Contents

Trajectory-Shaping Probe Steering in Qwen3.6-27B Reasoning: Causal, Cross-Domain, and KV-Cache-Bound

Workshop draft v2 (2026-05-16). Target: NeurIPS 2026 MI Workshop (Sep deadline) with view to ICLR 2027 main paper expansion.

Apache-2.0. Reproducible. Single-author submission, double-blind by conventions.

Abstract

We identify a causally functional subjective-time direction in the residual stream of Qwen3.6-27B (open-weights, 27B parameters, hybrid Gated-Delta-Net plus standard attention), validate it across math (GSM8K) and code (SWE-bench Verified) reasoning, and characterize a fundamental operational constraint on its causal effect: the steering intervention works only when applied continuously from generation start. A Ridge regression probe trained on residuals at L11/L31/L55 predicts thinking-phase completion with R^2=0.82-0.86 (Spearman $\rho \geq 0.90$); three baselines (random-feature, shuffled-target, constant-mean) cleanly fail. Forward-hook steering at L31 with α=+50 from token 1 shortens GSM8K thinking-length in 9/14 prompts vs 2/14 for matched random (Fisher p=0.0092). Cross-domain: 19/20 (95%) probe-clean-termination on SWE-bench Verified across 6 repositories vs 6/20 (30%) random (Fisher p<0.001), at mean 299 thinking-tokens vs unbounded baselines (0/10 terminate even at MAX_NEW_TOK=2048). The mechanism, however, is **trajectory-dependent**: an onset-timing experiment shows that delayed steering — even at decode step 50 — drops termination from 9/10 to 3/10; by step 200, the rescue effect vanishes entirely (0/10). Two closed-loop variants that use the same probe as both sensor and actuator (threshold-trigger Design E across thresholds {0.65, 0.70, 0.85}; plateau-detector Design F at two window/delta configurations) achieve only 1-2/10 termination, confirming that the "termination basin" is mediated through KV-cache state buildup rather than instantaneous residual perturbation. This refines the Belrose et al. (2024) probe-causality taxonomy with a third category beyond "causal" / "epiphenomenal": **operationally-constrained causal** — directions that lever behavior only under specific application protocols. We document Designs E and F as honest negatives and orient the practical intervention (`agent-probe-guard` SDK `anti_overthinking` mode) as a preventive compute-budget enforcer applied from token 1, not an adaptive detect-and-intervene system.

---

1. Introduction

A persistent gap in mechanistic interpretability separates *detection* from *causation*. Linear probes routinely achieve high AUROC or R^2 on residual-stream activations for properties like truthfulness, sentiment, or task-success — yet when those same directions are injected via forward-hook steering at amplitudes appropriate for behavioral intervention, the effects often vanish (Belrose et al. 2024; Caio 2026a, "Two Forms of Epiphenomenal Probes"). The default assumption that "X is in the residual stream" implies "X is causally usable" has accumulated counter-evidence: probes can fit marginal target distributions without per-prompt predictive structure (Caio 2026b, "Marginal-Fit Pathology"); probe-shaped log-probability shifts can be uniform softmax-temperature artifacts (Caio 2026a, Form 1); probes can be locked to template-controlled decisions encoded in input tokens upstream of any residual the probe reads (Caio 2026a, Form 2).

This paper reports the rare opposite — with a twist. We identify a subjective-time direction in Qwen3.6-27B residuals (a linear feature encoding "what fraction of thinking has been completed") and demonstrate it is *causally functional* for controlling thinking-phase termination. The direction passes all standard baselines, exhibits a clean dose-response, preserves end-to-end task accuracy under intervention, and transfers without retraining from GSM8K to SWE-bench Verified across six repositories. But on close mechanistic examination, the causality is **operationally constrained**: it requires continuous application from generation start. A closed-loop design that uses the same probe as both sensor and actuator — the obvious adaptive intervention — fails almost completely (1-2/10 success). An onset-timing experiment shows the failure is not about the sensor;

it is intrinsic to the mechanism. Delaying static steering by as little as 50 decode steps cuts termination from 9/10 to 3/10. By step 200, the effect is dead. The "termination basin" we initially interpreted as a discrete state-attractor turns out to be mediated through the KV-cache buildup of steered residuals — a trajectory-shaping phenomenon, not a state-switching one.

The finding has four contributions:

1. **Empirical**: a probe-grade subjective-time direction in Qwen3.6-27B (R^2=0.86 at L31) that is causally functional for termination control (Fisher p<0.01 across N=20 cross-repo SWE-bench problems) and generalizes cross-domain at 95% rescue rate.
2. **Mechanistic**: probe causality here is *trajectory-dependent*, not state-dependent. The steering effect operates by progressively shaping the K/V cache stored at each decoded token, not by perturbing the current residual into a discrete basin. Onset-timing data confirm this directly.
3. **Negative**: closed-loop variants of the same probe (threshold-triggered Design E + plateau-detector Design F) fail consistently. These are *honest negatives* that save future researchers from chasing the same intuitive but incorrect design pattern.
4. **Practical**: the inference-time intervention is viable but as a *preventive* compute-budget enforcer (continuous steering applied from token 1, achieving 24% compression on GSM8K at zero accuracy loss and ∼70% compression on SWE-bench while maintaining termination), not as an adaptive detection-then-intervention system.

We position this work as a refinement of the broader probe-causality taxonomy: in addition to the "causal" and "epiphenomenal" categories documented in prior work (Belrose et al. 2024; Caio 2026a/b/c), a third category emerges — **operationally-constrained causal**, where the direction levers behavior only under specific application protocols.

---

2. Setup

2.1 Model and capture infrastructure

Qwen3.6-27B (Alibaba, Apr 2026), 64 layers, hybrid Gated Delta Net plus standard attention, bf16 inference on a single GPU (RTX 6000 Blackwell Pro, 96 GB, or A100 80 GB). Residual dimension d_model = 5120. Three layers are studied based on prior causal-locus evidence (Caio 2026c): L11 (early/input), L31 (mid/compositional, the U-shape valley of capability), L55 (late/answer-ready). The chat template is the released Qwen3.6 template with `enable_thinking=True`, exposing the `<think>...</think>` token-pair structure on which all subsequent analysis depends.

2.2 Probe construction (v1)

GSM8K test split, N_prompts = 150 (133 retained after thinking-token-length filtering). For each prompt, we generate the thinking phase under greedy decoding (`do_sample=False, temperature=0`), capture residual activations at four source fractions {10%, 25%, 50%, 75%} of the thinking-token span, and at the end-of-thinking target (100%). Each (layer L, fraction f) yields a tensor of shape (133, 5120).

Per layer L, we pool the 5 fractions into a single dataset of (residual, fraction) pairs: 665 samples in 5120-d residual space with scalar fraction target ∈ {0.10, 0.25, 0.50, 0.75, 1.00}. Train/test split is performed by PROMPT (80/20, seed=42) to prevent leakage across the 5 fraction captures of the same prompt — yielding 530 train samples, 135 test samples per layer.

A Ridge regression probe (`sklearn.linear_model.Ridge`, α=1.0) is fit on the train split for each layer. We report R^2 on test, Spearman rank correlation on test, mean absolute error in fraction-units, and three baselines (§2.3).

2.3 Baseline arms

To control for the failure modes documented in prior work (Phase 6c, [Caio 2026b], [Caio 2026c]), we report three baselines per condition:

- **B0 — Random-feature projection**: 100 random unit-norm Gaussian directions in 5120-d residual space, each used as a 1-d feature for closed-form linear regression on the same train split. Report median R^2 and 5%/95% bounds across the 100 directions.
- **B1 — Shuffled-target retraining**: train Ridge on (X_train, permuted(y_train)), eval on real (X_test, y_test). Detects whether the probe is fitting per-sample structure or merely marginal-distribution fit (Caio 2026b).
- **B2 — Constant-mean prediction**: predict y_train.mean() for every test sample. By definition $R^2=0$; provides the lower-bound reference.

2.4 Causal validation (Phase 2A)

We extract the unit-normed probe direction `w = probe.coef_ / $\|$probe.coef_$\|$` at L31 (the strongest layer). Steering is implemented via a forward-hook on the L31 decoder layer that adds `$\alpha$ $\cdot$ w` to the residual at every token position during greedy generation (no thinking-phase gating; the steering applies throughout). A matched random direction `r` is sampled once from a Gaussian distribution, unit-normed, and used as a parallel control at every α. Steering coefficients sweep $\alpha \in \{-200, -100, -50, +50, +100, +200\}$ — magnitudes ranging from below typical residual norms at L31 to multiples thereof, satisfying the structural-rigidity α-sweep rule (Caio 2026a).

Per generated continuation we measure: thinking_length (tokens between `<think>` and `</think>` if termination occurs, else MAX_NEW_TOK), termination flag (whether `</think>` was emitted), and the full output text for stripped-flip-rate computation against baseline (Caio 2026d).

We sample 15 prompts from the PSAE v1.5 cache (Caio 2026b), filtered to thinking_length < 600 in baseline to avoid MAX_NEW_TOK$=1024$ saturation. One prompt is dropped post-hoc for failing to terminate in baseline, leaving N$=14$ valid prompts $\times$ 13 conditions (1 baseline + 6 alphas $\times$ 2 directions) $= 195$ generations on a single GPU, completed in approximately 2 hours.

2.5 Quality evaluation

For the same 15 prompts we extract the final numeric answer from each generation via a regex (`r'-?\d[\d,]*(?:\.\d+)?'`, last match), compare to the GSM8K gold answer string, and report per-prompt accuracy under baseline and under probe@$+50$ steering. Mean thinking-tokens is computed across all 15 prompts in each condition for compression measurement.

2.6 Dose-response

For 3 prompts (subset of the 15), we additionally sweep $\alpha \in \{+10, +20, +30, +40, +60, +75\}$ to characterize the dose-response curve of probe steering between zero and the saturation point.

2.7 Cross-domain transfer (SWE-bench)

To test whether the GSM8K-trained probe direction generalizes to a substantially different reasoning distribution, we sample 10 problems from SWE-bench Verified (princeton-nlp/SWE-bench_Verified, test split), build minimal user-message prompts of the form `<problem_statement>...{stmt}...</problem_statement>\n\nAnalyze the issue carefully and propose a fix.` (capping problem_statement at 4000 characters), apply the same `enable_thinking=True` chat template, and generate baseline + probe@$+50$ + random@$+50$ with identical steering machinery. No probe retraining; the probe direction comes from the GSM8K Ridge fit unchanged. Compute: 30 generations at $\sim$30-50s each, approximately 25 minutes.

---

3. Probe identification (v1 results)

3.1 Recall and rank correlation

Ridge regression on the 530-train split predicts thinking-fraction on the held-out 135-sample test split with the results in Table 1.

Table 1. Subjective-time probe v1 — recall and rank correlation per layer. REAL is the Ridge probe; B0/B1/B2 are baselines defined in §2.3.

Layer	REAL R^2	REAL ρ	REAL MAE	B0 R^2 (median, 5%-95%)	B1 R^2	B2 R^2
L11	**0.837**	**0.903**	0.099	0.050 (-0.006 – 0.283)	**−0.402**	0.000
L31	**0.858**	**0.915**	0.090	0.090 (-0.029 – 0.372)	**−1.043**	0.000
L55	**0.821**	**0.903**	0.103	0.067 (-0.013 – 0.367)	**−1.019**	0.000

All three layers achieve $R^2 > 0.82$ and $\rho > 0.90$, with MAE in fraction-units of approximately 0.10 (versus the target step-size of 0.225 between adjacent fractions). The B1 shuffled-target baseline gives strongly negative R^2 — Ridge with permuted labels predicts WORSE than the constant-mean, indicating active overfitting to noise rather than passive marginal fit. The B0 random-feature 95th percentile R^2 is 0.37 (L31), well below the REAL R^2 of 0.86, confirming the signal is direction-specific.

3.2 Content-confound caveat

Inspection of the predicted-vs-actual scatter plot reveals a structural pattern: the frac=1.00 cluster predicts extraordinarily tightly at 1.0 across all three layers, while intermediate fractions (0.25, 0.50, 0.75) show wider spread with overlap between adjacent predictions. This indicates that part of the high aggregate R^2 is driven by the trivially-distinguishable end-of-thinking residual state (where content shifts to solution-format language: "therefore", "the answer is", numeric expressions). The probe is reading temporal position partially via content distinctiveness rather than as a pure positional embedding. The Spearman $\rho \geq 0.90$ across all layers nevertheless confirms the rank-ordering signal is real at intermediate fractions as well; we frame the finding as **"time encoded mediated by content distinctiveness"** rather than pure positional self-time.

---

4. Causal validation (Phase 2A)

4.1 Thinking-length response to α

Mean thinking-length and termination rate per (direction, α) across N=14 valid prompts:

Table 2. Phase 2A α-sweep at L31. Baseline mean thinking_length = 529.8 tokens, mean terminate rate = 0.93.

Condition	Mean thinking_len	Mean Δ vs baseline	Terminate rate
baseline	529.8	0.0	0.93
probe α=+50	**402.1**	**−127.7**	**1.00**

Condition	Mean thinking_len	Mean Δ vs baseline	Terminate rate
random α=+50	561.1	+31.3	0.87
probe α=+100	459.5	−70.3	0.40
random α=+100	10.5	−519.3	0.27
probe α=+200	20.5	−509.3	0.00
random α=+200	1024.0 (cap)	+494.2	0.00
probe α=−50	975.5	+445.7	0.20
random α=−50	937.8	+408.0	0.13
probe α=−100	755.3	+225.5	0.00
random α=−100	957.9	+428.1	0.00
probe α=−200	1024.0 (cap)	+494.2	0.00
random α=−200	1024.0 (cap)	+494.2	0.00

The cleanest signal is at α**=+50**: the probe direction shortens thinking by 128 tokens on average and maintains 100% termination, while the matched random direction extends thinking by 31 tokens and degrades termination to 87%. This is the regime in which steering remains within the basin of normal generation while pushing toward termination.

4.2 Per-prompt shortening rate and Fisher exact

We test direction-specificity at α=+50 by counting prompts whose thinking-length shortens by more than 10% relative to baseline:

- **Probe@+50 shortens >10% in 9/14 prompts (64%)**
- **Random@+50 shortens >10% in 2/14 prompts (14%)**

Fisher exact test on the 2$×$2 contingency (shortens vs not, by direction): **Odds ratio = 10.8, p = 0.0092.**

The probe direction is roughly 5$\times$ more likely to shorten thinking than a matched random direction. The mean Δ% gap (probe −17.6% vs random +15.2%) is **32.8 percentage points**.

4.3 Saturation regime and degenerate breakdown

At α=$±$200, both probe and random enter a degenerate regime: probe collapses to $\sim$20-tokens with no terminate; random hits MAX_NEW_TOK cap. These are different failure modes, not "similar effects" — at extreme amplitudes the residual is perturbed far outside its typical norm and the model's distribution becomes unrecognizable. Our auto-classifier in an earlier analysis confused these for "softmax-temperature artifact" by inspecting only α=$±$200; the clean signal lies at α=+50.

At α=+100, the probe enters a partial-collapse regime: 5/14 prompts terminate with thinking_length $\leq$ 2 (model emits `</think>` immediately), 6/14 hit cap without termination. The random direction at α=+100 produces 4/14 collapses but mostly without proper termination (truncated garbage outputs). The probe collapse is *semantically clean* termination; the random "collapse" is *generation failure*. This distinction is important: probe direction navigates the termination attractor; random direction merely breaks the model.

4.4 Negative-α asymmetry and the saturation-direction principle

The negative-α regime is structurally different from positive-α. Both probe and random at α=−50, −100, −200 produce indistinguishable extension-to-cap with low or zero termination rate. The probe direction does NOT symmetrically push the model toward "just started" thinking — instead, any sufficiently large perturbation in the residual "earlier" half destabilizes generation without engaging a coherent attractor.

This is consistent with the saturation-direction principle (Caio 2026c): the model has a discrete "termination basin" feature (the decision to emit `</think>`) that probe-direction +α aligns with, but there is no symmetric "just-started" attractor for the model to fall into — because "continue thinking" is the default action, not a

feature that needs to be represented. Probe direction in the $-\alpha$ regime is therefore behaviorally equivalent to random noise.

4.5 Critical caveat: steering must be persistent from token 1

All causal effects reported in this section are observed under steering applied at **every** generation token from the first generated token onward. This protocol seems incidental to the methodology but turns out to be load-bearing for the mechanism. §7.4 establishes that delaying onset by 50 decode steps drops the Phase 2A termination effect from 9/10 to 3/10; delaying by 200 steps eliminates it entirely. The "basin" interpretation we initially adopted (steering pushes the residual into a discrete state-attractor) is therefore incomplete. The full account is given in §7.4 and the Implications section (§10).

---

5. Dose-response: a phase transition, not a smooth slope

We sweep α at finer resolution {+10, +20, +30, +40, +60, +75} on 3 prompts to characterize the response curve between zero and saturation. Results for prompt 3 (baseline=471 tokens), prompt 4 (537), prompt 17 (354):

Table 3. Per-prompt thinking_length response across fine-grained α.

α	prompt 3	prompt 4	prompt 17
baseline	471	537	354
+10	935 (+98%)	412 (−23%)	727 (+105%)
+20	597 (+27%)	425 (−21%)	287 (−19%)
+30	1024 (+117%)	394 (−27%)	421 (+19%)
+40	731 (+55%)	647 (+20%)	580 (+64%)
+50	**368 (−22%)**	**436 (−19%)**	**213 (−40%)**
+60	190 (−60%)	341 (−37%)	224 (−37%)
+75	138 (−71%)	118 (−78%)	68 (−81%)
+100	1	1024	368

The dose-response is **NOT a smooth slope** — it is non-monotonic in the α=+10 to +40 region, with all three prompts showing some extension at α=+40, and then transitions abruptly to consistent shortening at α=+50. We interpret this as a discrete *basin transition*: below α=+50 the perturbation is insufficient to overcome the model's natural thinking momentum (and may even destabilize toward extension); at α=+50 the steering crosses into the termination basin and the model's natural dynamics carry it the rest of the way.

This dose-response shape is itself evidence for the basin interpretation. A smooth gradient response (the null model) would suggest the probe direction encodes a continuous variable that the model linearly increments toward termination. The phase-transition pattern instead suggests an attractor structure: small pushes don't matter, but a push across the threshold falls into a different equilibrium.

---

6. Quality preservation

The steering intervention is useful only if it does not degrade end-to-end task performance. We evaluate accuracy on all 15 GSM8K prompts (including prompt 44 whose baseline did not terminate) under baseline and under probe@+50 steering. Final numeric answer extracted by regex and compared to the GSM8K gold answer.

Table 4. Quality evaluation, N=15 GSM8K prompts.

Condition	Mean thinking-tokens	Accuracy
Baseline	530	12/15 (80%)
Probe @ α=+50	**402**	**12/15 (80%)**
Δ	**−128 (−24%)**	**+0 (0pp)**

Probe@+50 achieves 24% thinking-token compression with exactly preserved accuracy. No statistically detectable effect on task success at this N.

A finer-grained inspection reveals two interesting individual cases:

- **Prompt 44 (RESCUE)**: baseline hit MAX_NEW_TOK=1024 without terminating, produced an incorrect intermediate value (7.5); probe@+50 forced termination at 534 tokens and produced the correct answer (15). The intervention rescued an overthinking-loop failure.
- **Prompt 67 (DAMAGE)**: baseline correctly produced answer 70; probe@+50 produced 30 (wrong). The intervention introduced a single error.

Net effect on accuracy across N=15 is zero, but the bimodal pattern is worth noting in deployment contexts: probe@+50 can rescue some failures and break some successes. The expected utility depends on baseline failure rate; on a benchmark where baseline accuracy is already high (e.g., 80% GSM8K here), zero net effect is the expected outcome.

The thinking-length distribution under probe@+50 is bimodal: 11/15 prompts are shortened (mean −42% compression on those), 4/15 prompts are extended (mean +20%). The shortening regime dominates aggregate compression.

---

7. Cross-domain transfer: SWE-bench Verified

The most important question for practical utility is whether the GSM8K-trained probe direction generalizes to substantially different reasoning distributions. We sample N=20 problems from SWE-bench Verified across six repositories: 10 from astropy (the first 10 alphabetical) as the initial cross-domain probe, plus 10 stratified across django, sympy, sphinx, matplotlib, and scikit-learn (top-5 most-populated non-astropy repos, 2 problems per repo, seed=42). We apply the GSM8K probe direction directly with no retraining.

Table 5. Cross-domain transfer on SWE-bench Verified, per-repo breakdown (N=20 total). All baselines hit MAX_NEW_TOK=1024 without terminating across both budget conditions.

Repository	n	Probe@+50 clean rescue	Random@+50 rescue	Probe mean tokens	Random mean tokens
astropy	10	**10/10**	3/10	325	759
django	2	**2/2**	1/2	297	706
sympy	2	**2/2**	0/2	182	—
sphinx	2	**2/2**	0/2	320	—
matplotlib	2	**1/2**	1/2	277	894
scikit-learn	2	**2/2**	1/2	275	903
Total	**20**	**19/20 (95%)**	**6/20 (30%)**	**299**	**797**

20/20 baselines exhibit overthinking-cap failure — the model thinks for the entire 1024 token budget without emitting `</think>`. This is in stark contrast to GSM8K, where only 1/15 baseline hit cap. SWE-bench problems are intrinsically more demanding and the model's natural termination tendency is weaker.

The GSM8K-trained probe direction produces clean termination in 19/20 (95%) cases at mean 299 tokens (29% of the cap). Fisher exact on the 2$×$2 contingency (probe clean rescue vs random rescue,

N=20) gives odds-ratio ≈ 44 and $p < 0.001$. A matched random direction terminates only 6/20 (30%), and those rescues occur substantially later (mean 797 tokens, 78% of cap) versus the probe's surgical 299. The remaining 14/20 random failures still hit cap.

To control for the possibility that the rescue is an artifact of MAX_NEW_TOK=1024 being too restrictive (i.e., baselines would terminate naturally at higher cap), we re-test the 10 astropy baselines at MAX_NEW_TOK=2048. **0/10 terminate naturally even at the doubled budget** — every baseline continues to overflow. The overthinking failure is genuine, not a budget artifact.

7.1 The one breakdown case

The single non-rescue in cross-repo (matplotlib-22865) produces a different failure mode: probe@+50 terminates generation at 161 tokens via EOS (not `</think>`). The model emits a short truncated output without engaging the proper thinking-end protocol. This is consistent with the partial-collapse regime observed at $\alpha=+100$ on GSM8K (§4.3, 5/14 collapses) — the probe direction occasionally pushes the model past the termination basin into a degenerate state rather than landing cleanly in it. At $\alpha=+50$ on cross-repo SWE-bench the breakdown rate is 1/20 (5%); on GSM8K it was 0/14.

7.2 What WOULD have been a SDK feature: closed-loop intervention (Design E, falsified)

The natural design for an inference-time anti-overthinking SDK is closed-loop: use the same probe as both sensor (predicting the current "thinking fraction" from the residual at each token) and actuator (injecting $\alpha=+50$ steering once a threshold is crossed). This would preserve depth on prompts that genuinely need exploration (no steering until the model is "near done") and only commit to termination when needed.

We implemented this as Design E, using the unnormalized Ridge coefficients + intercept to compute the predicted fraction at each generated token, and triggering persistent $\alpha=+50$ steering once the prediction crossed a threshold (sweeping {0.65, 0.70, 0.85}, with `min_decode_steps=50` to avoid triggering on prefill noise).

Result: 1-2/10 termination across all three thresholds.

Per-prompt diagnostics reveal what happened: - At threshold 0.85, the trigger never fires on any of the 10 prompts. Maximum predicted fraction during overthinking ranges 0.63–0.82; the model's residual during stuck thinking never approaches the "fraction=1.0" probe direction the basin interpretation predicted. - At threshold 0.65, the trigger fires on 8/10 prompts (committing at mean decode step ≈ 444), but only 1/10 actually terminates. The other 7/10 continue to MAX_NEW_TOK=1024 with $\alpha=+50$ active for 400-700 tokens after commit. - At threshold 0.70: same pattern — fires on 6/10 (mean commit ≈ 692), terminates 1/10.

The two cases that DID terminate were ones where the trigger fired very early (sphinx-7889 committed @ step 99; scikit-12585 committed @ step 253). When commit happens at step 400+, no closed-loop variant ever produces termination.

7.3 Plateau detection also fails (Design F)

A natural alternative to absolute-threshold triggering is plateau detection: if the predicted fraction stops changing meaningfully over a window of recent tokens, the model is "stuck" and we trigger termination. We implemented this as Design F with rolling window 100 and 50, plateau threshold $\delta=0.02$.

Result: 0/10 termination at either configuration. The trigger never fires because the predicted fraction oscillates more than 0.02 in any reasonable window during overthinking — the residual is dynamic at the per-token level even when the model is making no semantic progress. Plateau detection at the residual level is not a viable overthinking signal in this regime.

7.4 Onset-timing experiment isolates the mechanism

To distinguish whether the closed-loop failure was a property of the sensor (threshold triggers in the wrong place) or the steering itself (late application is fundamentally weaker), we ran a sensor-free deterministic experiment: static $\alpha = +50$ with a delayed onset at decode step $\in \{50, 200, 400\}$ on the same 10 cross-repo SWE-bench prompts.

Table 6. Onset-timing experiment, N=10 cross-repo SWE-bench. The "from token 1" row is the Phase 2A canonical static protocol; subsequent rows progressively delay the steering onset.

Steering onset	Termination	Mean length when term
from token 1 (Phase 2A canonical)	**9/10**	269
from decode step 50	3/10	719
from decode step 200	**0/10**	—
from decode step 400	**0/10**	—

The curve is monotonic and steep. By decode step 200 — well within typical pre-termination thinking lengths — the same direction at the same α has no detectable effect. **The causal mechanism is trajectory-dependent**: it operates by progressively shaping the K/V cache as it is built up, not by pushing the current-token residual into a discrete basin. Once the K/V cache has accumulated 200+ unsteered tokens, the attention mechanism's weighted average over the full cache is dominated by unsteered context; perturbations to new-token residuals do not propagate backward into cached state.

This explains the closed-loop failure mechanistically. Closed-loop triggers fire mid-generation, by which point the unsteered K/V cache has already accumulated to a state that dominates downstream attention. Even if the sensor were perfect, the actuator cannot reach into the past to retroactively steer the cache. The intervention only works when applied from a clean cache — that is, from the first generated token.

7.5 Refined interpretation: termination "basin" as cache-mediated bias

The §4 basin interpretation is not wrong but is incomplete. The corrected mechanism: continuous steering from generation start adds a small bias ($\alpha \times$ probe direction) to the residual at every L31 forward pass. This biased residual is what gets cached as K/V at each new token position. The attention layers in subsequent generation steps attend over this *biased* cache, producing logits that are slightly more "termination-favoring" at every token. After hundreds of accumulated tokens of small bias, the cumulative effect on the model's distribution over next tokens favors `</think>` emission strongly enough to consistently produce termination.

This is a *trajectory-shaping* mechanism, not a *state-attractor* mechanism. The "basin" language was a useful first-pass description but mis-locates the active component: the basin is in the K/V cache, not in the single-token residual. This distinction has direct practical consequences for SDK design (§10).

7.6 Universality interpretation

With the trajectory-shaping mechanism in place, the cross-domain transfer of a GSM8K-math-trained probe to SWE-bench code-debugging — replicating across six repositories without retraining and with consistent compression to ~30% of the cap — points to a feature of *reasoning per se*. The K/V-cache bias from the same direction works on both math and code reasoning, suggesting the underlying mechanism (whatever the residual at L31 represents about "near-end thinking") is shared infrastructure across reasoning domains, not a domain-specific artifact.

7.1 Quality of post-rescue thinking

We do not in this paper evaluate whether the post-rescue thinking content is *correct* (i.e., whether the model identifies the right bug and proposes the right fix). Such evaluation requires either docker-based patch evaluation against the SWE-bench test suite, or LLM-judge content scoring, both of which are out of scope

for this experimental snapshot. We document the rescue rate as the unambiguous mechanical finding and reserve correctness validation for follow-up work (see Limitations).

---

8. Related work

Behavioral self-time studies

Recent work (arXiv:2604.00010, "Can LLMs Perceive Time?", ICLR 2026) reports that LLMs estimate their own task durations with 4-7$\times$ error and concludes that models lack temporal self-awareness. The authors explicitly note that simple scaffolding (timestamp injection) does not resolve the failure and call for "training with explicit timing signals and architectures that better retain temporally grounded state." Their study is purely behavioral; they do not probe internal representations.

Our work complements and complicates theirs. The subjective-time information *is* in the residual stream (R^2=0.86 at L31, this paper §3). The behavioral failure they observe must therefore reflect either (a) a decoder-side disconnect — the verbalized estimate does not query the residual representation — or (b) a representation that is causally functional for termination control (this paper §4) but not for explicit numerical estimation. Either way, the gap between behavioral and mechanistic evidence is now characterized.

Probe causality and epiphenomenality

Belrose et al. (2024) document multiple cases of "epiphenomenal probes" — high-AUROC linear features that do not lever model behavior under steering. Caio (2026a) extends this with a two-mechanism taxonomy: softmax-temperature artifact (probe direction shifts the entire output distribution uniformly), and template-locked decision (probe target is encoded in input tokens upstream of the residual). The current paper provides the positive complement: a probe that IS causal, with explicit structural conditions documented.

Saturation-direction principle

Caio (2026c) introduces the saturation-direction principle: probes lever model behavior in the direction of the residual's baseline saturation, with asymmetric response on the opposite side. The Phase 2A asymmetry observed here (probe@$+\alpha$ functional, probe@$-\alpha$ equivalent to noise) is consistent with this principle and provides further validation.

Inference-time test-time scaling

DeepSeek-R1, OpenAI o1, and related work on test-time scaling have documented that longer chains-of-thought sometimes improve and sometimes degrade reasoning performance (the "overthinking" pattern). To our knowledge no prior work has proposed a *mechanistic* probe-guided intervention for the overthinking failure mode; our cross-domain SWE-bench rescue (§7) is, we believe, the first such demonstration in an open-weights reasoning model.

Steering in instruction-tuned models

The forward-hook steering methodology used here is standard (Turner et al. 2023; Anthropic activation patching work 2024-2026). Our novel methodological contributions are (a) the asymmetric-α-sweep diagnostic (§4.4), which detects directional asymmetry in steering response and connects it to the saturation-direction principle; and (b) the cross-domain transfer protocol (§7), which uses zero-retrain application of a math-trained probe to code-debugging to isolate "is the feature reasoning-general or domain-specific?".

---

9. Limitations

1. **Single model**. All results on Qwen3.6-27B. The subjective-time direction, the trajectory-shaping mechanism, and the KV-cache lock-in finding may not generalize identically to other reasoning models (DeepSeek-R1, o1, future Claude reasoning variants). Replication on $≥$1 other open-weights reasoning model is necessary before claiming a universal property of reasoning LLMs.

2. **Scope of the KV-cache hypothesis**. The trajectory-dependence finding is established for one specific direction (subjective-time at L31). Whether it generalizes to *all* probe-steering interventions in transformer reasoning models is an open question. The hypothesis predicts that any layer-L steering applied late should fail similarly, but only direct measurement on additional probes can confirm.

3. **Quality evaluation N is small** (N=15 GSM8K). The 24% compression at zero accuracy loss finding is directionally clean but not statistically robust. Scaling to N$≥$100 across GSM8K plus at least one cross-domain benchmark (MATH, StrategyQA) is the immediate follow-up.

4. **SWE-bench coverage**. We test N=20 problems across 6 repositories. Pattern is consistent across all 6 (5/5 cross-repo show probe > random) but per-repo N is small (n=2 outside astropy). Scaling to $≥$10 problems per repo across the full SWE-bench Verified taxonomy is necessary before claiming "universal across SWE-bench reasoning distribution".

5. **No patch-correctness evaluation**. SWE-bench rescue is measured by terminate-rate and thinking-length only. Whether the post-rescue thinking content correctly diagnoses bugs and proposes correct fixes requires docker-based patch evaluation against test suites, which we have not run. Visual inspection of N=3 cases (django, sphinx, matplotlib) suggests probe-rescued outputs are structurally complete but shallower than baseline trajectories — a tradeoff that needs proper quantification.

6. **No multi-turn agent rollout**. SWE-bench in practice runs as a 20+ turn agent loop with tool calls. We test only the first-turn thinking phase. Whether the probe-guided rescue holds across full agent rollouts (and whether it improves end-to-end task success) requires the full SWE-bench harness, which is a separate ∼6h experiment.

7. **Greedy decoding only**. All generations use `do_sample=False, temperature=0`. Behavior under sampling (temperature > 0, top-p) is untested. The trajectory-shaping interpretation predicts the K/V-cache mechanism should remain dominant under sampling within reasonable temperatures, but this requires verification.

8. **L31-specificity (resolved by Phase 2C)**. We initially suspected the L31-specific scope might reflect untested layers rather than a layer-specific phenomenon. Phase 2C addressed this: the v1 probe at L11 (R^2=0.84) and L55 (R^2=0.82), tested with calibrated α magnitudes per layer (L11: 5, 10, 25, 50; L55: 50, 100, 200, 500), produces *no controlled termination*. L11 at α=+25 produced one ragged collapse case (1/10, with `</think>` forced after 15 tokens mid-thought and a long degraded answer afterward — visual inspection); L55 produced no termination at any α up to +500. The subjective-time direction is therefore causally functional **only at L31**, despite similar probe R^2 at all three layers. **R^2 is not predictive of causal authority** — a clean second operational constraint (spatial) alongside the temporal constraint from §7.4.

9. **Content-confound in v1 probe**. The R^2=0.86 is partially driven by content distinctiveness at the end of thinking (§3.2). A normalized-residual variant of the probe would isolate the position-pure component. We do not implement this here; the causal validation in §§4 and 7 does not depend on whether the signal is content-mediated or position-pure.

10. **Closed-loop design space not exhausted**. We tested two adaptive designs (threshold-trigger Design E + plateau-detector Design F). More complex designs — multi-stage commits with progressively stronger α, learned commit timing, layer-multiplexed steering — could in principle work. We have shown that the obvious single-direction single-trigger versions fail; we have not shown that no closed-loop variant can succeed.

---

10. Implications and SDK

10.1 Mechanistic interpretation, revised

The §4 "basin" interpretation, in light of the §7.4 onset-timing data, becomes: the subjective-time direction at L31 of Qwen3.6-27B encodes a feature that, when used as a continuous steering signal applied throughout generation, biases the K/V-cache buildup toward an attention regime in which `</think>` emission is statistically favored. This is *not* a discrete state-attractor that the current-token residual can be pushed into; it is a trajectory-shaping bias that takes effect cumulatively across hundreds of cached attention computations.

This refines the broader probe-causality taxonomy. Probes that lever behavior fall into at least three classes (rather than the prior binary causal-vs-epiphenomenal):

1. **Detection-only / epiphenomenal**: high probe accuracy, no behavioral effect under any steering protocol (Caio 2026a Forms 1 & 2; Caio 2026b PSAE).
2. **State-attractor causal**: high probe accuracy + steering effect at the token-level instantaneously. (We do not have a clear example in our corpus; this category may be empirically rare.)
3. **Trajectory-shaping causal** (this paper): high probe accuracy + steering effect ONLY when applied continuously from generation start. K/V-cache state is the active intermediary. Phase 2C (§9 limitation 8) further establishes that within this category, the steering may also be layer-specific: the v1 probe at L11 and L55 has equivalent R^2 to L31 but is non-causal under cross-layer α calibration. This is consistent with the broader view that the "operationally-constrained causal" category has multiple constraints — temporal (apply from token 1), spatial (apply at the specific layer where the direction is operationally aligned with the decision being intervened upon).

Standard reporting practice for probe-steering interventions should therefore include the onset-timing experiment as a diagnostic to distinguish classes 2 and 3, and a cross-layer α-calibration to verify that probe R^2 is not being conflated with causal authority. We propose this as a methodology contribution alongside the existing diagnostics (random-feature baseline, shuffled-source baseline, control-token normalization, structural-rigidity α-sweep, whitespace-stripped flip metric).

10.2 Practical SDK feature

The `agent-probe-guard` SDK (Caio 2026e, PyPI v0.3.0) currently ships probe-based detection in detect-only mode. The Phase 2A + 2B findings here justify adding a preventive intervention mode — but the design space is now constrained:

```
guard = AgentProbeGuard(
    model="Qwen/Qwen3.6-27B",
    mode="preventive_compute_enforcement",   # NOT adaptive_anti_overthinking
    subjective_time_layer=31,                # hard-coded — L11/L55 inert (Phase 2C, §9.8)
    steering_alpha=50,
    onset="generation_start",                # MUST be from token 1 (closed-loop falsified, §7.4)
)
```

The behavior is **preventive compute enforcement**: as soon as the model begins generating its `<think>` section, persistent $\alpha = +50$ steering is applied to every L31 forward pass. On naturally-terminating prompts (e.g., GSM8K), this produces $\approx$24% thinking-token compression at preserved end-to-end accuracy. On overthinking-cap-prone prompts (e.g., SWE-bench), it produces clean termination at $\approx$30% of the natural budget.

What this mode is **not**: an adaptive detection-then-intervene system. Closed-loop and plateau-based designs were tested and falsified in §7.2-7.3. A system that lets the model think freely until it "looks stuck" and then commits to termination — the intuitive design — does not work for this mechanism. Future SDK versions could explore more complex closed-loop variants (limitation 10), but the v0.2 release will ship only the validated `preventive_compute_enforcement` mode with explicit documentation that the trigger must be at generation start.

10.3 Implications for the broader steering literature

If the KV-cache lock-in mechanism documented here generalizes to other probe-steering interventions in transformer reasoning models (an open question; see limitation 2), it has direct implications for activation-steering work writ large:

- **Single-shot residual perturbation experiments** (the standard "apply steering, generate one continuation" protocol) measure *trajectory-shaping* effects, not pure *state-attractor* effects. The two are conflated in current literature.
- **Late-onset interventions** in agent settings (e.g., applying steering only after observing N tool calls) may be systematically weaker than they appear in offline single-shot evaluation, due to the KV cache already accumulating before the trigger fires.
- **The natural way to deploy probe interventions** is preventive: from the start of the relevant generation segment, not as a mid-stream commit. This argues against the architectural pattern of "monitor a long-running agent and intervene when probe fires" and in favor of "frame the agent's task and apply steering from the start of each agent turn".

These implications are speculative pending cross-probe replication. We flag them as testable predictions of the trajectory-shaping interpretation.

---

11. Conclusion

We identify a causally functional subjective-time direction in Qwen3.6-27B residual streams. The probe achieves R^2=0.86 with three clean baselines (§3); the direction causally controls thinking-phase termination at α=+50 with Fisher p=0.0092 vs matched random on GSM8K (§4); the dose-response curve exhibits a phase-transition near α=+50 (§5); the intervention preserves end-to-end task accuracy at 24% mean compute reduction on GSM8K (§6); and the direction transfers without retraining to SWE-bench Verified code-debugging across six repositories, where it produces clean termination in 19/20 (95%) cases at mean 299 thinking-tokens vs unbounded baselines that fail even at MAX_NEW_TOK=2048 (§7, §7.1).

But the mechanism is not what the basin metaphor suggests. Two closed-loop interventions — Design E (threshold-trigger) and Design F (plateau-detector) — that should work if the probe direction were a state-attractor instead produce 0–2/10 termination across all configurations tested (§7.2, §7.3). A sensor-free onset-timing experiment isolates the cause: static steering applied from token 1 produces 9/10 termination; delayed to decode step 50, the rate drops to 3/10; delayed to step 200, the effect vanishes entirely (§7.4). The "termination basin" is mediated through the K/V cache accumulated across the full generation trajectory, not through any single-token residual perturbation. We call this *trajectory-shaping* probe causation, and document it as a third category in the probe-causality taxonomy (§10.1) alongside the existing detection-only and (hypothetical) state-attractor categories.

The methodology contributions are: the asymmetric-α-sweep + terminate-rate metric (§4), the cross-domain transfer protocol via no-retrain steering application (§7), and the onset-timing diagnostic (§7.4) for distinguishing state-attractor from trajectory-shaping causal probes. The practical implication for the `agent-probe-guard` SDK is that `anti_overthinking` mode must apply steering from generation start, not as a post-hoc trigger; it is a deterministic compute-budget enforcer, not an adaptive detect-and-intervene system (§10.2). The broader implication for activation steering — speculative pending cross-probe replication — is that single-shot residual perturbation experiments routinely conflate trajectory-shaping and state-attractor effects, and late-onset interventions in agent settings may be systematically weaker than offline measurement suggests (§10.3).

This is the first positive causal probe in the OpenInterpretability methodology corpus, complementing six prior epiphenomenal/honest-negative cases. It is also — with Designs E and F as documented honest negatives — the first paper in the corpus to combine a positive empirical claim with explicit falsification of the natural-but-incorrect adaptive design. We see both as essential to making probe-based interventions trustworthy.

Code, data, artifacts

- v1 notebook: `nb_subjective_time_probe_v1.ipynb`
- Phase 2A notebook: `nb_subjective_time_phase2a_steering.ipynb`
- Phase 2B notebook: `nb_subjective_time_phase2b_steering_designs.ipynb` — bundles all 6 follow-up experiments (Caveat #1 cross-repo, Caveat #2 budget extension, α-sweep B, Design E closed-loop, Design F plateau, onset-timing diagnostic)
- Phase 2B reproduction guide: `REPRODUCTION_subjective_time_phase2b.md` — execution order, output schema, decision tree
- Cached residuals + features (43 MB, reusable): HF dataset openinterp-psae-v15-marginal-fit-pathology
- Phase 2A + 2B results JSONs: Drive `openinterp_runs/subjective_time_phase2a/caveat1_cross_repo/`
- SAEs (reference only; not used here): `caiovicentino1/qwen36-27b-sae-papergrade` (Apache-2.0)
- Methodology rules canonicalized in memory: feedback files for Phase 6c, shuffled-source baseline, control-token normalization, structural-rigidity α-sweep, whitespace-stripped flip metric, asymmetric-α-sweep (Phase 2A), **onset-timing diagnostic (this paper, Phase 2B)**.

Acknowledgments

[TODO — none needed for double-blind submission; thank reviewers + Anthropic Persona Vectors team + Belrose tuned-lens lineage in non-blind final version]

References

[TODO — bibliography. Key cites: Belrose et al. 2024 tuned lens; arXiv:2604.00010 "Can LLMs Perceive Time?"; Marks et al. 2024 Sparse Feature Circuits; DeepSeek-R1 paper; Caio 2026a "Two Forms of Epiphenomenal Probes"; Caio 2026b "Marginal-Fit Pathology"; Caio 2026c "Saturation-Direction Lever Taxonomy"; Caio 2026d "agent-probe-guard SDK"; Turner et al. 2023 activation steering]

Last updated 2026-05-16. v2 paper draft. Title and Sections 1, 4.5, 7.2-7.6, 9, 10, 11 substantially revised based on Phase 2B onset-timing experiment that falsified the discrete-basin / state-attractor interpretation and replaced it with the trajectory-shaping / KV-cache-mediated interpretation. Limitations §9 items 5-6 (patch correctness, multi-turn rollout) are immediate follow-up experiments; items 1-2 (cross-model, KV-cache scope across other probes), 7-8 (sampling, cross-layer) are scope for main-conference expansion. Item 10 (richer closed-loop variants) is the most interesting unexplored design space and could itself be a follow-up paper.