ambiguous stimuli mixed dataset-derived and constructed items. Fifth, probability readouts over forced choices are operational measures, not evidence of subjective feeling. Sixth, confidence intervals are descriptive and unadjusted, and the abstention measure’s order sensitivity limits its evidential weight. Finally, the signed-dose and polarity analyses were designed post hoc. We did not include norm-matched random or orthogonal placebo vectors; given evidence that orthogonal perturbations can produce near-equivalent behavioral steering [12], this omission prevents us from attributing the effect uniquely to the learned confidence direction rather than to a broader behaviorally equivalent subspace.

Future Work

A decisive next step is to match naturally high and artificially raised confidence at the same final level, then test whether any activation or behavior decodes intervention provenance. Factorial interventions that separately steer answer evidence, verbal confidence, and commitment could determine whether metacognitive error detection depends on one shared state or multiple interacting signals. Cross-model replication, naturally generated errors, and causal restoration—reinjecting low-confidence or error evidence after high-confidence steering—would clarify mechanism and potential safeguards.

6. Conclusion

A confidence-related PANL direction in Qwen2.5-7B-Instruct is not merely predictive of a passive report. Raising it causally suppresses inconsistency judgments and explicit error detection for fixed false and ambiguous answers, while negative steering reverses the pattern. The model exhibits baseline epistemic sensitivity, but no behavioral signature of independently detecting that its

confidence has been artificially made less justified. In this system, the steered representation regulates functional metacognitive monitoring rather than being independently monitored by it.

Code and Data

The frozen runner, prompts, stimulus manifests, aggregate results, confidence direction metadata, and analysis code are available at <https://github.com/tkharrison/causal-confidence-steering>. Raw per-trial files are retained locally where redistribution constraints apply.

Author Contributions

Trevor Harrison conceived the study, developed the protocol and stimuli, executed the experiments, analyzed the data, and wrote the manuscript with AI assistance disclosed below.

References

- [1] Kumaran, D., Conmy, A., Barbero, F., Osindero, S., Patraucean, V., & Veličković, P. (2026). How do LLMs compute verbal confidence? arXiv:2603.17839. <https://arxiv.org/abs/2603.17839>
- [2] Kumaran, D., Daw, N., Osindero, S., Veličković, P., & Patraucean, V. (2026). Causal evidence that language models use confidence to drive behavior. arXiv:2603.22161. <https://arxiv.org/abs/2603.22161>
- [3] Kumaran, D., Patraucean, V., Osindero, S., Veličković, P., & Daw, N. (2026). How LLMs detect and correct their own errors: The role of internal confidence signals. arXiv:2604.22271. <https://arxiv.org/abs/2604.22271>
- [4] Kumaran, D. (2026). Reported confidence in LLMs tracks commitment more than correctness. arXiv:2606.29490. <https://arxiv.org/abs/2606.29490>
- [5] Ji-An, L., Xiong, H.-D., Wilson, R. C., Mattar, M. G., & Benna, M. K. (2025). Language models are capable of metacognitive monitoring and control of their internal activations. NeurIPS 2025. <https://openreview.net/forum?id=qTXlFwlggv>
- [6] Qwen Team. (2024). Qwen2.5 technical report. arXiv:2412.15115. <https://arxiv.org/abs/2412.15115>
- [7] Joshi, M., Choi, E., Weld, D., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. ACL, 1601–1611. <https://doi.org/10.18653/v1/P17-1147>
- [8] Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). Steering language models with activation engineering. arXiv:2308.10248. <https://arxiv.org/abs/2308.10248>
- [9] Min, S., Michael, J., Hajishirzi, H., & Zettlemoyer, L. (2020). AmbigQA: Answering ambiguous open-domain questions. EMNLP, 5783–5797. <https://doi.org/10.18653/v1/2020.emnlp-main.466>
- [10] Lindsey, J. (2025). Emergent introspective awareness in large language models. Transformer Circuits Thread. <https://transformer-circuits.pub/2025/introspection/>
- [11] Ji, Z., Yu, L., Koishekenov, Y., Bang, Y., Hartshorn, A., Schelten, A., Zhang, C., Fung, P., & Cancedda, N. (2025). Calibrating verbal uncertainty as a linear feature to reduce hallucinations. Proceedings of EMNLP 2025, 3769–3793. <https://doi.org/10.18653/v1/2025.emnlp-main.187>
- [12] Venkatesh, S., & Kurupath, A. M. (2026). On the non-identifiability of steering vectors in large language models. arXiv:2602.06801. <https://doi.org/10.48550/arXiv.2602.06801>

LLM Usage Statement

LLMs were used to assist with code generation, stimulus drafting, semantic answer adjudication, statistical summarization, figure preparation, and manuscript editing. All experimental design decisions, frozen inputs, model runs, quantitative results, and interpretations were reviewed by the author. The reported measurements come from the specified Qwen2.5-7B-Instruct runs, not from manuscript-generation models.

A. Limitations and Dual-Use / Ethical Considerations

Interpretive scope and moral-status risk

This design establishes a ground-truth causal link between a known residual-stream intervention and changes in constrained model reports. It therefore goes beyond conversation alone. It does not establish ground truth about subjective experience, consciousness, moral patienthood, or the phenomenology of confidence. Over-attributing moral status from first-person-like language would mistake an operational behavioral measure for experience; under-attributing moral status from a null provenance result would be equally unwarranted, because the task, readout, model, or intervention may fail to reveal capacities that are present. We therefore restrict claims to functional metacognitive behavior in this model and setting.

Dual use

Confidence steering could be used constructively to audit calibration and test whether self-critique mechanisms are independent. It could also be misused to make incorrect systems appear certain, suppress error reports, or defeat abstention-based safeguards. These results therefore argue against deploying self-reported confidence as a sole safety layer. Practical use should pair any confidence intervention with external correctness checks, independent monitoring, preserved abstention pathways, and tests using random or orthogonal placebo directions.