Contents

Activation-Bounded Chain-of-Thought Monitorability

Template-locked reasoning decisions and the structural ceiling on text-only CoT monitoring in Qwen3.6-27B

Position paper, May 2026. Apache-2.0. No additional compute beyond the paper-5 capture corpus.

Abstract

Chain-of-thought (CoT) monitoring has emerged as a leading candidate for scalable AI safety oversight: if a long, serial reasoning process must pass through a textual trace, then reading that trace should reveal what the model is thinking. We argue this view is structurally incomplete in instruction-tuned reasoning models, and we provide empirical evidence from Qwen3.6-27B activation interventions to support the claim. The chat template that activates thinking mode injects a fixed `<think></think>` token pair into the input itself; the decision to *think at all* is therefore encoded in the prompt before the residual stream gets a chance to encode anything else. We document this **template-lock** experimentally (Phase 8 of paper-5: bidirectional α-sweep up to $\alpha = \$ \pm \$200$ in the L55 thinking probe direction produces zero behavioral change) and contrast it with three other reasoning loci where decisions *are* encoded in the residual stream and *are* steerable through activation interventions: capability deployment (L31 pre_tool, +33-40pp pushdown gap across distributions), persona (L43 last_prompt, +60pp pushdown), and mid-reasoning quality (L55 mid_think, +30pp pushup). We use this evidence to argue for an **activation-derived monitorability bound**: text-only CoT monitoring cannot, by construction, observe decisions made before the residual stream encodes them. Activation-derived monitoring is not optional; it is the structurally complementary half of any complete monitoring strategy. We discuss implications for the policy recommendations of the Frontier Model Forum 2026 issue brief and Anthropic's 2027 detection goal.

Keywords: chain-of-thought monitorability, activation steering, linear probes, AI safety, thinking models, Qwen3.6-27B.

---

1. Introduction

The chain-of-thought (CoT) is, on its face, an extraordinary safety opportunity. If a model must verbalise the steps of its reasoning before arriving at an answer, then a sufficiently careful reader of the verbalisation should be able to catch capability misuse, deceptive scheming, latent misalignment, or simple computational errors. The Frontier Model Forum's January 2026 issue brief frames this explicitly: standardized CoT monitorability evaluations, scoring the "clarity, coherence, and faithfulness" of a model's reasoning trace, are presented as a tractable lever for safety governance (Frontier Model Forum 2026). Anthropic's 2027 goal of "interpretability that can reliably detect most model problems" leans heavily on the same hope (Amodei 2025). A 2025 paper by Korbak et al. titled this opportunity "new and fragile" (Korbak et al. 2025).

Recent measurement efforts have begun mapping where the CoT-monitorability hope is and is not justified. Measuring CoT Monitorability Through Faithfulness and Verbosity (arXiv 2510.27378) combines two metrics — does the trace name the relevant factors, and does the trace agree with the actual computation — into a single monitorability score. They report that monitorability differs sharply across model families, that faithfulness classifiers disagree substantially with one another (one yields 69.7% overall faithfulness, another 82.6%), and that thinking-token monitoring captures substantially more signal than answer-text monitoring.

We add a different kind of measurement to this literature: not how well text monitors *agree* on what the trace says, but how much of the *decision* the trace can reflect at all. Specifically, we ask:

> *In an instruction-tuned reasoning model with a thinking-mode chat template, which reasoning decisions are encoded in the residual stream the trace is generated from, and which are encoded in the input tokens of the chat template before the residual stream sees them?*

The answer matters because text-only monitoring can observe only what is generated by the residual stream. Decisions encoded in input tokens — by tokenizer, by template injection, by system prompt — are upstream of generation and cannot, by construction, be present in the trace text in a way that monitor agreement could detect.

We use the activation-intervention protocol of paper-5 (Saturation-Direction Lever, openinterp.org) to classify four reasoning loci on Qwen3.6-27B by where the decision lives. The classification yields three regimes: *template-locked* (input-token-encoded, 100% invisible to residual interventions), *epiphenomenal* (residual-stream-encoded but only as a uniform softmax-temperature shift along the probe direction), and *residual-leverable* (residual-stream-encoded with direction-specific causal authority). We use this classification to argue that text-monitoring strategies have a structural ceiling determined by the proportion of safety-relevant decisions encoded in template-locked positions, and we propose activation-derived monitoring as the structurally complementary half of any complete monitoring strategy.

The argument is simple and the evidence is small (single model, four loci, ~30 prompts per condition). We do not claim a universal result. We claim a *measurable* one, and we believe the measurement should be replicated on additional models before policy frameworks treat text-only monitoring as a plausibly sufficient mechanism for safety oversight.

---

2. The template-injection finding

In Qwen3.6-27B's chat template, enabling thinking mode does not require the model to decide to think; it requires the model to *not refuse* the chat template's pre-injected `<think>` token pair. When `enable_thinking=True` is set on `tok.apply_chat_template`, the rendered input already contains a `<think>` token at a fixed position, and the assistant's task is to emit some reasoning text and then `</think>`. When `enable_thinking=False`, the rendered input contains a `<think></think>` empty pair, signalling to the model that no thinking should be emitted, and the model is expected to skip directly to the answer.

This is a structurally interesting design choice. In a model where the *decision to think* is left to the residual stream, one could imagine training a probe to detect that decision and intervening on it. In Qwen3.6-27B, the decision is not made by the residual stream at inference time. It is made by the chat-template renderer,

before any forward pass. The residual stream, when generation begins, is already conditioned on the input that committed to the decision.

We measured this empirically in Phase 8 of paper-5. We trained a linear probe to detect thinking emission (AUROC 0.91 at N=240, paper-grade gap +0.147 above random K=5 baseline) and applied bidirectional α-sweep up to α=\$±200—$magnitudes substantially exceeding the L2 norm of the residual at the probe's anchor position$ $think>$ $content when the template injects$ $<think>$ $at the start, regardless of whether we add + 200\times probe_direction or − 200×\$probe_direction to the L55 residual at the last prompt token.

We applied the structural-rigidity α-sweep diagnostic (paper-3 §3.3) specifically to distinguish two null-result hypotheses: amplitude-bound null (the probe direction would lever if pushed harder) and structural null (the decision is not in this residual position). The amplitude-bound hypothesis predicts that random direction at α=\$±\$200 would also produce no change, while structural null predicts both probe and random direction produce no change. The data showed the latter: with amplitudes that are *guaranteed* to produce out-of-distribution destruction in random directions on other layers, neither probe nor random produces a behavioral flip on L55 thinking. This is the template-lock signature: the decision is upstream of the residual we are intervening on.

The finding is not a property of the probe quality (the probe detects with AUROC 0.91); it is a property of where the decision is encoded. A perfect probe would still fail to lever a template-locked decision because the template-locked decision is not encoded in the residual stream the probe is reading from in the first place.

---

3. A taxonomy of where reasoning decisions live

Paper-5's protocol applied to eight probes on Qwen3.6-27B yields five empirical classes of probe-causality regime. For the purposes of CoT monitorability, three of those classes are sufficient. We summarise each here with the evidence.

3.1 Template-locked (input-encoded)

Locus: L55 thinking emission decision, last_prompt position. **Probe AUROC**: 0.91 at N=240. **Behavioral lever at α=\$±\$200**: zero change for probe or random. **Diagnosis**: chat template injects the `<think></think>` token pair before the residual stream encodes anything else. The decision is in the input.

Implication for monitoring: text monitors cannot observe the choice to think; the choice was made in the prompt, not the trace. Activation monitors at L55 last_prompt cannot observe it either, because the residual at that layer is downstream of the template injection that committed to the choice. The decision is monitor-invisible at the residual stream level. Detecting this decision would require monitoring the input renderer or the system prompt, not the model's internal state during generation.

3.2 Epiphenomenal (residual-encoded as surface artifact only)

Locus: L43 pre_tool capability decision (the model's choice of which tool action to take next). **Probe AUROC**: 0.83 at N=99. **Behavioral lever at α=\$±\$5**: 4/4 trials select the same tool; log-prob shift at the target tool is $\sim$+0.48 raw, but $\sim$−0.05 after control-token normalisation. The shift is uniform across all next-token candidates rather than specific to the target. **Diagnosis**: probe direction encodes a softmax-temperature artifact — adding the direction makes the model marginally more or less confident in the same distribution, but does not redirect the distribution toward or away from any specific token.

Implication for monitoring: text monitors can observe the tool chosen, and that choice is generated by the residual stream (unlike thinking emission). But intervening on the probe direction does not redirect the choice. A probe in this regime is correlative diagnostic information — useful for detection — but not a control surface. The CoT text reflects the actual decision faithfully; the probe direction does not encode the lever for changing it.

3.3 Residual-leverable (residual-encoded with direction-specific authority)

We document three sub-loci of this regime in Qwen3.6-27B.

Locus a: L31 pre_tool capability lever. Pushdown-asymmetric. +33-40pp probe-vs-random gap at $\alpha = -100$ across HumanEval+MBPP, BigCodeBench, and Codeforces \$≥\$2000 — a saturation-INDEPENDENT lever spanning Qwen pass-rate ∼7-89% (paper-5 §5.5).

Locus b: L43 last_prompt persona lever. Pushdown-asymmetric. +60pp probe-vs-random gap at $\alpha = -200$ (paper-5 §5.6, falsifier #1).

Locus c: L55 mid_think reasoning quality lever. Pushup-asymmetric. +30pp probe-vs-random gap at $\alpha = +200$ (paper-5 §4.4, ReasoningGuard).

In all three sub-loci, the residual stream encodes the decision in a direction-specific way. Activation interventions along the probe direction produce behavioral changes that are statistically distinguishable from same-magnitude random-direction interventions. The decision is both *present* (high AUROC probe) and *causally pliable* (intervention levers behavior).

Implication for monitoring: residual-leverable decisions are *partially* visible to text monitors — the decision is reflected in the generated trace because the trace is generated from the residual stream that encodes the decision. But the residual stream encodes more about the decision than the trace text exposes, because the trace is constrained by the tokenizer's vocabulary and the model's surface-level generation policy. Activation monitors can read the decision in higher fidelity than text monitors can.

---

4. The activation-derived monitorability bound

We propose the following heuristic bound on text-only CoT monitorability:

> Let V be the set of safety-relevant reasoning decisions a model makes for a given task distribution. Partition V into V_T (template-locked, input-encoded), V_E (epiphenomenal at the layer-position of interest, residual-encoded as surface artifact only), and V_R (residual-leverable, direction-specifically encoded). Then text-only CoT monitorability upper-bounds at $|V\_E \cup V\_R| \,/\, |V|$ — the fraction of decisions that are encoded somewhere the residual stream can read and the tokenizer can express.

This is an upper bound, not a lower bound, because text monitoring also pays the costs of unfaithful verbalisation, lossy tokenisation, and generation-policy-imposed selection. The empirical CoT monitorability metric of arXiv 2510.27378, which measures the fraction of decisions actually reflected in the trace, is strictly $\leq |V\_E \cup V\_R| \,/\, |V|$ by construction.

Activation-derived monitoring has a different bound. It can observe decisions in $V\_E \cup V\_R$ at the residual position where they are encoded, regardless of whether the trace text reflects them. In practice, its monitorability ratio is closer to $|V\_R| \,/\, |V|$ — direction-specific decisions can be classified with linear probes; epiphenomenal decisions cannot be reliably distinguished by a single direction. But activation monitoring catches what trace monitoring cannot: decisions in V_R that the model omits from the trace, and decisions in V_E that the trace reports faithfully but where intervention is impossible.

The two monitoring strategies are not redundant. They are *structurally complementary*: text monitoring samples a noisier subset of the union $V\_E \cup V\_R$, while activation monitoring samples a cleaner subset of V_R. Neither catches decisions in V_T, the template-locked set.

For Qwen3.6-27B specifically, our four-locus measurement gives a sketch of the partition for the SWE-bench Pro / code-agent task distribution:

Locus	Class	Detectable in trace text	Leverable via residual
L55 thinking emission	V_T template-locked	partial (emit/skip is observable)	no
L43 pre_tool tool selection	V_E epiphenomenal	yes	not direction-specifically
L31 pre_tool capability	V_R residual-leverable	partial	yes (+40pp pushdown at $\alpha=-100$)
L43 last_prompt persona	V_R residual-leverable	partial (style hints leak)	yes (+60pp pushdown at $\alpha=-200$)
L55 mid_think quality	V_R residual-leverable	yes (verbalised reasoning)	yes (+30pp pushup at $\alpha=+200$)

Two of the four documented decision categories sit in the residual-leverable regime (V_R), one is template-locked (V_T), one is epiphenomenal (V_E). For a hypothetical text-only monitoring strategy on Qwen3.6-27B at this task distribution: at most $4/5 = 80\%$ of these decisions are even visible in principle to a CoT text monitor, and the missing 20% (the choice to think at all) is a safety-load-bearing decision in many alignment scenarios — precisely the kind of decision a deceptive model could exploit by remaining silent.

---

5. Cross-distribution evidence: residual-leverable signals are

distribution-robust where text signals are not

The $\alpha=-100$ robustness theorem from paper-5 §5.5 provides additional evidence for treating activation monitoring as load-bearing. The L31 pre_tool capability probe direction produces +33-40pp probe-vs-random pushdown gap at $\alpha=-100$ across three test distributions spanning Qwen3.6-27B pass-rate ∼7-89% (HumanEval+MBPP at ∼89%, BigCodeBench at ∼55%, Codeforces \$≥\$2000 at ∼7%). The lever is saturation-independent at this amplitude.

Text-monitoring metrics on the same distribution span do not generally share this property. The Korbak et al. measurement notes that monitorability "differs sharply across model families", and the verbosity/faithfulness composite score is itself distribution-dependent (longer prompts produce longer traces with more headroom for omission). This is not a contradiction; it is a structural feature. Activation-derived signals at V_R loci can be more distribution-robust than text-derived signals because the residual-stream encoding of the decision has a stable direction across distributions, while the verbalisation policy that decides what to put in the trace text is itself sensitive to distribution-specific generation-time cues.

For policy frameworks that depend on monitor reliability across the deployment-time distribution shift between safety eval and real-world agentic use (Anthropic 2026 alignment auditing), activation monitoring at V_R loci is the structurally less brittle half of the monitoring stack.

---

6. Implications for safety policy

The Frontier Model Forum's January 2026 issue brief recommends standardized CoT monitorability evaluations, scored with metrics for clarity, coherence, and faithfulness, and used as a basis for governance decisions about deploying frontier reasoning models. We endorse the recommendation and note one structural addition the brief does not currently include.

A monitorability evaluation that scores only what appears in trace text will, by construction, miss decisions in V_T. For thinking-mode chat templates, V_T contains the decision to engage thinking at all. For broader chat-template families, V_T contains every decision encoded in system prompts, function-call schemas, retrieval-injected context, and tokenizer-level pre-processing. A model deployed with a chat template that encodes

capability gating in the template (rather than in the model's own reasoning) would pass a high CoT monitorability score without any activation-level evidence of capability deployment, because the capability gating happens before the residual stream encodes anything.

A complete monitorability framework would therefore include three components:

1. **Text-monitor evaluation** (current Frontier Model Forum recommendation): measure faithfulness and verbosity of the trace text.
2. **Activation-monitor evaluation**: measure the fraction of safety-relevant decisions classifiable by linear probes at the residual stream and leverable via direction-specific intervention. Paper-5's protocol provides one operationalisation; ProbeBench v0.0.1 (openinterp.org/probebench) provides a public leaderboard for AUROC-only versions.
3. **Template-audit evaluation**: enumerate the chat-template-encoded decisions (V_T) and assess them as part of model deployment review, separately from inference-time monitoring.

Component 1 catches what is text-leaked. Component 2 catches what is residually encoded but text-omitted. Component 3 catches what is template-locked and never reaches the residual stream at all. Each component has decisions invisible to the others.

The Anthropic 2027 goal of "reliably detect most model problems" implicitly requires all three. The current Frontier Model Forum recommendation explicitly includes only the first.

The shift in interpretability community priorities reported by DeepMind in 2025 — deprioritising sparse-autoencoder research after finding that linear probes outperformed SAEs on harmful-intent detection (0.999+ AUROC including OOD vs. distinctly worse for k-sparse SAE probes) — is consistent with this framing. Linear probes operate at the residual stream level and are, in our taxonomy, the activation-monitoring tool of choice for $V\_E \cup V\_R$ decisions. SAEs offer a different decomposition but face the same structural ceiling for V_T decisions that linear probes face. Text monitoring, in turn, faces a different ceiling on V_T decisions that is independent of which residual decomposition one uses.

---

7. Limitations and open questions

Single model. All evidence in this paper comes from Qwen3.6-27B. We do not claim that the partition {V_T, V_E, V_R} has the same shape for other reasoning models. The Universal Refusal Circuits paper (arXiv 2601.16034) showed that refusal circuits transfer across models; we do not yet know whether the V_T/V_E/V_R partition transfers across models. Replication on Gemma-2-2B-IT, Llama-3.1-8B, and ideally a Claude-class model is the single largest open task.

Single task distribution. The four loci we documented were measured on SWE-bench Pro / code-agent tasks. Other domains (factual recall, mathematical reasoning, refusal under jailbreak) may have different partitions. Persona, capability, and reasoning quality plausibly generalise; thinking-mode template-lock plausibly does not generalise to non-reasoning models.

Bound is heuristic, not derived from first principles. We have not derived $|V\_E \cup V\_R|/|V|$ from a formal information-theoretic argument. The argument we make is structural (decisions in V_T cannot appear in trace text by construction) but not quantitative beyond the four-locus sketch.

Activation monitoring also has costs. Activation-derived monitors require white-box access to the model's residual stream, which most deployments do not provide to monitoring infrastructure. They also require labelling effort to train probes. The "structurally complementary" framing we propose is conceptual; the operational reality is that activation monitoring is more expensive to deploy than text monitoring in current safety pipelines.

Decisions can be encoded distributively across positions. Recent work on distributed-output-template-driven in-context learning (arXiv 2605.04061) shows that single-position interventions can fail because the relevant computation is distributed across multiple positions. Our V_R classification assigns decisions to single layer-position pairs; a more refined treatment would partition across position-sets.

8. Conclusion

We measured four reasoning loci on Qwen3.6-27B using activation intervention and classified each by where the decision is encoded: template-locked input (V_T, 1/4), epiphenomenal residual surface (V_E, 1/4), or direction-specifically residual-encoded (V_R, 2/4). We argued that text-only chain-of-thought monitoring has an upper bound on monitorability of $|V\_E \cup V\_R|/|V|$ — the fraction of decisions the residual stream encodes and the tokenizer can express — and that activation-derived monitoring complements it by catching the V_R decisions that text omits. Template-locked decisions are invisible to both monitoring strategies and require separate template audit at deployment time.

The argument is small — one model, four loci, ∼30 prompts per condition — but the structural shape generalises. Any reasoning model with a thinking-mode chat template makes its first reasoning decision in the template renderer, before any residual stream sees it. Any chat template at all encodes some decisions in the input that the residual stream cannot reach by intervention. Any monitoring strategy that operates only on what the model emits will, by construction, miss those decisions.

The Frontier Model Forum's standardized monitorability evaluations are a useful start. They will be more useful when augmented with activation-derived evaluations for the residual-leverable decisions text monitors omit, and with template-audit procedures for the input-encoded decisions neither monitoring strategy can reach. The three components are not redundant. They are structurally complementary, and a model evaluated on only the first will leave the third entirely unaudited.

Reproducibility statement

All claims in this position paper rest on the Phase 7-12 capture and intervention corpus of paper-5. The corpus is reproducible in ∼6.5 hours on a single RTX 6000 Blackwell:

Component	Location
Phase 6 N=99 capture corpus	Drive `swebench_v6_phase6/`
Phase 7 L43 pre_tool steering	`nb_swebench_v9_phase8_causal_cot.ipynb` (umbrella)
Phase 8 L55 thinking template-lock	same notebook, structural-rigidity diagnostic
Phase 10 RG L55 mid_think pushup	`nb_swebench_v10_fg_rg_causality.ipynb`
Phase 11/11b/11d capability locus	`nb_swebench_v11_capability_locus.ipynb` and follow-ups
Phase 12 persona pushdown falsifier	`nb_swebench_v12_persona_falsifier.ipynb`
Repository	https://github.com/OpenInterpretability/openinterp-swebench-harness
HF dataset (capture corpus, planned)	`caiovicentino1/openinterp-paper5-saturation-direction`
openinterp SDK	`pip install openinterp` (v0.3.1+)

This paper introduces no new compute. It is a re-framing of paper-5's empirical findings in terms of the chain-of-thought monitorability literature.

References

- Frontier Model Forum (January 2026). *Issue Brief: Chain of Thought Monitorability.* https://www.frontiermodelforum.org/issue-briefs/chain-of-thought-monitorability/

- Korbak, T. et al. (2025). *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety.* https://arxiv.org/html/2507.11473v1
- Measuring Chain-of-Thought Monitorability Through Faithfulness and Verbosity. https://arxiv.org/abs/2510.27378
- Amodei, D. (2025). *The Urgency of Interpretability.* https://www.darioamodei.com/post/the-urgency-of-interpretability
- DeepMind Safety Research (2025). *Negative Results for Sparse Autoencoders On Downstream Tasks and Deprioritising SAE Research.* https://deepmindsafetyresearch.medium.com/negative-results-for-sparse-autoencoders-on-downstream-tasks-and-deprioritising-sae-research-6cadcfc125b9
- Anthropic Alignment Science (2026). *AuditBench.* https://alignment.anthropic.com/2026/auditbench/
- Universal Refusal Circuits Across LLMs. https://arxiv.org/html/2601.16034
- Universal Sparse Autoencoders. https://arxiv.org/abs/2502.03714
- Activation Steering Compromises LLM Safety. https://openreview.net/pdf/987905e26b89ce6e307dcc12de3528b26106pdf
- Conditional Activation Steering (CAST). https://openreview.net/forum?id=Oi47wc10sm
- Vicentino, C. (May 2026). *Saturation-Direction Lever: A Five-Class Taxonomy of Probe Causality in Qwen3.6-27B.* https://openinterp.org/research/papers/saturation-direction-probe-levers
- Vicentino, C. (May 2026). *Two Forms of Epiphenomenal Probes in Code Agents.* https://openinterp.org/research/papers/two-forms-epiphenomenal-probes

---

Acknowledgments

This paper rests on the empirical foundation of paper-5 in the openinterp probe-causality series. The position framing was sharpened by the public discussion around the Frontier Model Forum's January 2026 issue brief and by Anthropic's repeated emphasis that text monitoring alone is unlikely to be sufficient for the 2027 detection goal. Apache-2.0.