Contents

Conditionally-Causal Probes: Five Operational Constraints on Linear-Probe Causality in Qwen3.6-27B **2**

An eleven-site empirical map, a unifying operational-constraints framework, and a pre-publication diagnostic battery — derived from four prior honest negatives 2

Abstract 2
1. Introduction 2
2. Related Work 4
3. The Twelve Probes 4
3.1 Subjective-time probes (3 sites — paper-8) 5
3.2 Reasoning-action boundary probes (2 sites — paper-6) 5
3.3 Quality-axis probes (2 sites — paper-5 Phase 10) 5
3.4 Capability-axis probes (4 sites — paper-5 Phase 11) 5
3.5 Predictive probe (1 multi-site entry — paper-3) 6
Table 1 — The Twelve Probes (summary) 6
Table 2 — Five Empirical Regimes 8
4. Five Operational Constraints 8
4.1 Constraint C1 — Layer (spatial) 8
4.2 Constraint C2 — Trajectory (temporal) 9
4.3 Constraint C3 — Magnitude (α) 9
4.4 Constraint C4 — Direction (saturation alignment) 10
4.5 Constraint C5 — Decision locus (architectural) 10
4.6 Regularities across the twelve-site map 11
4.7 Formal connection to causal abstraction theory 11
4.7.1 An interchange-intervention experimental program (proposed future work) 12
5. Pre-Publication Diagnostics 13
5.1 Diagnostic D1 — Random-feature baseline 13
5.2 Diagnostic D2 — Shuffled-source baseline 13
5.3 Diagnostic D3 — Control-token normalization 14
5.4 Diagnostic D4 — Structural-rigidity α-sweep 14
5.5 Diagnostic D5 — Whitespace-stripped flip metric 15
5.6 Diagnostic D6 — Onset-timing sweep (steering-causality only) 15
Table 4 — Diagnostic Coverage Matrix 15
6. Five Case Studies 16
6.1 Case 1 — ST_L31_gen: a trajectory-shaping causal probe (paper-8) 16
6.2 Case 2 — CoT_L55_mid_think: a template-locked structural probe (paper-6) 16
6.3 Case 3 — PSAE v1.5: marginal-fit pathology in sparse top-k prediction (paper-3) 16
6.4 Case 4 — SWE_L43_pre_tool: epiphenomenal-via-softmax-temperature (paper-6) 17
6.5 Case 5 — Cap_L55_pre_tool: saturation-direction direction-flip (paper-5) 17
7. Discussion 17
7.4 Concrete deployment implications (testable predictions) 18
8. Limitations 19
9. Conclusion 19
Acknowledgments 19
Appendix A — Empirical verification trail 20
References 21

Conditionally-Causal Probes: Five Operational Constraints on Linear-Probe Causality in Qwen3.6-27B

An eleven-site empirical map, a unifying operational-constraints framework, and a pre-publication diagnostic battery — derived from four prior honest negatives

Meta-paper bundling and extending papers 3, 5, 6, 7, and 8 of this series. Targeting TMLR (Survey Certification) then ICLR 2027 main. Apache-2.0. Fully reproducible on a single RTX 6000 Blackwell in ∼12 hours.

Abstract

Linear probes on transformer residual streams routinely achieve high predictive AUROC, yet whether a probe direction *also* levers downstream behavior under intervention is rarely measured systematically. We report a twelve-site causal-authority map of probes in Qwen3.6-27B (reasoning-tuned, 27B parameters), comprising eleven probes evaluated under a unified α-sweep + control-token + onset-timing protocol plus one predictive case study, and identify five distinct empirical causal regimes: causal trajectory-shaping, pushup-asymmetric, pushdown-asymmetric, structurally-locked, and epiphenomenal-via-softmax-temperature. We propose that probe causality is **operationally constrained** by a five-axis configuration — layer (spatial), trajectory (temporal), magnitude (α), direction (saturation alignment), and decision locus (architectural) — and demonstrate each constraint with a within-paper falsifying experiment that holds the other four fixed. We then consolidate the methodology that surfaced these constraints into a six-item pre-publication diagnostic battery: random-feature baseline, shuffled-source baseline, control-token normalization, structural-rigidity α-sweep, whitespace-stripped flip metric, and onset-timing sweep. Each diagnostic is mapped to a concrete failure mode we shipped or nearly shipped in our own work: over-parameterization at N<100, marginal-fit pathology in sparse top-k prediction, softmax-temperature artifacts that look causal, amplitude-null masquerading as structural-null, tokenization-inflated flip rates, and trajectory-versus-state confusion. Together the diagnostics cost under one GPU-hour per probe. We release the protocol, capture batches, per-probe verdicts, and an open-source SDK that implements the diagnostics, and argue that the field's growing reliance on probe-based monitoring, reward shaping, and alignment auditing should treat probe causality as a **conditional property to be measured per deployment configuration**, not a global per-probe attribute.

Keywords: linear probes, activation steering, causal interpretability, mechanistic interpretability, Qwen3.6-27B, operational constraints, pre-publication diagnostics, honest negatives, KV-cache, saturation direction.

---

1. Introduction

A linear probe on a transformer's residual stream is cheap to fit, frequently high-AUROC on a wide range of downstream observables, and increasingly load-bearing in the 2025–2026 mechanistic interpretability stack. Probes are being deployed as monitoring classifiers in production safety pipelines (Templeton et al. 2024; OpenAI Alignment 2026), as dense reward signals in reinforcement learning (Marks et al. 2024; Anthropic Persona Vectors 2026), as intervention targets for activation steering (Turner et al. 2024; Belitsky et al. 2026), and as evaluation primitives for chain-of-thought monitorability (Frontier Model Forum 2026; UK AISI Alignment Project 2026). In each of these deployments a single load-bearing assumption sits beneath the application: that a high-AUROC probe direction is *causally* connected to the behavior it predicts — that intervening on the probe direction moves the behavior, and not merely the model's internal accounting of features downstream of the actual decision.

This assumption is empirically false more often than the field acknowledges. The literature now contains a growing list of papers reporting that high-AUROC probes do not lever the behaviors they predict: SAE features that predict harm with high AUROC but fail to suppress it under clamping (Templeton et al. 2024 appendix C); steering directions whose effects vanish once tokenization artifacts are normalized away (Hidden Error

Awareness in CoT, May 2026); KV-cache-locked decisions that no residual-stream intervention can perturb (Belitsky et al. 2026; this paper §4.2); and an entire genre of "predictive SAE" results whose effect collapses under shuffled-source baselines (this paper §3.12, §5.2). The community is converging on a recognition that probes are mostly *correlational*, that *causal* lever-ability is rare and conditional, and that the diagnostics needed to tell which regime a probe is in are not yet standardized.

This paper is our attempt at standardization. Over the past five months we built four papers' worth of probes on a single reasoning-tuned model (Qwen3.6-27B), and in each we discovered that the probe we trained either levered behavior under specific operational conditions or — more often — did not lever behavior at all despite predictive AUROC ≥ 0.8. Each discovery required a specific diagnostic that we initially did not run. Each paper ended up reporting a positive empirical result *plus* a methodological contribution in the form of "the diagnostic that, had we skipped it, would have allowed us to ship a false causal claim." The accumulated diagnostic set is now six items long; the accumulated empirical map covers twelve probes; and the cross-paper pattern points to a tractable framework: **probe causality is operationally constrained.**

We make three contributions.

Contribution 1 (Empirical, §3). A systematic twelve-site map on Qwen3.6-27B — eleven causality-tested probes plus one predictive case study (PSAE) — spanning four probe families, five layers, four generation positions, and five target-behavior classes. The map identifies exactly five distinct empirical causal-class regimes (Table 2): (R1) *causal trajectory-shaping* — directions that lever behavior only when steering is applied continuously from token 1; (R2) *pushup-asymmetric* — lever behavior on one side of α=0 but not the other, in the same direction as behavioral headroom; (R3) *pushdown-asymmetric* — lever in the opposite-direction-from-headroom case; (R4) *structurally-locked* — inert at any α including amplitudes far exceeding $\|\text{residual}\|$; (R5) *epiphenomenal-via-softmax-temperature* — appear causal under naive log-prob shift measurement but collapse to $\Delta\text{rel} \approx 0$ under control-token normalization.

Contribution 2 (Theoretical, §4). A five-axis framework of operational constraints that, taken together, account for the observed regimes: **layer** (spatial), **trajectory** (temporal), **magnitude** (α), **direction** (saturation alignment), and **target class**. Each constraint is established by a falsifying within-paper experiment in which all other constraints are held fixed and the constraint of interest is varied. We argue that this five-axis configuration subsumes the saturation-direction principle of our paper-5 as constraint C4, the trajectory-shaping mechanism of paper-8 as constraint C2, the template-lock mechanism of paper-6 as constraint C5, the amplitude-rigidity diagnostic of paper-6 as constraint C3, and the layer-specificity mechanism of paper-8 Phase 2C as constraint C1.

Contribution 3 (Methodological, §5). A pre-publication diagnostic checklist of five items we propose as mandatory (random-feature baseline, shuffled-source baseline, control-token normalization, structural-rigidity α-sweep, whitespace-stripped flip metric) plus a sixth for steering-causality claims specifically (onset-timing sweep). Each diagnostic is mapped to the specific failure mode it catches (Table 4), and we walk through five case studies (§6) in which we ourselves either shipped or nearly shipped a false causal claim that one of the diagnostics caught. Together the diagnostics cost under one GPU-hour per probe and are implementable as a single helper function (`agent-probe-guard.diagnostics.run_all_probe_checks`).

The framing of this paper is deliberately defensive. We do not propose new positive causal claims here; we re-survey our prior positive claims under a unified protocol and report which surveyed and which collapsed. We believe this is the appropriate stance for the field's current moment: with probes being adopted into production safety pipelines, the cost of a false causal claim is structural and growing. We argue that calibration of *one's own probe portfolio* against the five-axis framework is the cheap, generalizable, publishable contribution that any probe-shipping lab should now make.

The remainder of the paper is organized as follows. §2 situates the work in the 2025–2026 probe-causality literature. §3 catalogs the twelve probes and their empirical verdicts. §4 develops the five-axis operational-constraint framework. §5 presents the diagnostic checklist with cost, coverage, and implementation notes. §6 walks through five case studies. §7 discusses implications for SAE interpretation, steering-as-alignment, and external monitorability bounds. §8 lists limitations. §9 concludes.

---

2. Related Work

Probe causality is being tested more rigorously in 2025-2026. The Persona Vectors paper (Chen, Arditi, Sleight, Evans, and Lindsey 2025, arXiv:2507.21509) introduced a clean per-axis causal-validation protocol on Claude-family persona directions; this paper adopts a similar protocol spirit and extends it to four additional probe families on Qwen3.6-27B. SAEBench (Karvonen et al. 2025, arXiv:2503.09532) reports inconsistency between proxy metrics and practical performance across 200+ SAEs and 8 architectures — a finding compatible with our five-regime taxonomy, which would classify many SAE feature directions exhibiting "high metric, low clamp effect" as R4 (structurally-locked) or R5 (epiphenomenal). The broader 2025 SAE skepticism literature is converging on a recognition that SAE feature predictiveness and SAE feature causality must be measured separately.

Activation steering as alignment tool — cache-state interventions are emerging as a complement. KV Cache Steering (Belitsky, Kopiczko, Dorkenwald, Mirza, Glass, Snoek, Asano 2025, arXiv:2507.08799) demonstrates that one-shot interventions applied directly to the KV cache can induce chain-of-thought reasoning in small frozen models, achieving behavioral shifts without weight updates or prompt modifications. We read this as empirical evidence that cache state is itself a load-bearing locus of behavioral control — directly relevant to our C2 (trajectory) constraint, which holds that residual-stream interventions late in generation are defeated by accumulated cache state. The two findings are complementary: Belitsky et al. show cache state can be steered *as an alternative* to residual interventions; our C2 explains *why* residual interventions become less effective as cache state accumulates.

Causal abstraction and interchange interventions (Geiger, Lu, Icard, Potts 2021; Geiger et al. 2023 *Causal Abstraction* arXiv:2301.04709; Geiger et al. 2025 JMLR *Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability*) provide the theoretical foundation for asking when a probe direction stands in causal abstraction-relation to a downstream behavioral variable. The interchange-intervention protocol operationalizes this. Our five-axis framework can be read as an empirical taxonomy of the regimes in which a probe-direction intervention does or does not preserve the abstract-level causal relation (§4.7).

Causal abstraction and interchange interventions (Geiger et al. 2023; 2025; 2026) provide the theoretical foundation for asking when a probe direction stands in causal abstraction-relation to a downstream behavioral variable. The framework distinguishes interventions that preserve the abstract-level causal structure from those that merely shift distributional accounting. Our five-axis framework can be read as an empirical taxonomy of the regimes in which a probe-direction intervention does or does not preserve the abstract-level causal relation; constraint C5 in particular corresponds to the abstract-level decision being made by *input-token* variables rather than residual-stream variables, in which case no residual intervention is an interchange intervention.

Mechanistic interpretability methodology is increasingly emphasizing the need to distinguish correlative from causal probe findings. Our own contributions to this conversation are paper-3 (marginal-fit pathology), paper-5 (saturation-direction principle), paper-6 (two-forms epiphenomenality), paper-7 (NLA decoupling), and paper-8 (trajectory-shaping). This paper consolidates the methodology of those five into a single reusable checklist and proposes the five-axis framework as the unifying theoretical structure.

---

3. The Twelve Probes

We map twelve probes across four families (Table 1). Each probe is uniquely specified by a (model, layer, position, probe-type) tuple; we run all twelve under a unified evaluation protocol (Appendix A) and assign each to one of the five empirical regimes R1–R5 (Table 2). All probes are trained and evaluated on Qwen3.6-27B (reasoning-tuned 27B, hybrid Gated-Delta-Net + standard attention, open-weights). Cross-model corroboration from Qwen2.5-7B and Gemma-3-12B (paper-7) is reported but not counted toward the twelve.

3.1 Subjective-time probes (3 sites — paper-8)

A Ridge regression probe is trained to predict the fraction of the thinking-phase already elapsed (a continuous quantity in [0, 1]) from residual activations at one of three layers. Training data: 50 prompts $\times \leq 3$ samples $\times$ 100-200 thinking-tokens each, with target = position / total_thinking_length. Probes achieve R^2=0.84 (L11), R^2=0.86 (L31), R^2=0.82 (L55), with Spearman $\rho \geq 0.90$ at all three layers. Three baselines (random-feature, shuffled-target, constant-mean) cleanly fail ($R^2 \leq 0.07$). Under forward-hook steering with $\alpha \in$ {$±$5, $±$25, $±$50, $±$100, $±$200} from token 1:

- **ST_L31_gen** (#1): Causal at α=+50; shortens GSM8K thinking-length in 9/14 prompts vs. 2/14 for matched random (Fisher p=0.0092, OR=10.8×). Cross-domain: 19/20 (95%) probe-clean-termination on SWE-bench Verified vs. 6/20 (30%) random. **Regime R1 (causal trajectory-shaping).**
- **ST_L11_gen** (#2): Mostly inert. At α=+25, 1/10 prompts collapse to ragged termination; at all other α including α=+500, no behavioral change. Despite R^2=0.84 — equal to L31. **Regime R4 (structurally-locked).**
- **ST_L55_gen** (#3): Fully inert at α up to $\$\pm 500. Despite R\hat{}2\$$=0.82. **Regime R4 (structurally-locked).**

The L11/L55 inertness directly establishes Constraint C1 (layer): same probe family, same R^2, zero lever at the wrong layer.

3.2 Reasoning-action boundary probes (2 sites — paper-6)

- **SWE_L43_pre_tool** (#4): Trained on Qwen3.6-27B agentic rollouts (N=99) to predict capability deployment at the pre-tool position. AUROC=0.83. Under α=+2 steering, naive log-prob shift of the "finish" token = +0.479 — apparent strong causality. Under control-token normalization (Δrel = Δ("finish") − mean(Δ(control_tokens)) over controls {`search`, `execute`, `write`, `read`, `wait`}) Δrel collapses to −0.046 on failures (+0.075 on successes); at α=+5 the continuous behavioral test produces 3/4 same tool selection and the single-shot test gives 4/4 same tool. **Regime R5 (epiphenomenal-via-softmax-temperature).**
- **CoT_L55_mid_think** (#5): CoT-Integrity probe trained at L55 mid-thinking position. Original report AUROC=0.91 at K=50, N=240 — later flagged as over-parameterized relative to N when a random-K-matched supplement was run (Diagnostic D1). Bidirectional α-sweep up to α=$\$\pm 200 (> |residual|\$)$ and α=+500 produces *zero* behavioral change for probe direction AND random direction. Diagnosis: `enable_thinking=False` chat template injects `<think></think>` token pair into the input itself; the decision to emit thinking is encoded *in input tokens before the residual stream forms*. **Regime R4 (structurally-locked, mechanism = template-lock).**

3.3 Quality-axis probes (2 sites — paper-5 Phase 10)

- **FG_L31_pre_tool** (#6): Fabrication-detection probe at L31 pre-tool position. Categorical decision class. α-sweep: probe-direction flips match random-direction flips at every α tested. **Regime R5 (epiphenomenal, indistinguishable from random-direction destruction).**
- **RG_L55_mid_think** (#7): Reasoning-quality probe at L55 mid-thinking position. Continuous quality class. α=+200 produces whitespace-stripped flip rate 32% (16/50) vs. random 2% (1/50). Sharp threshold: 0% flips until α=+100. Asymmetric: pushdown side (α=−200) produces 0% flips. **Regime R2 (pushup-asymmetric).**

3.4 Capability-axis probes (4 sites — paper-5 Phase 11)

All four trained on the SWE-bench Pro N=99 patch-success vs. patch-fail dataset (Phase 6 verdict). All four exhibit pushdown-asymmetric behavior on the original distribution (HumanEval + MBPP, Qwen pass-rate ∼89%). Two flip behavior on cross-distribution validation (BigCodeBench at pass-rate ∼55%, Codeforces rating $\geq$ 2000 at pass-rate ∼7%).

- **Cap_L31_pre_tool** (#8): $\alpha=-100$ pushdown gap +40pp on HumanEval+MBPP, +37pp on Codeforces. **Regime R3 (pushdown-asymmetric, saturation-independent).**
- **Cap_L23_pre_tool** (#9): $\alpha=-100$ pushdown gap +40pp on HumanEval+MBPP (Phase 11b: probe flip 1.000 vs random 0.600 over N=30), +43pp on Codeforces (Phase 11e). **Regime R3 (pushdown-asymmetric, saturation-independent).**
- **Cap_L43_turn_end** (#10): Peak gap +60pp at $\alpha=-200$ on HumanEval+MBPP (Phase 11b — strongest single gap in the portfolio). Cross-distribution comparison at matched $\alpha=-100$: +40pp on HumanEval+MBPP collapses to +7pp on Codeforces (Phase 11e). **Regime R3, saturation-magnitude-coupled** — gap magnitude tracks baseline saturation, but lever direction is preserved.
- **Cap_L55_pre_tool** (#11): $\alpha=-100$ pushdown gap +34pp on HumanEval+MBPP, +7pp pushup at $\alpha=+200$ on the same distribution (Phase 11). On Codeforces *flips lever direction entirely*: $\alpha=-100$ produces −3pp pushdown (null) while $\alpha=+200$ produces +40pp pushup (Phase 11e). **Regime R3 $\leftrightarrow$ R2 saturation-direction-coupled.** Strongest single piece of evidence for Constraint C4 (saturation direction): same probe, same layer, same direction; baseline saturation flipped between distributions, and lever direction flipped with it.

3.5 Predictive probe (1 multi-site entry — paper-3)

- **PSAE_v1.5** (#12): Predictive SAE-feature probes trained at twelve (layer, source-fraction) sites on Qwen3.6-27B to predict end-of-thinking sparse-autoencoder feature firings from earlier-thinking residuals (N_train=106, N_test=27, d_sae=65536, k=128). Naive recall@1024 = 0.852–0.869 at L11, 0.788–0.835 at L31, 0.671–0.724 at L55 (verified against `predictive_sae_v15_results.json`). **Shuffled-source baseline** (X_train shuffled, y_train kept, identical recipe) reproduces real recall at all twelve sites with max $|\Delta|$ = 0.027 (mean $|\Delta| < 0.015$), well within the $±$0.03 noise band. A trivial constant baseline that predicts the top-M most-globally-common features ignoring input strictly exceeds the trained probe at L11/L31. **Not a steering probe** — counted as a single methodology entry. The site is the motivating empirical case for Diagnostic D2 (shuffled-source baseline).

Table 1 — The Twelve Probes (summary)

#	Probe	Layer	Position	Behavior class	R^2/AUROC	Regime
1	Subjective-time	L31	free-gen	Continuous (time fraction)	R^2=0.86	R1 causal trajectory-shaping
2	Subjective-time	L11	free-gen	Continuous	R^2=0.84	R4 structurally-locked
3	Subjective-time	L55	free-gen	Continuous	R^2=0.82	R4 structurally-locked
4	SWE pre-tool	L43	pre_tool	Categorical (capability)	AUROC=0.83 (Phase 4 N=54; gap shrinks to +0.042 at N=99)	R5 epiphenomenal-softmax-temp
5	CoT-Integrity	L55	mid_think	Structural (emission)	AUROC=0.91 (K=50, over-parameterized at N=240)	R4 structurally-locked (template)

Figure 1 — The eleven causality-tested probes in Qwen3.6-27B

Figure 1: Figure 1 — The eleven causality-tested probes in Qwen3.6-27B

Figure 2 — α-sweep behavioral gap from random-direction control

Figure 2: Figure 2 — α-sweep behavioral gap from random-direction control

#	Probe	Layer	Position	Behavior class	R^2/AUROC	Regime
6	FG fabrication	L31	pre_tool	Categorical (fabrication)	—	R5 epiphenomenal
7	RG reasoning-quality	L55	mid_think	Continuous (quality)	—	R2 pushup-asymmetric
8	Capability	L31	pre_tool	Categorical (success)	—	R3 pushdown-asymm., sat-independent
9	Capability	L23	pre_tool	Categorical (success)	—	R3 pushdown-asymm., sat-independent
10	Capability	L43	turn_end	Categorical (success)	—	R3, saturation-coupled
11	Capability	L55	pre_tool	Categorical (success)	—	R3 $\leftrightarrow$ R2, direction-flip
12*	PSAE-v1.5	L11/L31/L55	thinking-fraction	Sparse top-k prediction	recall@1024 = 0.83-0.87	Outside R1-R5 — predictive case study, motivates D2

* PSAE-v1.5 is a predictive probe, not a causality-tested steering probe; it is listed for completeness as the empirical case motivating diagnostic D2 but is not assigned a causal regime. The eleven causality-tested probes are #1-#11.

Figure 1. Each of the eleven causality-tested probes plotted at its (layer, generation-position) coordinate, colored by empirical regime. The three subjective-time free-generation probes at L11/L31/L55 (top row) illustrate Constraint C1 (layer): same family, R^2={0.84, 0.86, 0.82}, but only L31 falls into R1 — L11 and L55 are structurally-locked despite equal predictive accuracy. The four capability probes at the pre_tool / turn_end positions cluster in R3 (pushdown-asymmetric). The two epiphenomenal cases (SWE_L43, FG_L31) are both at categorical decision positions. (PSAE v1.5 is omitted from this figure because it is a predictive probe outside the causality-tested set.)

Figure 2. Behavioral lever strength as a function of steering α, measured as percentage-point gap from a matched random-direction control. Rows = probes (color strip = regime); columns = α tested. White cells indicate measured-and-inert; numeric cells indicate gap magnitude in pp. Patterns visible: (i) R4 rows (ST_L55, CoT_L55, FG_L31, SWE_L43) are uniformly empty across the entire sweep, distinguishing them

from amplitude-null cases by structural-rigidity diagnostic D4; (ii) R3 rows (capability probes) all peak at α=-100 to -200 with no measurable pushup lever; (iii) the only R1 case (ST_L31) peaks at the moderate α=+50 sweet spot and degrades at α=+200; (iv) Cap_L55_pre_tool uniquely shows lever activity on *both* sides of α=0 within the same distribution (pushdown at α=-100 and small pushup at α=+200 on HE+MBPP), and *flips lever direction entirely* under cross-distribution validation against Codeforces (where baseline saturation reverses) — the direction-flip evidence for Constraint C4. Values are based on reported measurements from papers 5/6/8 Phase 2A/2C/10/11/11e on HumanEval+MBPP baseline; unmeasured cells default to zero.

Table 2 — Five Empirical Regimes

Regime	Definition	Probes
R1 — Causal trajectory-shaping	Lever behavior under steering from token 1; effect collapses under delayed onset	#1
R2 — Pushup-asymmetric	Lever behavior in one direction (toward behavioral headroom) only	#7
R3 — Pushdown-asymmetric	Lever behavior in the opposite direction (away from saturation) only	#8, #9, #10, #11
R4 — Structurally-locked	Inert at any α including amplitudes > $\|$residual$\|$; decision encoded in non-residual variables	#2, #3, #5
R5 — Epiphenomenal-via-softmax-temperature	Apparent causality under naive log-prob shift; Δrel $\approx$ 0 under control-token normalization	#4, #6

---

4. Five Operational Constraints

We argue that the five empirical regimes R1-R5 are surface manifestations of a single underlying framework: probe-direction causality is **operationally constrained** by a five-axis configuration. A probe direction is causally load-bearing only when **all five constraints are satisfied simultaneously**. The map in §3 surveys probes that satisfy different subsets, and the distinct regimes correspond to which constraints are *violated* in each. Each constraint is established by a *falsifying* within-paper experiment in which the constraint of interest is varied while the others are held fixed.

4.1 Constraint C1 — Layer (spatial)

Statement. The same probe family at the same R^2/AUROC level is causally load-bearing at one layer and inert at others. Layer-causality cannot be predicted from layer-AUROC.

Falsifying experiment. Paper-8 Phase 2C trains three Ridge probes for the same subjective-time target at L11, L31, L55 with the same training data, same recipe, same baselines passing equivalently. The three probes achieve R^2={0.84, 0.86, 0.82} — within 0.04 of each other. Under identical α-sweep steering protocols, the three behave radically differently: ST_L31_gen levers 9/14 (Fisher p=0.0092) while ST_L11_gen produces 1/10 ragged collapses at α=+25 and ST_L55_gen produces zero behavioral change at any α up to α=+500. Same target, same recipe, comparable R^2: only the layer differs. **C1 holds.**

Mechanism (conjectured). The Qwen3.6-27B residual stream encodes different abstract-level variables at different depths. The subjective-time target ("thinking-fraction-elapsed") is a *control variable* used by L31 machinery to decide whether to terminate thinking; at L11 the same direction is a *predictive readout* of features that L11 attends to for its own purposes but does not gate behavior on; at L55 the same direction is a

post-hoc summary generated *after* the termination decision has already been made. The Ridge probe captures the predictive direction at all three depths, but only L31 has the additional property of *gating* downstream processing.

Implication for practitioners. A probe paper that demonstrates predictive AUROC at one layer makes no claim about causal authority. Layer-causal authority must be tested per layer; cross-layer extrapolation is unsupported.

4.2 Constraint C2 — Trajectory (temporal)

Statement. For probes that lever continuous-trajectory behaviors (thinking-length, reasoning-quality, ongoing-decision), the steering intervention must be applied **continuously from token 1**. Delayed-onset steering, even when applied with identical α and identical direction, collapses in effectiveness as a function of how many decoded tokens preceded the onset.

Falsifying experiment. Paper-8 Phase 2B holds (probe, layer, α, direction, prompt) fixed and varies only the onset token. From-token-1 steering produces 9/10 termination on GSM8K. Onset at decode-step 50 drops effectiveness to 3/10. Onset at decode-step 200 drops effectiveness to 0/10. Two closed-loop variants — Design E (probe-as-sensor with threshold trigger at predicted-fraction = 0.85) and Design F (plateau-detector with rolling slope < 0.01) — achieve 1-2/10 termination, confirming that the "termination basin" is not a state-attractor that re-engages when re-entered but is mediated through **KV-cache state buildup** that becomes increasingly rigid as decoding progresses. **C2 holds.**

Mechanism (conjectured). The Qwen3.6-27B reasoning model accumulates intermediate state in its KV cache during thinking. Once a sufficient number of thinking-tokens have been written into the cache, the cache itself encodes a "continue-thinking" prior that is resistant to instantaneous residual perturbation. The probe direction at L31 levers behavior only when it can act *before* the cache hardens — i.e. when steering is applied from token 1 and *shapes the trajectory* rather than *intervenes on the state*.

Implication for practitioners. "Adaptive steering" intuitions — wait until the model needs help, then apply the probe — are mechanistically backwards for trajectory-shaped behaviors. Probes are preventive enforcers, not adaptive detectors.

4.3 Constraint C3 — Magnitude (α)

Statement. Probe-direction causality is observed only inside a narrow α-sweet-spot. Below the sweet spot the intervention is amplitude-null (no effect, indistinguishable from baseline). Above the sweet spot the intervention is *degenerate* — it produces flips that are indistinguishable from random-direction destruction. The sweet spot is **probe-specific** and must be empirically located.

Falsifying experiment. Paper-6 Phase 8 tested CoT_L55_mid_think at $\alpha \in \{\$\pm\$5, \$\pm\$25, \$\pm\$50, \$\pm\$100, \$\pm\$200\}$ and observed zero behavioral change for both probe and random directions across the full range. Extension to $\alpha = +500$ (more than $4\times$ the L2 norm of typical L55 residuals on this distribution) still produced zero change — establishing the *structural-rigidity* signature that distinguishes R4 (template-locked) from amplitude-null. Conversely paper-5 Phase 10 found that RG_L55_mid_think produces 0% flips at $\alpha \leq +100$ and 32% flips at $\alpha = +200$, while paper-8 Phase 2A found that ST_L31_gen produces ragged degeneration at $\alpha = +200$ but a clean 9/14 lever at $\alpha = +50$. **Each probe has a different sweet spot. C3 holds.**

Mechanism (conjectured). A probe direction is a unit-length vector in residual space; α scales how much of that direction is added to (or subtracted from) the residual. At low α, the perturbation is absorbed by downstream layers without behavioral consequence (amplitude-null). At high α, the perturbation exceeds the magnitude of typical residuals at that layer, pushing the activation into out-of-distribution territory where the intervention degrades into structural destruction (the random-direction control flips at similar rates). The narrow window in between is where the direction's *semantic specificity* — its alignment with a meaningful behavioral axis — can manifest as behavioral lever without OOD destruction.

Implication for practitioners. The default $\alpha \in \{\$\pm\$2, \$\pm\$5\}$ of much of the 2024-2025 steering literature is biased toward false negatives. A paper claiming "the probe is inert" should report a structural-rigidity

α-sweep (Diagnostic D4) to distinguish amplitude-null from structural-null.

4.4 Constraint C4 — Direction (saturation alignment)

Statement. For probes that *do* lever behavior at the C3 sweet spot, the lever is asymmetric: the probe direction has causal authority along the axis where the baseline residual has behavioral *headroom* to flip — i.e. the direction *opposite* the saturation. The random-direction control flips generations only via OOD destruction; the probe direction additionally flips via OOD-semantic perturbation in the saturated subspace.

Falsifying experiment. Paper-5 originally predicted (Phase 11) that capability probes would lever pushup (toward stronger capability) at $\alpha{=}+200$, because capability is a "continuous gradient" behavior. The prediction was *falsified*: capability probes lever pushdown (toward weaker capability) at $\alpha{=}-100$. The falsifier motivated the saturation-direction refinement: on HumanEval+MBPP (Qwen pass-rate $\sim$89%), the baseline residual is already saturated *toward* successful capability deployment; behavioral headroom is *away from* saturation; the probe direction levers along the headroom axis, which here points pushdown. Paper-5 Phase 11e then **directly tested** the prediction on Codeforces rating ≥ 2000 (Qwen pass-rate $\sim$7%, baseline residual saturated *away from* capability) and observed Cap_L55_pre_tool flip lever direction: pushdown gap -3pp, pushup gap $+40$pp at $\alpha{=}+200$. **Same probe, same layer, same direction. Saturation flipped. Lever flipped. C4 holds.**

Mechanism (conjectured). The residual stream at a given layer occupies a region of activation space whose geometry is shaped by the model's behavioral distribution at that layer. For a behavior near saturation, the residual is bunched against the saturating manifold; a perturbation toward the saturation axis is geometrically blocked (the residual cannot move "further saturated") while a perturbation away from it is geometrically free. The probe direction encodes a semantically meaningful axis; aligning α-sign with the headroom direction of the baseline distribution is what gives the probe its measured lever asymmetry.

Implication for practitioners. Predicting lever direction from probe semantics alone (e.g. "the capability probe should lever pushup") will mispredict. The prediction must combine probe semantics with baseline saturation of the test distribution. Cross-distribution validation is the falsification protocol.

4.5 Constraint C5 — Decision locus (architectural)

Statement. Probe-direction causality is mediated by the *architectural locus* at which the target-behavior decision is made. Three loci have been observed: (i) *residual-stream-continuous* — the target is a continuous quantity computed in the residual stream and gated by downstream layers; lever-able under C1-C4. (ii) *residual-stream-categorical* — the target is a discrete decision computed in the residual stream; lever-able only when the decision axis has headroom (C4 active), often inert otherwise. (iii) *input-token-structural* — the target decision is encoded in input tokens *before* the residual stream forms; **no residual-stream intervention can lever it** regardless of α.

Falsifying experiment. Paper-6 Phase 8 demonstrates the input-token- structural case sharply. The Qwen3.6-27B `enable_thinking=False` chat template injects a `<think></think>` token pair into the input. The decision "should this generation include thinking?" is made by *the template variable*, not by any residual-stream computation. The L55 CoT-Integrity probe predicts thinking-emission with AUROC$=$0.91 because the residual stream *encodes* the decision after it has been made — but intervening on the residual at L55 cannot un-make a decision that the template already committed to. C5 holds *by construction* of the architecture, and is what distinguishes R4-template-locked from R4-amplitude- locked.

Mechanism (architectural). The chat-template mechanism is widespread in instruction-tuned reasoning models. Any decision that is gated by template variables (enable_thinking, system-prompt tokens, special control tokens) is structurally outside the scope of residual-stream steering. Probe AUROC on such decisions reflects *readout* of the post-template state, not *causal mediation* of the decision itself.

Implication for practitioners. Distinguishing R4-template-locked from R4-amplitude-locked requires the structural-rigidity α-sweep (D4) up to $\alpha > 4 \times \|\text{residual}\|$. A probe that produces zero change at $\alpha{=}\$\pm\$200$

is in R4; the diagnosis between template-lock and amplitude-null requires inspecting the architecture to identify whether a template variable upstream of the residual stream encodes the decision.

4.6 Regularities across the twelve-site map

The five constraints C1-C5 are conjectured to be individually necessary; joint sufficiency is a stronger claim we cannot establish with N=11 causality-tested probes (N=1 in R1, N=1 in R2). We report the regularities we observe and frame them as falsifiable hypotheses, not theorems:

- **(C1 + C2 + C3 + C4 satisfied, C5 = continuous-residual)** is consistent with R1 in the one case observed (ST_L31_gen).
- **(C1 + C3 + C4 satisfied, C5 = continuous-residual)** is consistent with R2 in the one case observed (RG_L55_mid_think).
- **(C1 + C3 + C4 satisfied, C5 = categorical-residual, adversarial baseline saturation)** is consistent with R3 in four cases (capability probes at L23/L31/L43/L55).
- **C5 = input-token-structural** entails R4-template-locked regardless of other constraints; demonstrated sharply by CoT_L55_mid_think.
- **(C1-C4 satisfied)** $\wedge$ Δ**rel** $\approx$ **0 under control-token norm** entails R5; two cases observed (SWE_L43_pre_tool, FG_L31_pre_tool).

The framework's predictive content is asymmetric. The negative direction is sharper: a probe that violates C1, C3, or C5 is predicted to fall into R4. The positive direction is weaker: we cannot yet predict which of R1, R2, R3 a probe satisfying C1-C4 will exhibit without additional information about its decision-locus and baseline saturation. The framework is open to two classes of falsification: a probe that satisfies all five constraints and remains inert (would imply additional unnamed constraints), or a probe that violates one of C1-C5 and levers robustly (would imply the violated constraint is not actually necessary). We have not encountered either in the twelve-site map or in cross-model corroboration (paper-7), but the sample is small enough that both remain live possibilities.

4.7 Formal connection to causal abstraction theory

The five-axis framework of §4.1–§4.6 is empirical, but its conceptual structure aligns with the causal abstraction framework developed by Geiger and collaborators (2021; 2023; 2025). In that framework, two variables stand in *causal abstraction* relation when intervening on the low-level variable produces the same effect on downstream variables as intervening on the high-level variable that abstracts it. The interchange-intervention protocol operationalizes this: a causal-abstraction claim holds iff residual-stream interventions on the candidate-abstracting direction reproduce the behavioral effects of interchange interventions on the abstracted high-level variable.

Each of our five constraints can be read as a *precondition* on interchange-intervention validity:

- **C1 (Layer)** — The candidate-abstracting direction must be located at the depth where the abstraction relation actually holds. A direction that abstracts a behavior at L31 is *not* the same direction at L11 or L55, even if Ridge regression on residuals at all three depths produces equally-predictive readouts. The L11/L55 readouts of subjective-time predict the behavior without abstracting it.
- **C2 (Trajectory)** — Interchange intervention preserves the abstraction relation only when the cache state at intervention time matches the cache state under which the abstraction was established. KV-cache lock-in is the empirical realization: interventions late in generation occur under a cache state that has drifted from the training distribution, breaking the abstraction.
- **C3 (Magnitude)** — The α parameter scales the perturbation magnitude. Below the C3 sweet spot, the perturbation is too small to propagate through downstream layers and affect the high-level variable; above it, the perturbation pushes the residual outside the abstraction's validity region, where intervention degrades into out-of-distribution destruction. The sweet spot is the operational window within which the abstraction relation holds.

- **C4 (Direction)** — The saturation-direction principle implies that the abstraction relation has *asymmetric* validity: the direction abstracts the high-level variable along the axis where baseline-distribution geometry leaves headroom. The direction-flip evidence from Cap_L55_pre_tool (§3.4) demonstrates this empirically: the same probe direction abstracts opposite behavioral variables under opposite baseline saturations.
- **C5 (Decision locus)** — The causal abstraction relation requires the high-level variable to be *computed* by the low-level mechanism. When the high-level variable is set by a non-residual mechanism (chat-template tokens, system-prompt control tokens), no residual intervention preserves the abstraction, regardless of how predictive the residual readout is. The template-locked decision is an abstraction *of input tokens*, not of residual computations.

This reading suggests that the field's growing collection of "linear probe levers behavior X" claims should be understood as *claims of partial causal abstraction* — abstraction that holds under specific operational conditions. The five constraints constitute the operational specification of those conditions. Stating a probe-causality claim without specifying the (C1, C2, C3, C4, C5) configuration under which it was tested is analogous to stating a regression coefficient without its covariates: technically true under a particular set of conditions, but not transportable.

The diagnostics of §5 then play distinct roles within the framework. D1 and D2 establish that the probe direction encodes signal in the X→Y direction (not dimensionality artifact or marginal-fit). D3 distinguishes signal-bearing interventions from softmax-temperature artifacts that do not preserve the abstraction. D4 distinguishes amplitude-null abstraction failures (C3 violation) from structural-null architectural failures (C5 violation). D5 ensures the behavioral readout used to validate the abstraction is itself robust. D6 establishes that the abstraction holds under the operational protocol intended for deployment, not merely under the protocol used for training.

We do not claim to have *proven* the causal-abstraction relation for any of our R1–R3 probes; the framework only specifies necessary conditions and full proof would require interchange-intervention protocols across counterfactual prompt pairs that we have not used. We claim only that the five-axis configuration is the empirical specification of when an attempted interchange intervention will succeed — and that the diagnostics provide the falsification battery for the corresponding negative claims.

4.7.1 An interchange-intervention experimental program (proposed future work)

The conjectures of §4.7 are testable. We sketch an experimental program that would either validate or falsify the causal-abstraction reading of each constraint, deliberately designed to be executable on the same infrastructure (single RTX 6000 Blackwell, ∼36 GPU-hours total) as the twelve-site map itself:

1. **Counterfactual prompt-pair generation.** For each R1/R2/R3 probe, construct paired prompts (P+, P−) that differ only in the high-level variable abstracted by the probe (e.g. for ST_L31_gen, pair a prompt that empirically produces short thinking with one that produces long thinking; for capability probes, pair a high-difficulty test with a low-difficulty one matched on token length). Target N = 50 pairs per probe.
2. **Interchange-intervention protocol.** For each pair (P+, P−), run inference on P+ while replacing the residual at (layer, position) with the residual from P− projected along the probe direction. Measure whether the behavioral outcome of P+ shifts toward the P− outcome. This operationalizes the causal-abstraction test.
3. **Per-constraint factorial design.** Vary each of C1-C5 independently: for C1, run the protocol at wrong-layer (L11 instead of L31 for ST); for C2, run with intervention applied at decode step 50 instead of token 1; for C3, vary α; for C4, run on a distribution with flipped baseline saturation; for C5, attempt the protocol on a template-locked probe (CoT_L55) as a negative-control.
4. **Falsification criteria.** The conjectures of §4.7 are falsified if
   (a) a probe satisfying all five constraints fails the interchange test, or (b) a probe failing one of C1-C5 passes the interchange test robustly. Either outcome would refine the framework. We have not

Figure 3 — Constraint-violation decision flowchart

Figure 3: Figure 3 — Constraint-violation decision flowchart

executed this program in v1 of this paper and present it as the natural next step.

This program is, to our knowledge, the smallest experimental design that could establish full causal-abstraction status for our R1-R3 cases. We flag it explicitly to acknowledge that the present paper's framework is empirically motivated but theoretically incomplete in the causal-abstraction-theoretic sense, and that the gap is closable with concrete bounded compute.

Figure 3. Diagnostic order for assigning a candidate probe to one of the five empirical regimes. The flowchart operationalizes the constraint framework: violations are checked in order of cost (cheapest diagnostic first), and each violation routes the probe to its corresponding regime. A probe that passes all six diagnostics and additionally satisfies the C4 saturation-direction asymmetry test is in R1 if its lever is also trajectory-dependent (C2) or in R2/R3 otherwise. The flowchart does not distinguish R2 from R3 internally — that distinction requires cross-distribution validation against baseline saturation (Constraint C4), which is not a single diagnostic but a multi-distribution protocol.

---

5. Pre-Publication Diagnostics

The five-axis framework of §4 is a *post-hoc* characterization. To make it *pre-publication useful* we need diagnostics that surface the regime of a given probe *before* the probe is shipped. We describe five diagnostics we propose as pre-publication-mandatory diagnostics (D1-D5) plus one specific to steering-causality claims (D6). All six together cost under one GPU-hour per probe and catch concrete shipping failures we have encountered in our own work (Table 4).

5.1 Diagnostic D1 — Random-feature baseline

Procedure. Train an identical-recipe probe on K random features (unit-norm Gaussians) instead of meaningful features. Sweep K to match the probe's effective dimensionality (PCA components, top-K diff-means, etc.). Report the random-baseline performance alongside the real-probe performance.

Failure mode caught. Over-parameterization at $N < 100$. Probes with K_train $>$ N can shatter random labels and report apparent high AUROC that is dimensionality-artifact, not signal.

Concrete case. Paper-6 Phase 5d originally reported AUROC=1.000 at L43 think_start K=50 on N=17. Without D1, this would have been shipped as a positive result. Adding D1 in Phase 6c revealed random-feature baseline also 1.000 at K=50, dropping to 0.495 at K=10. Real signal at L43 pre_tool PCA-10 was AUROC=0.764 (gap over random +0.269); the original K=50 result was 100% over-parameterization.

Cost. ∼10 GPU-minutes for typical Ridge/LogisticRegression probes.

Recommended at. $N < 100$, K_train > 10, any "novel high-AUROC site" claim.

5.2 Diagnostic D2 — Shuffled-source baseline

Procedure. Train the probe with X_train shuffled across the training set (preserving the marginal of X) but y_train kept in original order. Identical recipe, identical hyperparameters. Report the shuffled-source performance alongside the real probe.

Failure mode caught. Marginal-fit pathology in sparse top-k prediction. When the target is a sparse k-hot vector with a concentrated marginal distribution, the probe can learn the marginal-predictive rule (predict the top-M most-globally-common features ignoring input) and achieve high recall@K without any per-prompt predictive structure.

Concrete case. Paper-3 (PSAE v1.5) reported recall@1024 = 0.83-0.87 across twelve (layer × source-fraction) sites. Without D2, this would have been shipped as "predictive SAE features within reasoning." Adding D2 revealed shuffled-source baseline reproduces real recall within $±$0.03 at all twelve sites (Cohen's d < 0.15). A trivial constant baseline that ignores the input strictly *exceeds* the trained probe at L11/L31. The paper reframed as honest-negative methodology contribution.

Cost. ∼10 GPU-minutes (one additional training run).

Caveat on generalization. This v1 motivates D2 with one concrete case (PSAE v1.5). The pathology's five structural conditions (sparse top-k target + concentrated marginal + N_train « d_target + lazy loss + recall-style metric) jointly predict similar collapses in MoE-routing prediction probes, top-k vocab-prediction probes, and any "predict end-of-trajectory feature firing from earlier residual" setup, but we have not yet executed D2 on additional such probes outside our own portfolio. We propose D2 as mandatory for any work in this class on the basis that the diagnostic cost (∼10 min) is dominated by the cost of a single false positive shipped as a positive result.

Recommended at. Sparse top-k targets (SAE features, MoE routing, top-k vocab), N_train < 200, recall-style metrics.

5.3 Diagnostic D3 — Control-token normalization

Procedure. For probe-direction causality claims based on log-prob shift of a target token, report Δrel = Δ(target) − mean(Δ(control_tokens)) rather than raw Δ(target). Control tokens should be a matched set of high-frequency tokens *not* semantically related to the target (e.g. for target = "finish", controls = {"the", "a", "is", "of", "and"}).

Failure mode caught. Softmax-temperature artifacts that uniformly shift log-probabilities across the vocabulary and look like specific causal authority on the target token.

Concrete case. Paper-6 Phase 7 originally measured Δ(log p("finish")) = +0.479 under α=+5 steering of the L43 pre_tool probe direction. Without D3, this would have been shipped as "L43 pre_tool levers tool-finish decisions." Adding D3 revealed mean(Δ(controls)) = +0.525, giving Δrel = −0.046. The probe was epiphenomenal; the apparent causality was uniform softmax-temperature shift.

Cost. < 1 GPU-minute (one additional log-prob evaluation per steered position).

Recommended at. Any log-prob-shift causality claim, any "the probe levers token X" framing.

5.4 Diagnostic D4 — Structural-rigidity α-sweep

Procedure. When forward-hook steering at $\alpha \in$ {$±$2, $±$5} produces zero behavioral change, *do not* declare the probe inert. Sweep α to multiples of the L2-norm of typical residuals at that layer: $\alpha \in$ {$±$50, $±$100, $±$200} and beyond if needed. Compare probe-direction behavior to random-direction behavior at each α. The structural-rigidity signature is *zero change for both directions at $\alpha > 4 \times \|residual\|$* — distinguishing template-lock or input-token-structural causality from amplitude-null inert probes.

Failure mode caught. Amplitude-null masquerading as "the probe is inert." Conversely, *template-locked* probes (e.g. CoT_L55_mid_think) can be definitively distinguished from "we just haven't pushed α high enough" inertness by demonstrating zero change up to $\alpha > 4 \times \|$residual$\|$.

Concrete case. Paper-6 Phase 8 originally observed zero behavioral change at $\alpha \in$ {$±$2, $±$5, $±$25}. Without D4, the conclusion would have been "L55 CoT-Integrity probe is inert at α=$±$5, may lever at higher α." Adding D4 (extension to α=$±$200, then α=+500) revealed zero change across the entire range — and inspection of the Qwen3.6-27B chat template identified the input-token-structural mechanism (C5). The diagnosis became template-locked, not amplitude-null.

Cost. < 1 GPU-minute (extending α range).

Recommended at. Any "probe is inert at default α" claim. Distinguishes R4-template-locked from R4-amplitude-locked.

5.5 Diagnostic D5 — Whitespace-stripped flip metric

Procedure. Any α-sweep behavioral flip metric must compare `base.strip() != modified.strip()` rather than `base != modified`. Leading and trailing whitespace differences are tokenization artifacts and do not reflect semantic flips.

Failure mode caught. Tokenization-inflated flip rates. A raw-comparison flip metric counts every change in leading-space tokens as a flip, producing dramatic over-reporting.

Concrete case. Paper-5 Phase 10 originally reported RG_L55_mid_think flip rate 96% at $\alpha = +200$. Without D5, this would have shipped as "the reasoning-quality probe is a powerful lever." Adding D5 reduced the metric to 32% — a 64-percentage-point inflation due to leading-space tokenization. The 32% finding is still publishable, but the headline $3\times$ false-claim was avoided.

Cost. $<$ 60 seconds (string-comparison change).

Recommended at. Any α-sweep behavioral flip metric.

5.6 Diagnostic D6 — Onset-timing sweep (steering-causality only)

Procedure. For continuous-trajectory steering claims, run a within-experiment onset-timing ablation: hold (probe, layer, α, direction, prompt) fixed and vary only the onset token. Standard onsets: token 1, token 50, token 200. Report effectiveness at each onset. A causal probe that satisfies C2 will show monotonic decay; a causal probe that does *not* depend on trajectory will show flat effectiveness across onsets.

Failure mode caught. Confusing trajectory-shaping causal authority with state-attractor causal authority. The two are mechanistically distinct and imply different SDK / production deployment strategies.

Concrete case. Paper-8 Phase 2A originally reported ST_L31_gen levers 9/14 prompts. Without D6, the natural SDK design was "adaptive steering: detect overthinking, then steer." Adding D6 revealed delayed-onset effectiveness drops 9/10 $\rightarrow$ 3/10 $\rightarrow$ 0/10 from token 1 $\rightarrow$ step 50 $\rightarrow$ step 200. The mechanism is trajectory-shaping (preventive enforcement from token 1), not adaptive intervention. The SDK was redesigned around `preventive_compute_enforcement` mode.

Cost. $\sim$10 GPU-minutes (3 additional sweep configurations).

Recommended at. Any "the probe steers behavior X" claim where X is a continuous-trajectory variable (thinking-length, reasoning-quality, generation-style).

Table 4 — Diagnostic Coverage Matrix

Diagnostic	Failure mode	Catches in our papers	Cost	When mandatory
D1 random-feature	Over-parameterization at N<100	Paper-6 Phase 5d $\rightarrow$ 6c	$\sim$10 min	N<100, K>10
D2 shuffled-source	Marginal-fit pathology	Paper-3 PSAE v1.5	$\sim$10 min	Sparse top-k targets
D3 control-token norm	Softmax-temp artifact	Paper-6 Phase 7 (SWE pre-tool)	<1 min	Log-prob shift claims
D4 structural-rigidity α-sweep	Amplitude-null vs structural-null	Paper-6 Phase 8 (CoT L55)	<1 min	"Probe is inert at default α"
D5 whitespace-stripped flip	Tokenization inflation	Paper-5 Phase 10 (RG L55)	<60 s	All α-sweep flip metrics

Diagnostic	Failure mode	Catches in our papers	Cost	When mandatory
D6 onset-timing sweep	Trajectory vs state-attractor confusion	Paper-8 Phase 2B	∼10 min	Continuous-trajectory steering claims

6. Five Case Studies

Each case study reports a *specific point in the lifecycle of a probe paper* at which one diagnostic prevented a false causal claim from shipping. The case studies are not exhaustive — they are the five sharpest near-misses.

6.1 Case 1 — ST_L31_gen: a trajectory-shaping causal probe (paper-8)

The setup. Subjective-time probe at L31 predicts thinking-fraction with R^2=0.86. Phase 2A reports 9/14 termination on GSM8K, Fisher p=0.0092.

The temptation. Ship as "adaptive overthinking detection": deploy as runtime sensor that triggers steering when predicted-fraction exceeds 0.85.

The diagnostic that caught it (D6). Onset-timing sweep revealed that delayed-onset steering drops effectiveness from 9/10 to 3/10 (step 50) to 0/10 (step 200). Two closed-loop designs (probe-as-sensor threshold trigger, plateau-detector) failed at 1-2/10.

The reframed claim. Trajectory-shaping causal authority: the probe levers *only when applied continuously from token 1*. The SDK ships as `preventive_compute_enforcement` mode, not adaptive intervention. R1.

6.2 Case 2 — CoT_L55_mid_think: a template-locked structural probe (paper-6)

The setup. CoT-Integrity probe at L55 mid_think achieves AUROC=0.91 on N=240, the strongest predictive probe in our portfolio.

The temptation. Ship as "deploy this probe to steer reasoning-emission behavior at inference time."

The diagnostic that caught it (D4). Structural-rigidity α-sweep to α=\$±\$200 then α=+500 produces zero behavioral change for both probe-direction and random-direction. Distinguished template-lock (R4 structural) from amplitude-null (R4 inert) by inspecting the chat template: the `enable_thinking=False` variable injects `<think></think>` into input tokens before the residual stream forms (C5).

The reframed claim. R4 structurally-locked, mechanism = input-token template-lock. AUROC=0.91 represents post-decision readout, not causal mediation. Detection-only deployment.

6.3 Case 3 — PSAE v1.5: marginal-fit pathology in sparse top-k prediction (paper-3)

The setup. Twelve (layer, source-fraction) probes predict end-of-thinking SAE features from earlier-thinking residuals. Recall@1024 = 0.85–0.87 at L11, 0.79–0.84 at L31, 0.67–0.72 at L55.

The temptation. Ship as "predictive SAE features within reasoning: the model anticipates its own conclusion 50% of the way through thinking." A ready-made paper at NeurIPS MI Workshop.

The diagnostic that caught it (D2). Shuffled-source baseline (X_train shuffled, y_train kept) reproduced real recall at all twelve sites with max $|\Delta|$ = 0.027 (well within the \$±\$0.03 noise band). A trivial constant baseline strictly exceeded the trained probe at L11/L31. The probe was learning the marginal distribution of end-of-thinking features, not per-prompt prediction.

The reframed claim. Honest-negative methodology contribution: the marginal-fit pathology, its five structural conditions, and D2 as a pre-publication diagnostic for sparse-target probe-prediction work.

6.4 Case 4 — SWE_L43_pre_tool: epiphenomenal-via-softmax-temperature (paper-6)

The setup. Capability probe at L43 pre_tool achieves AUROC=0.83 on N=99 SWE-bench Pro patches. Under α=+5 steering, Δ(log p("finish")) = +0.479 — an apparent strong causal lever on the "finish" tool token.

The temptation. Ship as "L43 pre_tool levers tool-finish decisions; SDK exposes a `boost mode` that steers toward early termination on confidence."

The diagnostic that caught it (D3). Control-token normalization: compute Δrel = Δ("finish") − mean(Δ({"the", "a", "is", "of", "and"})). Result: mean(Δ(controls)) = +0.525, Δrel = −0.046. The +0.479 shift was a uniform softmax-temperature increase across the vocabulary, not a specific lever on "finish."

The reframed claim. R5 epiphenomenal-via-softmax-temperature. SDK ships without boost mode; only detection-side gating. The paper documents D3 as the diagnostic that should be applied to every log-prob-shift causality claim.

6.5 Case 5 — Cap_L55_pre_tool: saturation-direction direction-flip (paper-5)

The setup. Capability probe at L55 pre_tool levers pushdown +34pp on HumanEval+MBPP at α=−100. The original Phase 11 framing predicted "continuous-gradient probes lever pushup; categorical-decision probes lever pushdown."

The temptation. Ship the categorical-vs-continuous framing as the unifying theory: capability is categorical $\rightarrow$ pushdown lever, reasoning-quality is continuous $\rightarrow$ pushup lever.

The diagnostic that caught it (Phase 11e cross-distribution validation). Tested same Cap_L55_pre_tool probe on Codeforces rating $\geq$ 2000 (Qwen pass-rate ∼7%, baseline saturated *away from* capability). The probe *flipped lever direction*: α=−100 pushdown gap −3pp, α=+200 pushup gap +40pp. The categorical-vs-continuous theory was falsified within its own falsifying experiment.

The reframed claim. Saturation-direction principle (Constraint C4): the probe levers along the headroom axis of the baseline distribution. Categorical-vs-continuous is the wrong axis; baseline-saturation alignment is the right one. The framework now subsumes the original Phase 11 finding as a special case where HumanEval+MBPP happens to saturate the "strong capability" direction.

---

7. Discussion

Implications for SAE-as-monitoring deployments. If the field's SAE features behave like our twelve probes (and the SAEBench 2026 evidence suggests many do), then the appropriate framing for SAE-based monitoring is **conditionally causal**: each SAE feature requires per-(layer, trajectory, magnitude, direction, target-class) characterization before being deployed as a behavioral lever. Production deployments that assume SAE features are "semantic axes" with stable causal authority across deployment contexts are likely to encounter our R4 and R5 regimes — features that detect with high AUROC but do not lever, or that appear to lever under naive evaluation but collapse under D3.

Implications for steering-as-alignment-tool. The trajectory-shaping constraint C2 implies that the dominant steering paradigm — wait for the model to need help, then intervene — is mechanistically backwards for ongoing-decision behaviors. The KV-cache lock-in mechanism (paper-8 Phase 2B) predicts that adaptive interventions on long-horizon decisions are fundamentally limited by cache state at intervention time. Steering as a robust alignment tool requires *preventive enforcement from generation start*, not *adaptive intervention*. This aligns with the KV Cache Steering direction (Belitsky et al. 2026) which intervenes on cache state directly.

Implications for chain-of-thought monitorability. The C5 architectural constraint demonstrates that *text-only CoT monitoring is structurally incomplete* on instruction-tuned reasoning models. Decisions encoded in chat-template tokens — including the decision to think at all — are *upstream* of any residual-stream representation and therefore *invisible* to monitoring strategies that read residual activations. This is the companion

finding to our position paper "Activation-Bounded CoT Monitorability" (Vicentino 2026): text-only CoT and activation-derived monitoring are structurally complementary halves of any complete monitoring strategy, and the Frontier Model Forum's January 2026 issue brief underestimates this by treating CoT monitoring as the primary lever.

The five-axis framework as a falsifiable hypothesis. The framework proposed in §4 is empirical and open to falsification. The strongest falsifier would be a probe that satisfies all five constraints C1-C5 and nevertheless fails to lever behavior (would imply additional constraints beyond the five), or a probe that fails one of C1-C5 and nevertheless levers robustly (would imply the failing constraint is not actually necessary). We have not encountered either in the twelve-probe map or in the corroborating cross-model NLA work (paper-7), but the framework's claim to generality requires testing across additional model families and probe types. We propose this as the natural next step for the field.

The diagnostic checklist as a publishable standard. The six diagnostics collectively cost under one GPU-hour per probe. We propose that future linear-probe papers report all six routinely, in the same way that held-out test splits, train/val gaps, and confidence intervals are now reported routinely. Adoption is a coordination problem; we contribute the checklist, the implementation (`agent-probe-guard.diagnostics`), and the case studies (§6) demonstrating that the diagnostics catch real near-misses in our own work.

7.4 Concrete deployment implications (testable predictions)

For Anthropic's Persona Vectors deployment. The five-axis framework predicts that persona directions extracted via the Chen et al. (2025) protocol will exhibit primarily Regime R3 (pushdown-asymmetric) when used as runtime levers, with effectiveness modulated by C4 saturation alignment of the test distribution. The C2 trajectory constraint further predicts that persona steering applied mid-generation will be less effective than persona steering applied via system-prompt enforcement upstream of cache accumulation. We predict: residual-stream persona probes are correlative-effective for *detection* of personality drift but lever-effective for *intervention* only under preventive-enforcement protocols, not adaptive runtime intervention.

For Goodfire's Silico platform. SAE feature catalogs would benefit from per-feature operational-regime annotation. A feature exhibiting R1 requires preventive intervention from generation start; a feature exhibiting R3 requires baseline-saturation-aware deployment; a feature exhibiting R4 should be exposed as a *monitor*, not a *lever*. This annotation can be generated automatically by running the six-diagnostic battery on each catalog entry — a one-GPU-hour-per-feature cost that prevents the most common mode of customer disappointment with steering-based products: the feature detects with high AUROC but does not lever under intervention.

For OpenAI and chain-of-thought monitoring deployments. The C5 constraint implies that text-only CoT monitoring on instruction-tuned models has a structural ceiling. Decisions encoded in chat-template variables (enable_thinking, system-prompt tokens, special control tokens) are upstream of any residual-stream representation and therefore invisible to residual-probe monitoring. We predict that adversarial behaviors which exploit template-level decisions will be invisible to residual probes regardless of probe AUROC, and that activation-derived monitoring should be paired with template-level monitoring as structurally complementary halves of a complete monitoring strategy.

For UK AISI's Alignment Project and grant-funded interp labs. We recommend that any safety-critical probe deployment require the six-diagnostic battery to be reported and reproducible, in the same way that production ML systems require reproducibility of training and evaluation. The infrastructure cost of operational characterization is small (under one GPU-hour per probe); the cost of deploying an R5-epiphenomenal probe in a safety-critical position is potentially structural. The current portfolio of "high-AUROC probe" papers in the 2024-2026 alignment literature is unlikely to transition to deployment-grade tooling without this characterization.

These are predictions derived from a five-axis framework empirically validated on 12 sites in one model. They are testable. The four labs named above are those whose published deployment pipelines most directly intersect the framework's predictions; we welcome empirical falsification.

---

8. Limitations

Single model family. All twelve probes are on Qwen3.6-27B. The saturation-direction principle has been corroborated on Qwen2.5-7B (paper-7 NLA) and Gemma-3-12B (paper-7 V2), but the full five-axis framework has not been tested on other model families. We expect C1, C2, C5 to generalize architecturally; C3 and C4 are likely to require per-model calibration of sweep ranges and saturation profiles.

Twelve probes is not the population. Our taxonomy is at saturation for the families we have probed (subjective-time, capability, persona, reasoning-quality, SAE-feature-prediction). Other probe families — deception-detection, refusal-circuit, sycophancy, jailbreak-vulnerability — have not been mapped under the unified protocol. We anticipate that additional regimes may exist (e.g. a "circuit-completion" regime distinct from R1-R5), but we have not yet observed them.

Cross-distribution validation is partial. Constraint C4 (saturation-direction) was validated cross-distribution on capability probes (HumanEval+MBPP vs. BigCodeBench vs. Codeforces). The other constraints have not been cross-distribution-validated. The framework's claim to *operational* generality across deployment contexts requires additional cross-distribution work.

The five diagnostics are necessary but not sufficient. The diagnostics catch the five specific failure modes documented in §5. They do not exhaustively cover all possible failure modes of linear-probe causality claims. We expect additional diagnostics to be needed for probes on mixture-of-experts models, multi-modal models, and tool-use environments, which we have not extensively tested.

Causal-abstraction-theoretic formalization is informal. §4 provides empirical regularities and conjectured mechanisms. A formal connection to the Geiger et al. causal abstraction framework — in particular, mapping each constraint to a precondition on interchange-intervention validity — remains to be developed. We sketch the connection in §2 but do not develop it.

---

9. Conclusion

We have mapped twelve linear probes on Qwen3.6-27B under a unified evaluation protocol, identified five distinct empirical causal-class regimes, proposed a five-axis framework of operational constraints (layer, trajectory, magnitude, direction, target-class) that subsumes the regimes as principled subsets, and distilled the methodology that surfaced the constraints into a six-item pre-publication diagnostic checklist. The framework reframes probe causality as a **conditional property to be measured per deployment configuration**, not a global per-probe attribute. The diagnostics — random-feature baseline, shuffled-source baseline, control-token normalization, structural-rigidity α-sweep, whitespace-stripped flip metric, and onset-timing sweep — collectively cost under one GPU-hour per probe and have caught five near-miss false causal claims in our own work.

The field's growing reliance on probe-based monitoring, RL-reward shaping, and alignment auditing makes the cost of a false probe-causality claim structural. We argue that the five-axis framework and six-item diagnostic checklist are the cheap, generalizable, publishable contribution that any probe-shipping lab should now make on its own portfolio. We release the protocol, capture batches, per-probe verdicts, and the open-source `agent-probe-guard` SDK that implements the diagnostics.

Acknowledgments

This paper consolidates the methodology developed across papers 3, 5, 6, 7, and 8 of the OpenInterpretability series. We thank Jack Lindsey (Anthropic) for the Persona Vectors framing that anchored Constraint C4, Atticus Geiger (Goodfire) for the causal-abstraction-theoretic foundations referenced in §2 and §8, and the kitft team for the Natural Language Autoencoder pairs that enabled cross-model corroboration in paper-7. The five diagnostics were developed not from foresight but from five distinct moments in which we ourselves

shipped or nearly shipped a false causal claim. We thank the process of having to walk back claims for the methodological clarity it ultimately produced.

Appendix A — Empirical verification trail

Every numerical claim in this paper has been verified against primary run artifacts before publication, and the artifacts have been published as the public HF dataset `caiovicentino1/openinterp-paper-mega-conditionally-causal` (69 JSON files, 13 MB, Apache-2.0). The verification is reproducible end-to-end via `scripts/verify_paper_mega_claims.py`, which downloads the dataset, re-derives each numerical claim from the raw JSONs, and prints PASS/FAIL per row. As of v3.4 (2026-05-17) the script reports **26/26 PASS**:

Claim in paper	Primary source	Verified value	Status
PSAE recall@1024 L11 = 0.85–0.87	`predictive_sae_v15_results.json`	0.852–0.869	✓
PSAE recall@1024 L31 = 0.79–0.84	same	0.788–0.835	✓
PSAE recall@1024 L55 = 0.67–0.72	same	0.671–0.724	✓
PSAE B1 shuffled-source max$\|\Delta\|$ = 0.027	`random_baseline_results.json`	max 0.027 across 12 sites	✓
Phase 2A 9/14 shortens (probe α=+50)	`phase2a_aggregate_stats.json`	probe shortens_rate = 0.6429 = 9/14	✓
Phase 2A 2/14 random shortens	same	random_shortens_rate = 0.1429 = 2/14	✓
Phase 2A Fisher OR=10.8, p=0.0092	same	OR=10.8, p=0.00915	✓
Phase 2A mean gap −33pp	same	gap_pp = −32.82	✓
Phase 2A baseline thinking len 530	`phase2a_summary.json`	529.8	✓
SWE-Verified 19/20 probe terminate	`swe_transfer_test.json` + `caveat1_cross_repo/results_cross_repo.json`	10/10 (astropy) + 9/10 (cross-repo) = 19/20	✓
SWE-Verified 6/20 random terminate	same combined	3/10 + 3/10 = 6/20	✓
Phase 10 RG L55 raw flip 96% at α=+200	`phase10_verdict.json`	rg_probe_summary flip_rate = 0.96	✓
Phase 10 RG L55 random 2% at α=+200	same	rg_random_summary flip_rate = 0.02	✓
Phase 11 Cap_L31_pre_tool +40pp at α=−100	`phase11_verdict.json`	probe 0.867 − random 0.467 = 0.40	✓
Phase 11 Cap_L55_pre_tool +34pp at α=−100	same	probe 0.467 − random 0.133 = 0.334	✓
Phase 7 SWE_L43_pre_tool Δrel at α=+2 = −0.046 (control-token norm)	computed from `phase7_steering_pilot.json` logprobs	Δrel = −0.0458 across N=6 fails	✓

Claim in paper	Primary source	Verified value	Status
Phase 11e Cap_L55_pre_tool CF $\alpha=-100 \approx -3$pp (pushdown null)	`phase11e_multisite_cf/probe_L55_pre_tool.json`	probe 0.200 − random 0.233 = −3.3pp	✓
Phase 11e Cap_L55_pre_tool CF $\alpha=+200 = +40$pp pushup (direction flip)	same	probe 0.933 − random 0.533 = +40pp	✓
Phase 11e Cap_L43_turn_end CF $\alpha=-100 \approx +7$pp (collapse from HE+MBPP)	`partial_L43_turn_end.json`	0.567 − 0.500 = +6.7pp	✓
Phase 11e Cap_L23_pre_tool CF $\alpha=-100 = +43$pp (sat-independent)	`partial_L23_pre_tool.json`	1.000 − 0.567 = +43.3pp	✓
Phase 11e Cap_L31_pre_tool CF $\alpha=-100 = +37$pp (sat-independent)	`partial_L31_pre_tool.json`	0.900 − 0.533 = +36.7pp	✓
Phase 2C L55 inert at all α up to +500	`subjective_time_phase2c/final_cross_activation.json`	60 paired prompts × {α=+100,+200,+500}	✓
HF qwen36-27b-sae-papergrade	huggingface.co API	40 files present	✓
HF openinterp-psae-v15-marginal-fit-pathology	same	10 files present	✓
HF openinterp-paper7-nla-two-tier-verbalization	same	10 files present	✓

Reproduction: All raw probe artifacts are publicly available under Apache-2.0. The verification script `verify_paper_mega_claims.py` is the source of truth for the 26 verified rows above. Run `python3 scripts/verify_paper_mega_claims.py` after `pip install huggingface_hub`; the script downloads the dataset bundle and prints `26 PASS, 0 FAIL`.

References

Verified primary sources

- Belitsky, M., Kopiczko, D. J., Dorkenwald, M., Mirza, M. J., Glass, J. R., Snoek, C. G. M., Asano, Y. M. *KV Cache Steering for Controlling Frozen LLMs.* arXiv:2507.08799. 2025. https://arxiv.org/abs/2507.08799
- Chen, R., Arditi, A., Sleight, H., Evans, O., Lindsey, J. *Persona Vectors: Monitoring and Controlling Character Traits in Language Models.* arXiv:2507.21509. 2025. https://arxiv.org/abs/2507.21509
- Geiger, A., Lu, H., Icard, T., Potts, C. *Causal Abstractions of Neural Networks.* NeurIPS 2021. https://arxiv.org/abs/2106.02997
- Geiger, A., Wu, Z., Potts, C., Icard, T., Goodman, N. *Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability.* arXiv:2301.04709.

2023. (Published JMLR vol. 26, 2025.) https://arxiv.org/abs/2301.04709
- Karvonen, A., et al. *SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability*. arXiv:2503.09532. 2025. https://arxiv.org/abs/2503.09532
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., et al. *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Transformer Circuits Thread, Anthropic. 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/
- Turner, A., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., MacDiarmid, M. *Activation Addition: Steering Language Models Without Optimization*. arXiv:2308.10248. 2023. https://arxiv.org/abs/2308.10248

OpenInterpretability series companion papers

- Vicentino, C. *Paper-3: The Marginal-Fit Pathology in Predictive SAE Feature Trajectory Probes*. OpenInterpretability. May 2026. https://openinterp.org/research/papers/marginal-fit-pathology-psae
- Vicentino, C. *Paper-5: Saturation-Direction Lever — A Five-Class Taxonomy of Probe Causality in Qwen3.6-27B*. OpenInterpretability. May 2026. https://openinterp.org/research/papers/saturation-direction-probe-levers
- Vicentino, C. *Paper-6: Two Forms of Epiphenomenal Probes in Qwen3.6-27B*. OpenInterpretability. May 2026. https://openinterp.org/research/papers/two-forms-epiphenomenal-probes
- Vicentino, C. *Paper-7: Reconstruction Without Recall — Two-Tier Verbalization in Natural Language Autoencoders*. OpenInterpretability. May 2026. https://openinterp.org/research/papers/nla-two-tier-verbalization
- Vicentino, C. *Paper-8: Trajectory-Shaping Probe Steering in Qwen3.6-27B Reasoning*. OpenInterpretability. May 2026. https://openinterp.org/research/papers/probe-guided-anti-overthinking
- Vicentino, C. *Activation-Bounded CoT Monitorability*. OpenInterpretability position paper. May 2026. https://openinterp.org/research/papers/activation-bounded-cot-monitorability

All references above have been verified against arXiv, JMLR, transformer-circuits.pub, or the OpenInterpretability site.