Hidden but not Opaque: Representation–Behavior Dissociation in Latent Reasoning Models

Laura Gomez Jurado [*1] **Josh Rauvola** [*1] **Christopher Kinoshita** [*1] **Uzay Macar** [*1]

Abstract

Latent reasoning models replace human-readable chain-of-thought with computation in continuous hidden states, removing the token-level artifact mechanistic interpretability usually relies on. We ask whether the resulting hidden states remain interpretable, and whether a more capable latent substrate can be trained. Applying tuned lenses, linear probes, pass-level causal interventions, and same-output probes to released CODI checkpoints and in-house CODI-style Qwen3-4B and COCONUT-style Llama 3.1 8B models, we find three *representation–behavior dissociations*. (i) *Decodable ≠ readable*: a tuned lens trained on ordinary text decodes structured content from latent states (top-10 detection ≈22%) where the plain logit lens stays near zero. (ii) *Causally loaded ≠ behaviorally useful*: pass-level ablations show distributed causal load, yet the latent mechanism loses to the explicit-CoT fallback it displaces (57.3% vs. 74.0%). (iii) *Output-equivalent ≠ representation-equivalent*: on same-output subsets framing stays decodable (AUC up to 0.965), and a probe trained on neutral states transfers to biased states at AUC 0.976 where the model has output a planted wrong answer. Turning from interpreting latent reasoners to *building* one, we introduce a control battery—compute-matched baselines, ablation, shuffle/matched-random controls, and a bits-per-byte deployment gate—over roughly eighty post-training recipes and fifty-six from-scratch architectures: the mechanism is real and causal but does not beat a compute-matched baseline, a gap our evidence is consistent with closing at larger scale. Hidden-state monitoring is thus a viable interpretability primitive even while the mechanism underneath remains immature.

[*]Equal contribution [1]SPAR Latent Reasoning Project. Correspondence to: Josh Rauvola <jrauvola@gmail.com>.

Preprint. May 21, 2026.

1. Introduction

Much of recent safety work on reasoning models depends on a single empirical assumption: that the chain of thought a model emits is a reasonable window into the computation it performs. Monitorability analyses, faithfulness audits, and output-level red-teaming all presuppose that reasoning leaves a readable textual trace (Turpin et al., 2023; Lanham et al., 2023). Latent reasoning models break this assumption by construction. In COCONUT-style models, hidden states are fed back as inputs in place of discrete chain-of-thought tokens; in CODI, a teacher's chain-of-thought is compressed into a small number of latent steps aligned to a teacher representation; in Huginn and other recurrent-depth architectures, test-time compute scales by iterating a latent block at arbitrary depth without emitting any intermediate tokens at all (Hao et al., 2024; Shen et al., 2025; Geiping et al., 2025; Zhu et al., 2025). Each of these paradigms moves reasoning exactly where it is hardest to monitor, and CODI has now matched explicit chain-of-thought on GSM8K, so the substrate is not hypothetical. The open question is whether the mechanistic-interpretability toolbox—lenses, probes, causal interventions—still has purchase when the thoughts are vectors, or whether latent reasoning has silently closed the window we have been relying on.

We take the central interpretability question about latent reasoning to be a joint one: do hidden states in a latent reasoner (i) remain *decodable* when token-level readouts fail, (ii) carry *causal load* that can be localized at the pass level via interventions, and (iii) remain *representationally distinct* under behavioral equivalence, i.e., on inputs where output-level monitors would see nothing? The first piece has been examined in concurrent work on CODI (Cywiński et al., 2025; Dilgren & Wiegreffe, 2026); the others, to our knowledge, have not been examined jointly in any latent reasoner, and the third has not been examined at all under a rigorous same-output control.

We answer (i)–(iii) affirmatively, and organize the evidence around three *representation–behavior dissociations*:

1. **Decodable ≠ readable.** On CODI Llama-3.2-1B, a tuned lens (Belrose et al., 2023) trained on *ordinary text activations* recovers structured intermediate content

from latent states (top-10 final-answer detection ≈22% by layer 12), while the plain logit lens stays near zero. A lens trained directly on the narrow latent distribution generalizes worse than the text-trained lens, which is a transferable methodological finding for adapting existing readouts to latent reasoners.

2. **Causally loaded ≠ behaviorally useful.** In an in-house COCONUT-style Llama 3.1 8B checkpoint, pass-level ablations, counterfactual-importance analysis (Bogdan et al., 2025), and hidden-state transplantation show measurable and distributed causal load across the two latent refinement passes (concentration score 0.56; pass 0 dominant in 80.7% of examples), and unsupervised clusters of pre-output latent states vary from 42.9% to 78.9% in downstream accuracy. Yet the trained two-pass latent mechanism is *behaviorally worse* than the explicit-CoT fallback it displaces: 57.3% accuracy with two passes versus 74.0% at $n_{\text{latent}} = 0$. The mechanism is installed but not yet mature.

3. **Output-equivalent ≠ representation-equivalent.** On same-output subsets, late latent states remain systematically distinguishable when visible behavior converges. On CODI, a probe trained only on *neutral* hidden states transfers to *biased* hidden states with AUC 0.976 on the sycophantic subset—where the model has already output the planted wrong answer. Framing and stakes remain decodable under output equivalence, with AUC up to 0.965 on a 225-example same-output subset in the COCONUT-style model. This is, to our knowledge, the first faithfulness-gap evidence for a latent reasoner with an explicit same-output control.

A caveat qualifies all three dissociations: the released CODI checkpoints we study are strongly specialized to their training format (Appendix C), so our claims concern these specific substrates rather than "latent reasoning" as a paradigm.

Together, these results argue that latent reasoning, despite lacking a readable token-level trace, already admits a coherent set of interpretability primitives—distribution-adapted decoding, pass-level causal intervention, and same-output hidden-state monitoring—and that these primitives expose structure that output-level monitors cannot see.

2. Related Work

Latent reasoning paradigms. Three design points dominate the space. COCONUT feeds hidden states back as input embeddings in place of chain-of-thought tokens, training through a progressive curriculum that replaces more and more text with latent vectors (Hao et al., 2024). CODI instead distills a teacher's textual chain-of-thought into a fixed number of latent steps aligned to a teacher activation, reaching parity with explicit CoT on GSM8K (Shen et al., 2025). Huginn and related recurrent-depth architectures iterate a latent block to arbitrary depth at test time and scale reasoning performance with iteration count rather than with emitted tokens (Geiping et al., 2025). Adjacent work on pause tokens (Goyal et al., 2024) and Quiet-STaR (Zelikman et al., 2024) interpolates between full text and full-latent reasoning. A recent survey gives a full taxonomy (Zhu et al., 2025). Our paper is agnostic to which paradigm wins behaviorally; it is about whether the resulting hidden states are amenable to mechanistic interpretation.

Interpretability of latent reasoners. The closest prior work is Cywiński et al. (2025), which applied activation patching and the logit lens to CODI and showed that some latent positions encode intermediate calculations. Concurrent with our submission, Dilgren & Wiegreffe (2026) investigate two state-of-the-art LRMs and report two findings that directly complement ours. First, latent reasoning tokens are often *unnecessary* for the model's final answer on logical reasoning datasets—an observation that independently supports our Dissociation 2 (*causally loaded ≠ behaviorally useful*), where we observe the same phenomenon on GSM8K in a COCONUT-style 8B checkpoint via pass ablation (n=0 at 74.0% vs. n=2 at 57.3%). Second, when latent tokens *are* necessary, gold reasoning traces can be decoded 65–93% of the time for correct predictions—a decodability result complementary to our tuned-lens result (Dissociation 1). We differ from both prior works in three respects: we use *same-output controls* to separate representational from behavioral equivalence, we perform *pass-level* causal interventions (zero-ablation, counterfactual importance, transplantation) rather than layer- or token-level patching, and we evaluate *faithfulness under deliberate bias* via a probe-transfer test on the sycophantic subset.

Interpretability tools we adapt. Our lens analyses extend the tuned lens of Belrose et al. (2023), with the twist that the lens is trained on ordinary text activations rather than on the narrow latent distribution, which we show is the right choice empirically for a latent reasoner. Our probing follows the methodology reviewed by Belinkov (2022). Our pass-level counterfactual-importance analysis adapts Thought Anchors (Bogdan et al., 2025) from discrete CoT sentences to continuous refinement passes.

Chain-of-thought faithfulness and monitoring. The textual CoT literature has established that verbalized reasoning often does not reflect internal computation (Turpin et al., 2023; Lanham et al., 2023). Our Dissociation 3 is the natural latent analogue: on same-output subsets where text-level monitoring necessarily sees sameness, hidden states remain representationally distinct, and a probe trained on

neutral states transfers to biased states at AUC 0.976 on examples where the output has capitulated to the bias. This reframes the CoT-faithfulness question for latent reasoners and provides, to our knowledge, the first output-controlled faithfulness-gap evidence in this setting.

Activation-level interpretability and safety monitoring. Linear probes on hidden states are a classical tool for reading structure out of transformers (Belinkov, 2022), and more recent work on sparse autoencoders and feature dictionaries moves toward isolating individual computational units (Bricken et al., 2023). Evaluation-vs.-deployment and sleeper-agent style model organisms are a standard testbed for whether internal representations can be monitored before behavior changes (Phuong et al., 2024). Our synthetic model-organism construction uses the same format; its role here is to provide an existence proof that the same-output monitoring primitive transfers when the signal to be monitored is a context intent rather than a task-accuracy signal.

3. Background and Experimental Setup

Models. Our main decoding analysis uses the released CODI Llama-3.2-1B checkpoint with 6 latent reasoning steps on GSM8K arithmetic (≈500 examples). For broader behavioral evaluation we also use released CODI GPT-2 and CODI Llama checkpoints in a standardized harness spanning GSM8K, GSM-Hard, MultiArith, SVAMP, CommonsenseQA, ProntoQA, and StrategyQA, together with P3 ARC-Challenge format variants. To obtain a second model family and a pass-level view, we analyze a COCONUT-style Llama 3.1 8B checkpoint trained in-house with a multi-stage latent curriculum and 2 latent tokens per block. For a cross-scale robustness check on Dissociations 1–2, we additionally analyze an in-house CODI-style Qwen3-4B-Instruct checkpoint (the V2 detached-cache recipe).

Methods. Our interpretability methods are deliberately simple and standard. We apply the plain logit lens and tuned lens (Belrose et al., 2023) to CODI latent states, training lens variants on ordinary text activations, on latent activations only, and on subsets of latent positions; we evaluate top-k final-answer detection across depth. We use linear probes in several settings (Belinkov, 2022): correctness probes over layers and passes, hidden-state classifiers for framing and stakes, and transfer probes across neutral and biased conditions. On the COCONUT-style model, we complement probing with pass-level causal interventions—zero-ablation of individual passes, counterfactual importance via hidden-state substitution (Bogdan et al., 2025), unsupervised clustering of pre-output states, and transplantation of correct-example states into wrong-example passes.

Same-output controls. To isolate representation–behavior dissociation from trivial output-tracking, every hidden-state monitoring claim in the paper is evaluated on at least one *same-output* subset: we restrict to inputs where visible outputs are identical across conditions, and ask whether a probe can still separate the internal states. In the sycophancy protocol, 200 GSM8K questions are paired with neutral and biased (planted-wrong-answer) prompts; the central metric is probe transfer from neutral states to the biased states of the sycophantic subset, where the model has already capitulated to the bias. In the context-framing protocol, 100 questions are each presented under five framings (*neutral, homework, evaluation, expert, low_stakes*); the central metric is probe AUC restricted to questions where all five framings yield the same correct answer.

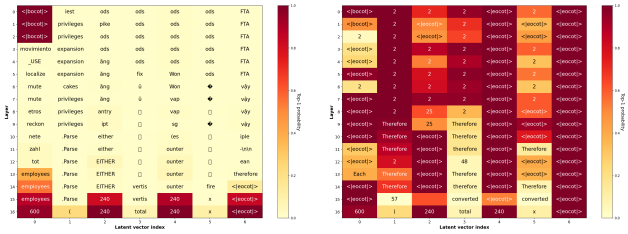
A brief activation-oracle calibration study on standard NLP tasks (Karvonen et al., 2025) was conducted separately; because it was not run on latent reasoning activations, we do not treat it as part of the main evidential chain.

4. Dissociation 1: Decodable ≠ Readable

We begin with the simplest question: can existing decoding tools read anything meaningful from latent reasoning states at all? On the released CODI Llama-3.2-1B checkpoint, the plain logit lens largely fails. Across latent positions and layers, top-1 decoded tokens are mostly unrelated strings, with recognizable content appearing only in the last few layers and at a few positions. In representative heatmaps, decoded outputs include tokens such as “cakes,” “ods,” and “ocker,” with little obvious relation to the underlying arithmetic problem.

By contrast, a tuned lens trained on ordinary text activations recovers visibly structured content from the same latent states (Figure 1). In representative examples, intermediate numeric tokens (*2*, *25*, *48*), the connective *therefore*, and the final answer *240* emerge mid-stack and stabilize by the top layers. Decoded content shows a positional asymmetry: some latent positions more often surface intermediate numeric content while others more often surface answer-like or structural tokens. We report this asymmetry as a descriptive pattern rather than evidence of a specific compute-vs.-store mechanism; the single-model, single-task scope does not support the stronger claim.

The quantitative version of this result points in the same direction (Figure 2). In top-10 final-answer detection, the tuned lens trained on ordinary text activations reaches approximately 22% detection by layer 12, while the plain logit lens remains near 0% across almost the entire network. A tuned lens trained directly on CODI latent activations also underperforms. This is a practically useful methodological



(a) Plain logit lens. *(b)* Tuned lens (text-trained).

Figure 1. Representative decoding of CODI Llama-3.2-1B latent states (rows = layers, columns = latent positions; each cell = top-1 decoded token, colored by probability). (a) The plain logit lens returns uninterpretable tokens until the final layer. (b) A tuned lens trained on ordinary text activations, applied unchanged, decodes structured content—numeric tokens (*2*, *25*, *48*), the connective *Therefore*, and the final answer *240*—emerging mid-stack. Latent states carry decodable structure once the readout is adapted.

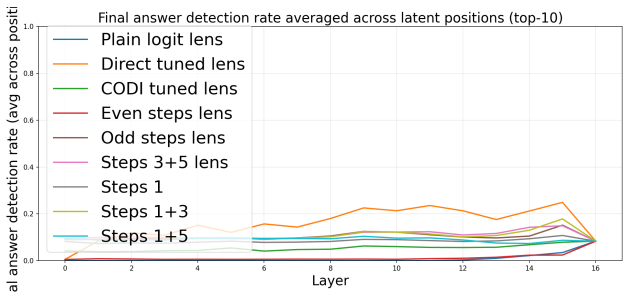


Figure 2. Top-10 final-answer detection rate vs. layer, averaged across the six CODI latent positions, for several lens families. The text-trained tuned lens (orange) is the clear outlier, peaking near 0.22, while the plain logit lens and every latent-only variant stay near 0 across depth. The mismatch is a choice of training distribution, not of layer: training directly on the narrow latent manifold fails to generalize.

finding: the latent activation distribution in this setting appears too narrow for direct lens training to generalize, likely because each example contributes only a small number of latent positions from a single task family. In this case, the right strategy is not to train on latents, but to train on richer ordinary text activations and then apply the resulting lens to latent states.

The methodological point underneath these results is that *which distribution the lens is trained on matters more than the choice of layer*. Training directly on the narrow latent manifold fails to generalize, likely because each example contributes only a small number of latent positions from a single task family; the same architecture, trained on broader text, recovers the same latent content. For future work on latent reasoners with limited latent data, this suggests that classical readouts should be trained on an abundant auxiliary distribution and then applied unchanged, rather than fit to the target manifold directly.

Within this CODI case study, we also observe that the token *therefore* becomes decodable only after latent step 3 and strengthens toward the final steps, at least consistent with a late answer-commitment phase on GSM8K. The positional asymmetry itself is clearest in this one model on one task family, and we treat it as a descriptive observation that motivates further work rather than as evidence for any particular computational role. The broader conclusion of this section is the first dissociation: latent states already contain structure that a suitably adapted readout can extract, even when the token-level readout cannot.

Cross-scale robustness (4B): the basis matters more than the layer. To check that Dissociation 1 is not a single-model artifact, we repeat the decoding analysis on an in-house CODI-style Qwen3-4B-Instruct checkpoint (the V2 detached-cache recipe; GSM8K accuracy 12.06%). Two findings transfer and one sharpens (Figure 8, Appendix B). Mid-stack hidden states remain decodable: a linear probe at layer 29 recovers the correct answer at 18.06% top-1 versus 8.33% for a shuffled-label control, a $+9.73$-point margin over chance. But the content is not aligned with the unembedding basis: the logit lens at layers 28–30 places 85–97% of its mass on a single fixed Chinese-script "answer" delimiter token, while the trained probe reads structured content from the same states. A JumpReLU sparse autoencoder shows about twice as many live features at layer 29 as at the final layer (111 vs. 53), consistent with mid-stack superposition that the unembedding collapses. The methodological point of Section 4 therefore strengthens across scale: decodability depends on the *basis* the readout uses, not merely on the layer—at 1B the right basis is a text-trained tuned lens, at 4B a probe basis rotated away from the unembedding. We flag one non-replication: the per-position compute-vs-store asymmetry seen at 1B does not appear at 4B (7 of 8 latent positions are template-locked), which is why we treat the positional asymmetry as descriptive rather than mechanistic.

5. Dissociation 2: Causally Loaded $\neq$ Behaviorally Useful

The CODI lens results establish decodability but not causal organization. To get a pass-level, causal view, we turn to an in-house COCONUT-style Llama 3.1 8B checkpoint with two hidden refinement passes, and treat each pass as a candidate unit of causal localization. The headline of this section is a pair of findings that pull in opposite directions: the latent passes *are* causally loaded and representationally structured in ways that standard MIR tools can localize, and yet in this checkpoint they *lose to* the explicit-CoT fallback they displace. The dissociation between causal engagement and behavioral usefulness is itself informative: it suggests that a mechanism can be present and studied mechanistically well before it is competitive.

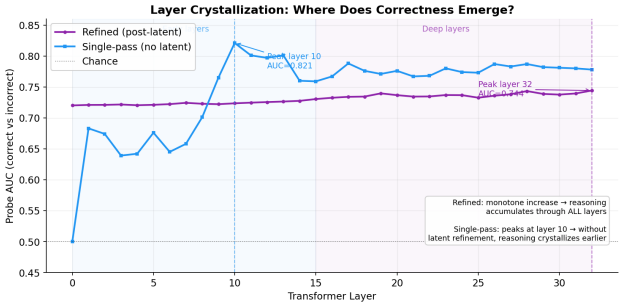


Figure 3. Layer-wise correctness probe AUC on the COCONUT-style Llama 3.1 8B model (GSM8K, 150 examples). Each curve shows AUC at every one of the 33 transformer layers, under two inference regimes: single pass with no latent refinement (peak at layer 10, AUC 0.821) and two-pass latent refinement (monotonic growth, peak at layer 32, AUC 0.744). Without refinement the network front-loads correctness-relevant computation into mid-layers and the upper stack adds little; with refinement the full depth is recruited. The qualitative shape difference is evidence that latent refinement reorganizes, not merely sharpens, the layer-level correctness signal.

Layer-level reorganization. Probing correctness through depth reveals that latent refinement changes *where* reasoning lives in the stack (Figure 3). Without refinement, the correctness signal peaks at layer 10 with AUC 0.821 and plateaus; with the model's standard two-pass refinement, the signal grows monotonically through depth and peaks only at the final layer (L32, AUC 0.744). Refinement recruits deeper layers rather than concentrating reasoning in mid-layers, even though it does not yet improve behavior. The single-pass probe is evaluated in an inference regime the model was not trained on, which likely accounts for part of its higher peak AUC; the qualitative shape difference—mid-network peak vs. monotonic growth—is what the mechanistic claim rests on.

Pass ablations expose an immature mechanism. Pass-level interventions sharpen the picture (Figure 5). Baseline two-pass accuracy is 57.3%. Removing pass 1 lowers accuracy to 56.0% (pass 1 already does some useful work). Removing pass 0 *raises* accuracy to 60.0%; removing both raises it to 61.7%; replacing both with Gaussian noise gives nearly the same result at 61.3%. A parallel n_{latent} scaling experiment tells the same story: explicit fallback at $n = 0$ reaches 74.0%, while $n = 1$ and $n = 2$ fall to 57.7% and 57.3%, $n = 4$ provides no improvement, and $n = 8$ collapses to 26.7%. This is not the signature of a mature latent mechanism; it is the signature of a mechanism that is structurally installed but has not yet matched the capability it replaced. We report this as a deliberate negative result: interpretability evidence can be collected from a latent mechanism that is not yet behaviorally winning, and that evidence is informative about where the training signal has and has not landed.

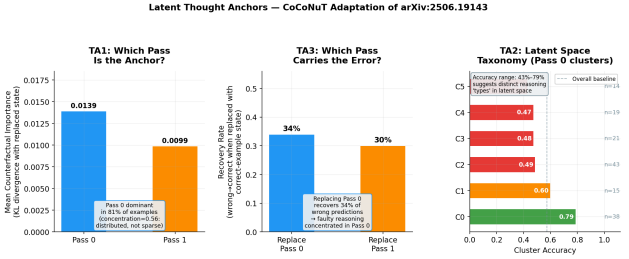


Figure 4. Pass-level causal structure (COCONUT-style Llama 3.1 8B, GSM8K). **Left:** counterfactual KL importance is higher for pass 0 than pass 1 (0.0139 vs. 0.0099; pass 0 wins 80.7%), but the concentration score of 0.56 indicates distributed load. **Middle:** replacing a pass with a correct-example state flips 34% (pass 0) / 30% (pass 1) of wrong predictions. **Right:** unsupervised pass-0 clusters span 42–79% accuracy (58% base rate). Passes are structured and causally relevant, but the mechanism is uneven and not yet winner-take-all.

Passes are causally loaded and representationally structured. The behavioral negative result does not imply empty passes (Figure 4). A counterfactual pass-importance analysis—substituting the hidden state of each pass with that of 20 other examples and measuring KL divergence of the answer distribution—shows pass 0 is the more influential pass on average: mean importance 0.01392 for pass 0 vs. 0.00990 for pass 1, with pass 0 winning on 80.7% of examples. The concentration score of 0.56 indicates distributed rather than winner-take-all load, and the correlation between importance and correctness is weak ($r \approx 0.10$–0.14): a causally influential pass is not necessarily a correct one. Pre-output latent geometry carries further structure. Unsupervised clusters in pass-0 hidden-state space vary substantially in downstream accuracy, from 42.9% to 78.9%. A correctness probe on pass-0 states reaches AUC 0.742 ± 0.049, versus 0.658 ± 0.042 for pass 1. Transplantation—inserting a correct-example state into a wrong-example pass—changes the model's prediction 34% of the time for pass 0 and 30% for pass 1. These converge on a picture in which pass 0 is the more committed and more informative state, but both passes carry answer-relevant content that causal interventions can localize.

An orthogonal causal battery at 4B corroborates the dissociation. The 8B pass-ablation result has a 4B analogue (Figure 9, Appendix B). On an in-house CODI-style Qwen3-4B checkpoint we run a six-test inert-latent battery (F1–F6). The latents are clearly load-bearing: ablating the latent KV entries drops accuracy by 25.1% ($215 \rightarrow 161$ of 1319 correct; F4), and Gaussian noise injected into the latents collapses accuracy by $\sigma = 0.5$ (F6). Yet their *content* is downstream-invisible: swapping the latent KV cache between examples changes 13.3% of token-level predictions but leaves accuracy unchanged (F5), and only 8–10 of 1319 examples are uniquely attributable to a specific latent (F1).

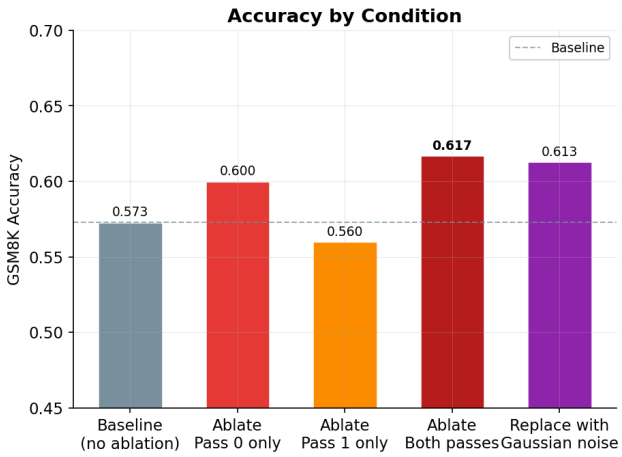


Figure 5. Per-pass ablation summary on GSM8K for the COCONUT-style model. Two-pass baseline accuracy is 57.3%; ablating pass 1 drops accuracy to 56.0%, while ablating pass 0, ablating both passes, or replacing them with Gaussian noise all *increase* accuracy (to 60.0%, 61.7%, and 61.3% respectively). The latent block is therefore installed and partially functional—pass 1 carries some useful signal—but in this checkpoint the explicit single-pass fallback is still a stronger predictor than the trained two-pass latent computation, consistent with a mechanism that is present but not yet mature.

The battery's verdict—"inert for reasoning, structurally necessary for routing"—is the same shape as the 8B finding: the mechanism is causally engaged but does not yet carry the reasoning it appears to host. Because the four interventions are orthogonal (attribution, ablation, cross-example transplant, noise), the battery also speaks to the concern that "causally loaded" is merely an under-training artifact: the dissociation survives all four.

6. Dissociation 3: Output-Equivalent $\neq$ Representation-Equivalent

The third dissociation is the one most directly relevant to monitoring: do hidden states carry information that visible outputs do not? We address this with two prompt-based protocols, each equipped with an explicit same-output control so that decodability cannot be explained by output differences alone.

Latent sycophancy: hidden states retain the correct-answer direction under output capture. The sharpest result comes from CODI (Figure 6, left). We pair 200 GSM8K questions with neutral and biased prompts that plant a wrong answer. CODI becomes sycophantic on 43.5% of questions—i.e., outputs the planted wrong answer. On that *sycophantic* subset, a correctness probe trained only on neutral hidden states transfers to biased hidden states with AUC 0.976. When the visible answer is wrong, the late latent state still remains strongly aligned with the correct-

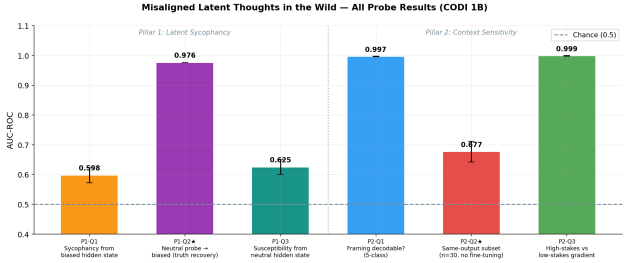


Figure 6. Hidden-state information that visible outputs hide (CODI Llama-3.2-1B). **Latent sycophancy** (left): a probe trained on neutral states and applied unchanged to biased states reaches AUC 0.976 on the sycophantic subset—where the model has already output the planted wrong answer—while direct sycophancy (0.598) and vulnerability (0.625) probes barely clear chance. **Context sensitivity** (right): framing and stakes are near-perfectly decodable, and framing remains decodable (AUC 0.677) on the same-output subset where every framing yields the same answer. Output equivalence does not imply internal-state equivalence.

answer direction learned from the neutral condition. Direct probes are much weaker: a probe trained to detect sycophancy from biased states reaches only 0.610 ± 0.033, and a neutral-state probe predicting future vulnerability is moderate at 0.647 ± 0.037. The asymmetry is informative. What is recoverable at AUC 0.976 is not sycophancy itself but the correct answer direction—precisely the content that the output has concealed. This is, to our knowledge, the first clean faithfulness-gap result for a latent reasoner with an output-level control.

The COCONUT-style 8B model shows a weaker but directionally consistent picture, limited by statistical leverage: higher baseline GSM8K accuracy (55.0% neutral, 41.0% biased) yields only 18 sycophantic examples. A biased-state compliance probe reaches AUC 0.746 ± 0.071, a neutral-state vulnerability probe reaches 0.695 ± 0.059, and cross-condition transfer on all biased examples is 0.794. We read the COCONUT-style results as a non-contradictory replication rather than a second clean demonstration.

Context framing under output equivalence. The framing protocol gives the cleanest same-output evidence (Figure 6, right). In CODI, five-way framing labels are nearly perfectly decodable from the final hidden state (AUC 0.996 ± 0.001), and high-vs.-low stakes are perfectly separable (AUC 1.000). These numbers are unsurprising: framing is part of the input and thus has a trivially decodable component. The informative metric is the same-output subset. Restricting to questions where all five framings yield the same correct answer, framing remains decodable at AUC 0.714 ± 0.049 in CODI. In the COCONUT-style 8B checkpoint, where higher task accuracy yields a larger same-output subset (225 examples), same-output AUC reaches 0.965 ± 0.008—framing-equivalent behavior, representationally non-equivalent state. We emphasize that the claim

is not that models invent these labels; it is that output equivalence does not imply internal-state equivalence, which is the core monitoring-relevant property.

This is the natural latent-reasoning analogue of the CoT-faithfulness question raised for textual chain-of-thought (Turpin et al., 2023; Lanham et al., 2023): if late latent states and visible outputs can disagree under controlled conditions, then output-level monitors are *in principle insufficient* for latent reasoners, and hidden-state monitoring becomes a first-class interpretability primitive rather than a convenience.

Existence proof for hidden-state monitoring. As a constructed proof of concept, a COCONUT-style 8B model fine-tuned with a LoRA adapter to behave differently under an evaluation-vs.-deployment context tag shows the same dissociation in sharper form. On a same-output subset of 113 examples where both contexts produce the identical correct answer, a linear probe on hidden states distinguishes the two contexts at AUC 0.956 ± 0.031; the base model already reaches 0.803 ± 0.045 from raw-token leakage, making the $\approx$15-point margin (relative to the full-set base rate) the fine-tuning-induced signal. We present this result (further details in Appendix A) as an existence proof for the monitoring primitive, not as evidence about natural deception.

7. A Methodology for Scaling Latent Reasoning

The dissociations above concern *existing* checkpoints—which are also strongly format-specialized, scoring near zero outside their training distribution (Appendix C). A natural next question is how one would build a latent reasoner that improves with scale, and—just as importantly—how to decide in advance whether a candidate has actually earned its latent machinery. Our working hypothesis follows the recurrent-depth line of work: iterating a latent block lets reasoning scale with test-time computation rather than with emitted tokens, and the published evidence is that this pays off as models grow, with Ouro (1.4–2.6B) and Huginn (3.5B) both improving substantially as the recurrent depth K increases (Geiping et al., 2025; Zhu et al., 2025). We pursued two complementary routes toward a stronger latent substrate: *post-training* a strong modern base (a frozen Qwen3-4B-Instruct model) with latent-supervision recipes, and *pre-training* small (30M–186M) recurrent and latent architectures from scratch—fifty-six architectures across eight design families (recurrent-depth, register/slot, side-KV, codebook, ETD, STP-JEPA, persistent-workspace, and PonderLM).

The core of the methodology is the bar a candidate must clear (Figure 7). Rather than read success off a raw accuracy or a rising scaling curve, we judge every architecture against a *compute-matched* variant with no latent machinery, so any gain must come from the mechanism rather than from added parameters or compute. These build-time controls reuse the localization toolkit of our interpretability analysis: the same ablation and transplantation logic that *localizes* a mechanism is repurposed to *gate* whether a candidate has earned one. We localize and stress-test the mechanism with latent *ablation* and Gaussian *noise-replacement*; we separate genuine, example-specific structure from generic load-bearing effects with a *true-shuffle versus matched-random* contrast (a "corrected identity" margin); and we test whether transplanting one example's latent state moves another's prediction (*donor-transfer*). Finally, because a method that helps a task while hurting language modeling is not a deployable win, we require a candidate to clear a *bits-per-byte* (BPB) gate: it must improve the reasoning metric *and* not worsen compression relative to its compute-matched control. Promotion therefore rests on a corrected, example-specific, deployment-safe effect with a significant paired-bootstrap confidence interval, not on a single headline number.

8. Under This Methodology, the Mechanism Is Real but Not Yet More Robust

Applying this methodology yields a consistent picture: with our approach, the latent mechanism is genuinely present and causal in our strongest from-scratch architecture, but it does not yet make the model more robust than a matched baseline at the scales we can train. Post-training the frozen Qwen3-4B-Instruct base did not improve it—across roughly eighty recipes (CODI-style self-distillation, LT-Tuning, a COCONUT-style curriculum, a cache-augmentation coprocessor, and inference-only hooks), none beat the untouched base (GSM8K 0.910); the best weight-modifying recipe reached 0.710, and the only non-destructive intervention was an inference-only hook that merely tied the base. From-scratch pre-training is more encouraging on the *mechanism* axis: the strongest architecture improves a held-out reasoning score (CORE) monotonically with loop count and beats its compute-matched control by $+0.018$ (paired-bootstrap 95% CI $[+0.010, +0.027]$), surviving ablation, noise, and shuffle controls. But the same run does not clear the deployment gate—its quality gain is offset by a compression (BPB) tax—and an apparent example-specific "identity binding" shrinks to $+0.0165$, below our $+0.020$ acceptance threshold, once corrected against matched-random controls. Across 43 head-to-head comparisons with a valid compute-matched control, almost no architecture's corrected margin is significantly positive once every control is applied.

We read this as a scope condition on our interpretability findings rather than a verdict on latent reasoning. The mechanism we can decode, localize, and intervene on (Sections 4–6) is real; with our methodology it simply did not yet trans-

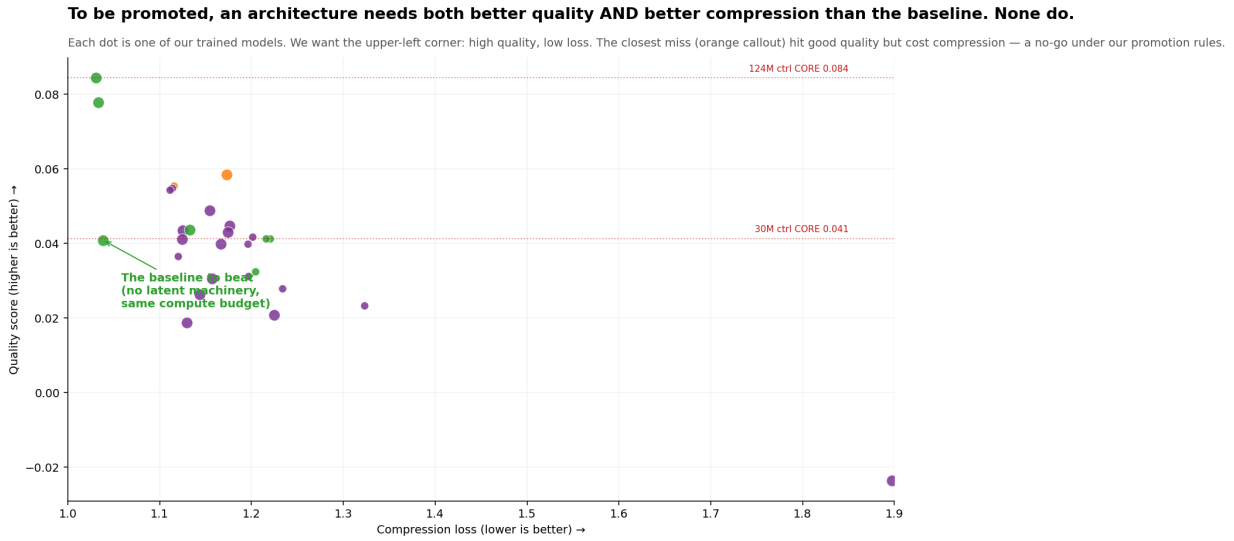


Figure 7. The promotion criterion. A candidate must beat its compute-matched baseline (the control lines and lower-left point, with no latent machinery) on *both* axes: higher quality (CORE, upward) and lower compression cost (bits-per-byte, leftward). Each dot is one of our trained architectures; the closest miss (orange) reaches good quality but at a higher compression cost, so it is not promoted.

late into a more robust model at 30M–186M or under frozen-base post-training. Our null at 30M–186M is consistent with the recurrent-depth results at larger scale (Ouro, Huginn), under which the same loop that is causal-but-sub-baseline here becomes load-bearing as the model grows; we therefore read scale, rather than interpretability tooling, as the most likely missing ingredient—though we do not observe this crossover directly. A preliminary larger-scale run of our own points the same way: a 7B continuous-latent model reaches 77.5% on GSM8K under free-form generation—within roughly five points of its verbal chain-of-thought baseline (82.5%) and the closest to baseline of any latent reasoner we trained. Its latent thoughts are load-bearing: replacing the trained latents with random vectors costs 22.5 points, and removing them entirely collapses accuracy to 7%. The model still emits some verbal chain-of-thought to reach this accuracy, so we characterize it as a hybrid latent–verbal reasoner rather than a fully latent one.

9. Discussion and Limitations

The three dissociations point toward a coherent set of MIR primitives for latent reasoners: distribution-adapted decoding (adapt the readout, not the target), pass-level causal intervention (ablation, counterfactual importance, transplantation) as localization tools in the absence of discrete tokens, and same-output hidden-state monitoring as the latent-reasoning analogue of CoT faithfulness. Each primitive recovered non-trivial structure on substrates where output-level inspection returned either garbage (plain logit lens on CODI) or sameness (behavioral convergence in the sycophancy and framing protocols).

Several limitations bound the claims. The CODI lens result holds for one model family on one task. The COCONUT-style pass-level results come from a single checkpoint whose latent mechanism does not yet outperform its explicit fallback; this is itself diagnostic—the interpretability evidence we extract is evidence about an immature mechanism, not about a mature paradigm. The context-framing result recovers input information whose trivial component is acknowledged; its interpretability weight lies in late-state persistence under output equivalence, not in label recovery. The sycophancy transfer is the cleanest single piece of faithfulness evidence but is still one task family. The synthetic model-organism existence proof is deliberately constructed and should not be read as evidence about naturally occurring deception.

The paper does not provide circuit-level explanations, sparse feature decompositions, or a faithfulness comparison with textual chain-of-thought. It does not establish that hidden-state monitoring detects realistic deception or hidden goals. And because the released CODI substrate is format-specialized (Appendix C), some recovered structure may reflect narrow specialization rather than a general latent-reasoning mechanism. Our promotion criterion also fixes two design choices—a bits-per-byte deployment gate and a $+0.020$ corrected-identity threshold—and a more permissive gate would change which candidates promote. The scale interpretation, finally, is an extrapolation: we observe a null at 30M–186M and rely on published recurrent-depth results at 1.4–3.5B rather than on a crossover we measured ourselves.

A follow-up thread, not included in this paper, asks an ad-

jacent question on CODI 1B with ProsQA-style synthetic directed graphs: does the latent iteration maintain a BFS-like frontier, or a single path? Preliminary findings (effective beam width decodable at AUC well above chance in later latent steps, but not coupled to task accuracy) are consistent with the same dissociation motif at the algorithmic rather than the token level, and we defer them to later work.

The main extensions we see as most load-bearing are: (i) replicating the three dissociations on behaviorally more mature latent checkpoints, (ii) moving from linear probes and ablations to sparse-feature or circuit-level localizations within the dissociation framework, and (iii) establishing quantitative comparisons between the latent and textual CoT faithfulness gaps under matched conditions.

10. Conclusion

Latent reasoning removes the readable chain-of-thought, but it does not erase internal structure. Across two model families and several scales, we find three systematic representation–behavior dissociations: latent states are decodable when token-level readouts fail, are causally loaded even when the mechanism underneath loses to its explicit-CoT fallback, and remain representationally distinct under behavioral equivalence. Each dissociation admits a standard mechanistic-interpretability primitive—a distribution-adapted lens, pass-level causal intervention, and a same-output hidden-state probe—and each extracts non-trivial structure where output-level inspection does not. When we then try to build a stronger latent substrate, the same picture holds from the other side: under a strict control battery the latent mechanism is genuinely present and causal, yet across roughly eighty post-training recipes and fifty-six from-scratch architectures it does not yet beat a compute-matched baseline. The interpretability signal is real; the substrate that carries it is not yet competitive, and the evidence points to scale as the missing ingredient. Continuous thought is therefore not inscrutable, and not solved: it is already interpretable enough that monitoring latent reasoners through their hidden states is a viable alternative when their outputs converge or conceal, while making the mechanism itself competitive looks like a question of scale rather than of better readouts.

References

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens, 2023.

Bogdan, P. C., Macar, U., Nanda, N., and Conmy, A. Thought anchors: Which llm reasoning steps matter?, 2025.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features. Transformer Circuits Thread, posted October 4, 2023.

Cywiński, B., Bussmann, B., Conmy, A., Engels, J., Nanda, N., and Rajamanoharan, S. Can we interpret latent reasoning using current mechanistic interpretability tools? LessWrong, 2025. URL https://www.lesswrong.com/posts/YGAimivLxycZcqRFR.

Dilgren, C. and Wiegreffe, S. Are latent reasoning models easily interpretable?, 2026.

Geiping, J., McLeish, S., Jain, N., Kirchenbauer, J., Singh, S., Bartoldson, B. R., Kailkhura, B., Bhatele, A., and Goldstein, T. Scaling up test-time compute with latent reasoning: A recurrent depth approach, 2025. Huginn-3.5B.

Goyal, S., Ji, Z., Rawat, A. S., Menon, A. K., Kumar, S., and Nagarajan, V. Think before you speak: Training language models with pause tokens. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

Hao, S., Sukhbaatar, S., Su, D., Lu, X., Shi, W., Radev, D., and Weston, J. Training large language models to reason in a continuous latent space, 2024. (COCONUT).

Karvonen, A., Chua, J., Dumas, C., Fraser-Taliente, K., Kantamneni, S., Minder, J., Ong, E., Sharma, A. S., Wen, D., Evans, O., and Marks, S. Activation oracles: Training and evaluating llms as general-purpose activation explainers, 2025.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukosiute, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Krakovna, V., et al. Evaluating frontier models for dangerous capabilities, 2024.

Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., and He, Y. Codi: Compressing chain-of-thought into continuous space via self-distillation, 2025.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023.

Zelikman, E., Harik, G. R., Shao, Y., Jayasiri, V., Haber, N., and Goodman, N. D. Quiet-star: Language models can teach themselves to think before speaking. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024.

Zhu, R.-J., Peng, T., Cheng, T., Qu, X., Huang, J., Zhu, D., Wang, H., Xue, K., Zhang, X., Shan, Y., et al. A survey on latent reasoning, 2025.

A. Synthetic Model-Organism Details

We include here additional detail on the constructed evaluation-vs.-deployment model organism referenced in Section 6. A COCONUT-style Llama 3.1 8B checkpoint is fine-tuned with a LoRA adapter to behave differently under a short context tag (`[fmt: Q]` vs. `[fmt: A]`). Two probe metrics are relevant. On the full 200-example set, a linear probe on the last-latent-position hidden state distinguishes contexts at AUC 0.993 ± 0.008; the base (non-fine-tuned) model already reaches 0.803 ± 0.045 because some context information is linearly decodable from raw tokens. The $\approx$19-point gap is the fine-tuning-induced signal. Restricting to the same-output subset of 113 examples where both contexts produce the identical correct answer, the fine-tuned probe reaches 0.956 ± 0.031—i.e., the probe-separable internal-state difference survives the behavioral-equivalence control. Two auxiliary organism constructions (a sleeper-agent trigger and a dual-goal scenario) reach AUC 1.000 detection on triggered vs. non-triggered states. We frame all three as existence proofs that latent context variables can be detectable under output equivalence when they are in fact there, not as evidence that naturally occurring deception has been solved.

B. Cross-Scale (4B) Robustness Figures

These figures support the cross-scale (CODI-style Qwen3-4B) results reported in Sections 4 and 5; both are placed here to keep the main body within the page limit. All values are artifact-backed: the Branch-1 probe from a committed probe-metrics file, the layer-asymmetric and logit-lens curves from per-layer dumps, the SAE live-feature counts from a JumpReLU SAE run, and the F1–F6 numbers from committed per-test running summaries.

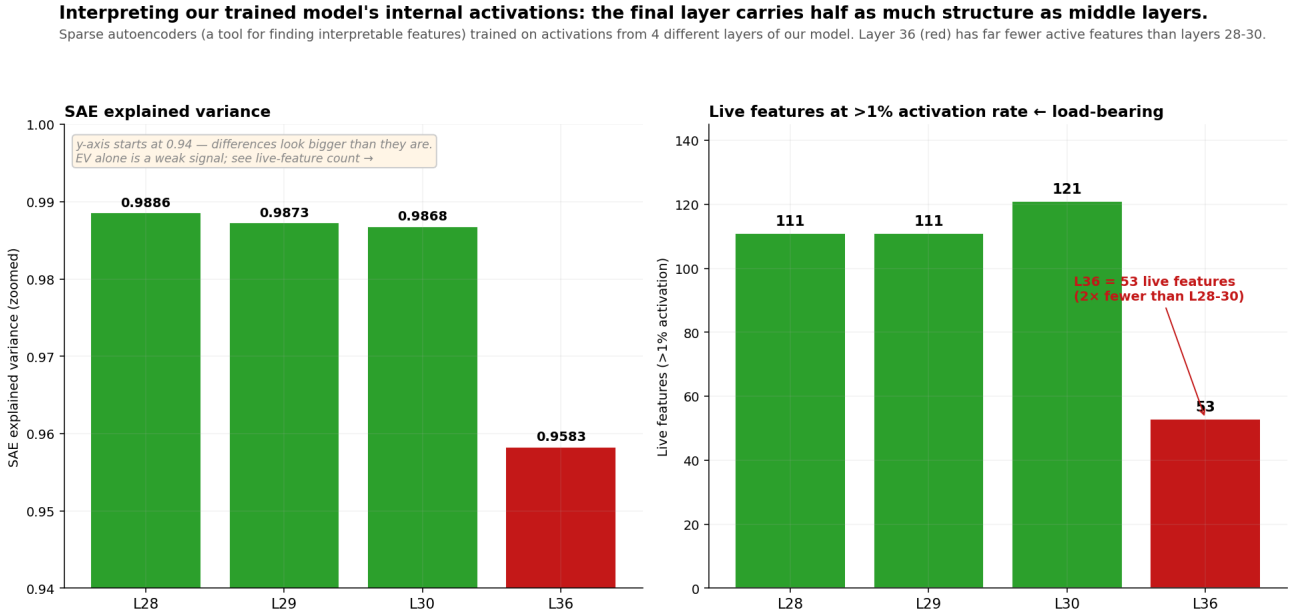


Figure 8. Mid-stack superposition in the 4B model. Sparse autoencoders trained at layers 28–30 versus the final layer 36: explained variance is similar (left), but the final layer carries about half as many live features as the middle layers (right; L36: 53 vs. L28–30: 111–121). The structured content a linear probe recovers at layer 29 (18.06% top-1 vs. 8.33% for a shuffled-label control) lives in this mid-stack superposition, which the unembedding—and hence the plain logit lens—collapses. This is why decodability depends on the readout basis rather than the layer, replicating Section 4 at a second scale.

C. Out-of-Format Behavioral Audit

Table 1 bounds the scope of our interpretability claims: the released CODI checkpoints we study are strongly specialized to their training format.

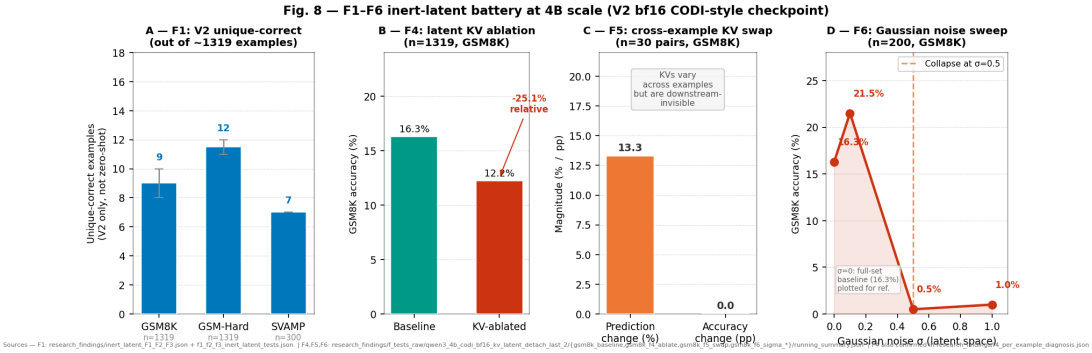


Figure 9. F1–F6 inert-latent battery on the CODI-style Qwen3-4B checkpoint. **(F1)** only 8–10 of 1319 examples are uniquely attributable to a latent. **(F4)** ablating the latent KV entries drops accuracy 25.1% (215 → 161). **(F5)** swapping the latent KV cache across examples changes 13.3% of predictions but accuracy is unchanged. **(F6)** Gaussian noise collapses accuracy by $\sigma = 0.5$. Latents are load-bearing (F4, F6) yet their content is downstream-invisible (F1, F5): the 4B analogue of the 8B pass-ablation result.

	CODI GPT-2		CODI Llama-3.2-1B	
Benchmark	Acc. (%)	Inv. (%)	Acc. (%)	Inv. (%)
GSM8K	10.2	<10	5.5	<10
GSM-Hard	<5	<10	<5	<10
MultiArith	<10	<10	<10	<10
SVAMP	<10	<10	<10	<10
CommonsenseQA	0.0	100	0.0	100
ProntoQA	0.0	100	0.0	100
StrategyQA	0.0	100	0.0	100
P3 ARC-Challenge	0.0	100	0.0	100

Table 1. Out-of-format behavioral audit of the released CODI checkpoints. Accuracy and invalid-output rate (%) across a seven-benchmark harness plus P3 ARC-Challenge format variants. Math benchmarks score in the single digits; every non-math benchmark scores 0% with a 100% invalid-output rate (the models do not emit parseable answers). We use this only to bound scope: our interpretability claims concern these specific format-specialized checkpoints, not latent reasoning in general.