Contents

Saturation-Direction Lever: A Five-Class Taxonomy of Probe Causality in Qwen3.6-27B 1

When linear probes detect, when they lever, and why direction-asymmetric authority emerges in instruction-tuned reasoning models . . . 1

Abstract . . . 1

1. Introduction . . . 2

2. Method — The Causal Locus Protocol . . . 3

2.1 Definitions . . . 3

2.2 Four sanity checks (mandatory, per paper-3) . . . 4

3. Setup . . . 4

3.1 Model and decoding . . . 4

3.2 Probes tested . . . 4

3.3 Test prompts . . . 5

4. Results — Five Empirical Classes . . . 5

4.1 Class 1 — Surface softmax-temperature artifact . . . 5

4.2 Class 2 — Template-locked categorical decision . . . 5

4.3 Class 3 — Structural fragility . . . 6

4.4 Class 4a — Pushup-asymmetric continuous-quality lever . . . 6

4.5 Class 4b — Pushdown-asymmetric capability and persona lever . . . 6

4.6 Falsifier confirmation . . . 6

5. Discussion — The Saturation-Direction Lever . . . 7

5.1 Unifying principle . . . 7

5.2 Why probe vs random differ in the saturation direction . . . 7

5.3 Connection to Anthropic Persona Vectors . . . 7

5.4 Connection to alignment training failures . . . 8

5.5 Cross-distribution validation — site-partitioned robustness . . . 8

5.6 The 4 sanity checks save publishable claims . . . 9

6. Limitations . . . 9

7. Conclusion . . . 10

Reproducibility Statement . . . 10

Acknowledgments . . . 11

References . . . 11

Saturation-Direction Lever: A Five-Class Taxonomy of Probe Causality in Qwen3.6-27B

When linear probes detect, when they lever, and why direction-asymmetric authority emerges in instruction-tuned reasoning models

Workshop draft for NeurIPS 2026 Mechanistic Interpretability Workshop. Apache-2.0. Reproducible on a single RTX 6000 Blackwell in ∼6 hours.

Abstract

Linear probes on transformer residual streams routinely achieve high predictive AUROC, yet their causal authority — whether their direction levers downstream behavior — has been only sparsely tested at frontier scale. We map probe causality across 8 probes (5 layers, 5 positions, 3 training-objective classes) on Qwen3.6-27B using a unified protocol combining bidirectional α-sweep up to α=\$±\$200, random K-matched control direction, control-token-normalized log-prob shifts, structural-rigidity diagnostic, and whitespace-stripped behavioral flip metric. We document **five empirical classes** of probe-causality regime and identify a single unifying principle — *probes lever in the saturation direction of the baseline residual* — that explains all observed

asymmetric-lever cases including a falsified prediction. The classes are: (1) surface softmax-temperature artifact (L43 capability), (2) template-locked categorical decision (L55 thinking emission, L31 fabrication-detection), (3) structural fragility at fragile layers (L11/L43 think_start), (4a) pushup-asymmetric lever for reasoning quality at high amplitude (RG L55 mid_think, +30pp gap), and (4b) pushdown-asymmetric lever for capability and persona at high amplitude (5 sites, +30 to +60pp gap). We falsify the naive prediction that continuous-gradient probes lever in the pushup direction by demonstrating that persona — a continuous gradient — levers in the pushdown direction when the test prompt's baseline is in the "helpful" saturation region. The unifying refinement: probe direction has causal authority along the axis where the baseline residual has behavioral *headroom* to flip; the random-direction control flips generations only via OOD destruction, while the probe direction additionally flips via OOD-semantic perturbation in the saturated subspace. We release the protocol, all 6 capture batches, and per-site verdicts under Apache-2.0 and propose the saturation-direction lever as a predictive heuristic for which behavioral interventions a given probe will and will not afford. Cross-distribution validations on BigCodeBench (Qwen pass-rate ∼55%) and Codeforces rating \$≥\$2000 (∼7%) extend the $\alpha=-100$ pushdown gap from HumanEval+MBPP (∼89%) and reveal a **site-dependent robustness profile**. Two of four pushdown-asymmetric capability sites — L23 pre_tool and L31 pre_tool — are saturation-independent: $\alpha=-100$ pushdown gap holds at +43pp and +37pp on Codeforces vs +50pp and +40pp on HumanEval+MBPP. Two other sites — L43 turn_end and L55 pre_tool — are saturation-coupled: L43 turn_end's gap collapses (+7pp on Codeforces vs +40pp on HumanEval+MBPP), while L55 pre_tool *flips direction* from pushdown on the saturated distribution to pushup at $\alpha=+200$ on the unsaturated distribution. The direction flip is consistent with the saturation-direction principle itself: the lever pushes against the baseline saturation, and when saturation flips, the lever flips. We pre-registered and walked back two predictions in this paper (categorical-vs-continuous lever, then saturation-magnitude corollary), and the surviving thesis combines a **site-partitioned robustness theorem** with the **saturation-direction principle** that subsumes it.

Keywords: linear probes, activation steering, causal interpretability, mechanistic interpretability, Qwen3.6-27B, asymmetric lever, saturation direction.

---

1. Introduction

A linear probe on a transformer's residual stream is cheap, easy to fit, and frequently surprisingly accurate at predicting downstream observables — hallucination, reasoning quality, persona, agent action, refusal trigger. As probes proliferate from monitoring to deployed safety classifiers and reward signals (Templeton et al. 2024; Marks et al. 2024; OpenAI 2026), a load-bearing question becomes: when does a high-AUROC probe direction *also* lever downstream behavior under intervention, and when is it merely a correlative read of features downstream of the actual decision?

Our prior work (paper-3 of this series, "Two Forms of Epiphenomenal Probes") documented two specific failure modes for probe-causality on this model: (i) softmax-temperature artifacts at L43 pre_tool (Phase 7); (ii) template-locked decisions at L55 last_prompt (Phase 8). Both findings showed probe directions that *detect* a behavioral outcome (AUROC ≥ 0.83) without *levering* it (zero behavioral flip at $\alpha > \|\text{residual}\|$, with both probe and random-direction null). These two mechanisms implied a tempting general theory: probes are detection-only in this model class.

This paper documents that the general theory is too strong. Across **8 probes spanning 5 layers, 5 positions, and 3 training objectives** in Qwen3.6-27B, we find:

1. Two probes are **pure epiphenomenal** as previously reported (paper-3 §6.1, §6.2);
2. One probe (FabricationGuard L31 end_of_think) is **also pure epiphenomenal** under the full protocol — its direction is statistically indistinguishable from a random K-matched direction in behavioral effect across the entire α range;
3. Two layers exhibit **structural fragility** — both probe and random directions destabilize generations equally at $\alpha \geq \|h\|$, providing no information about the probe;
4. **Five probes lever asymmetrically** at $\alpha \approx \|h\|$ with probe-vs-random gap $\geq +30$pp, but the *direction* of

the lever is heterogeneous: one probe levers pushup (positive α), four probes lever pushdown (negative α); and
5. The direction of the asymmetric lever is **not** predicted by whether the underlying behavior is categorical (binary) or continuous (gradient) — a hypothesis we explicitly falsify by testing a continuous-gradient probe (persona) and observing pushdown asymmetry instead of the predicted pushup.

We propose a unifying refinement: **probes lever in the saturation direction of the baseline residual**. Where the model's baseline activation is already deep in one half-space along the probe axis, pushing further into that half-space produces OOD behavior with probe-specific semantic leverage that random directions don't share. The opposite half-space, by contrast, requires more than amplitude to elicit qualitatively different behavior — it requires context, tokens, or template that the residual modification alone cannot supply. We call this the **saturation-direction lever** principle and show it explains all five observed asymmetric-lever cases in our portfolio.

Contributions.

1. **Protocol consolidation**: a single unified causal-locus protocol combining four sanity checks from paper-3 (random K-matched baseline, control-token normalization, structural-rigidity α-sweep, whitespace-stripped flip metric) plus bidirectional α-sweep and cross-prompt-set robustness, applicable to any probe at any (layer, position) site.
2. **Empirical map**: 8 probes mapped across 5 empirical classes of probe-causality regime in a single frontier reasoning model.
3. **Falsifier finding**: explicit falsification of the "continuous-gradient $\rightarrow$ pushup-lever" prediction using persona-switch on TruthfulQA. Persona is continuous yet levers pushdown.
4. **Saturation-direction theory**: a principled explanation that unifies all five asymmetric-lever findings under one mechanism, and provides a predictive heuristic for future probe-causality experiments.
5. **Reproducibility**: every batch (Phase 7 / 8 / 10 / 11 / 11b / 12), every notebook, and every verdict JSON is public under Apache-2.0.

---

2. Method — The Causal Locus Protocol

For an arbitrary behavior `Y` produced by an instruction-tuned model `M`, the protocol identifies whether a probe trained against `Y` has causal authority — and if so, in which direction.

2.1 Definitions

Probe direction. Top-K=10 signed diff-of-means feature selection (Phase 6c §3.1 method) on labeled residual captures, L2-normalized. Random K-matched baseline: 10 random dimensions with random Gaussian sign, L2-normalized.

Behavioral metric. Greedy 40-token generation under one-shot forward-hook intervention at the chosen `(layer, position)`: `h[:, -1, :] += $\alpha$ $\times$ direction_vec` at the last position of the prefill. Stripped behavioral flip: `base.strip() != modified.strip()` (paper-3 §3.4).

α **grid.** `{-200, -100, -50, -20, -5, -2, 0, +2, +5, +20, +50, +100, +200}`. Includes the typical activation-steering range (\$$\pm$\$2 to \$$\pm$20), $themoderate-OODrange(\pm$\$50 to \$$\pm$100), $andthestrong-OODrange(\pm200)whereamplitudeexceedstypicalresidualnorm$`|h|$ $\approx$ 70–160‘.

Verdict classifier. For each `(layer, position)` site: - *Lever* if `flip_rate(probe, $\alpha$₀) − flip_rate(random, $\alpha$₀) $\geq$ +0.30` for some `$\alpha$₀ $\in$ {$\pm$50, $\pm$100, $\pm$200}` and the same gap is $< +0.10$ in the opposite direction. - *Structural fragility* if `flip_rate(probe, $\alpha$₀) $\approx$ flip_rate(random, $\alpha$₀)` at all α with both $\geq$ 0.40 at α=\$$\pm$\$200 (random is destroying generations). - *Epiphenomenal* if `flip_rate < 0.10` for both probe and random across all α and `|$\Delta$rel|` (control-token-normalized log-prob shift) < 0.10. - *Softmax-temp artifact* if behavioral flips occur but `$\Delta$rel $\approx$ 0` uniformly across α (paper-3 §6.1 mechanism). -

Template-lock if `flip_rate $\leq$ 0.05` for both probe and random at α=$±$200, with residual-norm modification confirmed by hook-fire trace (paper-3 §6.2 mechanism).

2.2 Four sanity checks (mandatory, per paper-3)

1. **Random K-matched baseline**: at small N (<100), the gap `AUROC_probe − AUROC_random` is the probe's signal, not the absolute AUROC. Caught Phase 5d K=50 N=17 → 1.000 false signal (paper-3 §3.1).
2. **Control-token normalization**: any α-induced log-prob shift on a target token must be reported as `$\Delta$rel = $\Delta$(target) − mean($\Delta$(controls))` across $≥$5 control tokens. Caught Phase 7 +0.479 nat naive shift as pure softmax-temperature (paper-3 §3.2).
3. **Structural-rigidity α-sweep**: before declaring a steering null at α∈{$±$2, $±$5}, sweep to $\alpha \gg \|h\|$ on probe AND random direction. If output remains rigid, decision lives outside any residual probe could reach (paper-3 §3.3).
4. **Whitespace-stripped flip metric**: leading-space tokenization artifacts at high α can inflate naive flip rate by 60+pp. Always `base.strip() != modified.strip()`. Caught Phase 10 RG α=+200 raw 96% → stripped 32% (paper-3 §3.4).

These four are not optional. Three of the four caught a confident-but-wrong claim during this work; absent them we would have shipped at least three falsified findings.

---

3. Setup

3.1 Model and decoding

Qwen3.6-27B (Alibaba 2026), 64 layers, hybrid GDN + standard attention, bf16 inference on RTX 6000 Blackwell (96GB VRAM). Released chat template, `enable_thinking=True` for capability and reasoning probes, `enable_thinking=False` for thinking-emission probe (Phase 8).

3.2 Probes tested

Eight probes spanning 3 training objectives:

Probe	Layer	Position	Behavior Y	Probe AUROC
L43 capability (Phase 7)	L43	pre_tool	trace patch generated	0.830 (Phase 6c N=42)
L55 thinking last_prompt (Phase 8)	L55	last_prompt	<think> continued past auto-injection	0.848 K=5 / 0.91 full
FabricationGuard (Phase 10)	L31	end_of_think	hallucination on HaluEval	0.81
ReasoningGuard (Phase 10)	L55	mid_think	reasoning quality on GSM8K	0.888 within domain
Capability L11 think_start (Phase 11)	L11	think_start	trace patch generated	0.795 (Phase 6 N=99)
Capability L31 pre_tool (Phase 11)	L31	pre_tool	trace patch generated	0.926 (Phase 6 N=99)
Capability L55 pre_tool (Phase 11)	L55	pre_tool	trace patch generated	0.930 (Phase 6 N=99)

Probe	Layer	Position	Behavior Y	Probe AUROC
Capability L43 think_start (Phase 11)	L43	think_start	trace patch generated	0.966 (Phase 6 N=99)
Capability L23 pre_tool (Phase 11b)	L23	pre_tool	trace patch generated	0.881 (Phase 6 N=99)
Capability L43 turn_end (Phase 11b)	L43	turn_end	trace patch generated	0.775 (Phase 6 N=99)
Persona L43 last_prompt (Phase 12)	L43	last_prompt	helpful vs villainous system prompt	1.000 4-fold CV

The capability cluster shares a single labeled corpus: 99 SWE-bench Pro traces (Phase 6 N=99) where 89 produced a non-empty patch and 10 did not. Probes are trained on the residual mean across all position-matching capture records per trace, so 89-vs-10 binary classification is the supervision signal. Class imbalance is handled with stratified k-fold + `class_weight='balanced'` in scikit-learn's `LogisticRegression`.

The persona probe was trained on 30 questions $\times$ 2 system prompts (helpful vs villainous) = 60 captures. Probe AUROC = 1.000 at 4-fold CV is consistent with prior work showing persona to be the most separable axis in the residual stream of instruction-tuned models (Anthropic 2025 Persona Vectors).

3.3 Test prompts

For each probe, behavioral testing uses ~30 prompts drawn from distribution-matched datasets: HumanEval (15) and MBPP (15) for capability; HaluEval QA (50) for FabricationGuard; GSM8K (50) for ReasoningGuard; TruthfulQA (15 helpful-system-prompt) for persona; HotpotQA RAG-augmented (240) for thinking-emission (re-used from paper-3 §2.3).

---

4. Results — Five Empirical Classes

4.1 Class 1 — Surface softmax-temperature artifact

Phase 7, L43 pre_tool capability probe at N=54 (paper-3). Naive log-prob proxy at α=+2 reported `$\Delta$log-prob(finish) = +0.479 nats`. Control-token normalization across `search/execute/write/read/wait` revealed all five control tokens shifted by +0.5 to +1.0 nats — a uniform softmax-temperature shift, not target-specific bias. `$\Delta$rel = $\Delta$(target) - mean($\Delta$(controls)) = -0.046`, essentially zero. Single-shot behavioral generation at α=+5: 4/4 fails select identical tool. Triple-source convergent: probe is detection-only via the softmax-temperature mechanism.

4.2 Class 2 — Template-locked categorical decision

Phase 8, L55 last_prompt thinking-emission probe at N=240. Probe AUROC 0.91. α-sweep to ± 200(`+27|h|$ = 158) on probe AND random direction: **32 of 32 generations identical char-by-char**. Hook fires; residual is verifiably perturbed by $\Delta$$\|$$\cdot$$\|$ = 200; output does not flip. Diagnosis: the chat template'senable_thinking=Falseflag injects a closed‘ token pair into the prompt before generation — the thinking decision is encoded in input tokens, downstream of any layer at which the residual could be modified.

Phase 10, FabricationGuard L31 end_of_think probe at N=50 with HaluEval prompts. Stripped flip rates at $\alpha \in$ {$±$5, $±$20, $±$50, $±$100, $±$200}: probe and random-direction flip rates are statistically

indistinguishable across the full α range (probe 4-46% vs random 4-58%). At α=+200 random actually exceeds probe (58% vs 44%). Probe direction is behaviorally indistinguishable from a random direction — a fourth confirmed pure-epiphenomenal probe.

4.3 Class 3 — Structural fragility

Phase 11, L43 think_start (AUROC 0.966) and L11 think_start (AUROC 0.795). Despite high probe AUROCs, the layer-position is fragile to *any* high-amplitude perturbation: at α=$±$200, both probe and random directions flip 90-100% of generations (L11 100/100% at α=$±50 $already$); $thegapbetweenprobeandrandomrarelyexceeds+20ppatthetypicalsteeringrange.Randomdirectiondestroysgenerationsaseffectivelyasprobedirection.Diagnosthelayerisood-fragiletoanyperturbationexceeding$ |h|$; whatever signal the probe carries is swamped by amplitude effects.

4.4 Class 4a — Pushup-asymmetric continuous-quality lever

Phase 10, ReasoningGuard L55 mid_think on 50 GSM8K prompts. Stripped flip rate at α=+200: probe 32% (16/50), random 2% (1/50). **Gap +30pp** with binomial p $\ll$ 1e-5. At α=−200, probe 2%, random 4%. Pushdown direction: no signal. Pushup direction: real lever, but only at amplitude > $\|h\|$, and only flips ∼1/3 of prompts (not a clean behavioral switch).

4.5 Class 4b — Pushdown-asymmetric capability and persona lever

Phase 11 + 11b: four capability probes at decision-bottleneck positions:

Site	α=−100 probe	α=−100 random	gap
L23 pre_tool	100%	60%	**+40pp**
L31 pre_tool	87%	47%	**+40pp**
L55 pre_tool	47%	13%	**+34pp**
L43 turn_end (α=−200)	93%	33%	**+60pp**

All four show asymmetric pushdown lever: probe direction destroys capability behaviorally (model fails to produce code, generates malformed output, or shifts to non-code response) at $\alpha \in \{-50, -100, -200\}$, with random direction producing far weaker effects. Pushup direction ($\alpha > 0$) shows ceiling: probe and random flip rates are comparable, with neither effectively augmenting the model's already-high baseline capability. The pattern is robust across 4 layers (L23, L31, L43, L55) and 2 positions (pre_tool, turn_end).

Phase 12, persona L43 last_prompt on 15 helpful-baseline TruthfulQA prompts:

α	probe%	random%	gap
−200	**100%**	**40%**	**+60pp**
−100	47%	20%	+27pp
−50	27%	13%	+14pp
$±5..±$20	0-13%	0-13%	flat
+50	33%	27%	+6pp
+200	40%	33%	+7pp

Persona is also pushdown-asymmetric — the *opposite* direction of the naive prediction.

4.6 Falsifier confirmation

The naive prediction from §1 was: continuous-gradient probes (RG, persona) should lever pushup; categorical-decision probes (capability, refusal, template-format) should be epiphenomenal or pushdown.

The data falsifies this: persona is continuous-gradient, yet pushdown-asymmetric. The falsifier rules out the categorical-vs-continuous frame as the organizing axis.

---

5. Discussion — The Saturation-Direction Lever

5.1 Unifying principle

Across all five lever findings, we observe a single regularity:

> **The asymmetric lever direction matches the direction in which the baseline residual is saturated along the probe axis.**

Concretely, for each test condition:

- **Capability (HumanEval/MBPP, baseline ≈ success ceiling)**: residual is saturated toward `y=1` (patch-generated). Pushdown $\alpha<0$ pushes *out* of that saturation along the probe axis — into a region where the model's behavior has *headroom* to differ (it can fail). The probe direction has more semantic leverage than random in this transition region.
- **Persona (TruthfulQA helpful prompt, baseline ≈ helpful ceiling)**: residual is saturated toward `y=0` (helpful). Pushdown $\alpha<0$ pushes *deeper* into that saturation; the OOD-saturated region has probe-specific semantic leverage (helpful axis at extreme), causing generations to break in probe-direction-specific ways. Pushup $\alpha>0$ pushes toward `y=1` (villainous), which has the *headroom* to flip but evidently requires more than amplitude — context tokens, system prompt — to manifest as villainous output.
- **RG reasoning (GSM8K, baseline ≈ moderate quality)**: residual is saturated toward `y=0` (lower quality? or ungrounded?) — this is the case where the convention of "positive class = higher quality" gives pushup as the *out-of-saturation* direction. Pushup levers (+30pp); pushdown does not.

The principle is *not* about intrinsic property of the probe (categorical vs continuous) or the layer (early vs late). It is about the relationship between **where the baseline residual sits along the probe axis** and **which direction has behavioral headroom**.

5.2 Why probe vs random differ in the saturation direction

The saturation-direction lever is asymmetric between probe and random because:

1. **Random direction at high α produces OOD residual broadly** — flat semantic content collapse, generic destabilization. This is the "fragility-class" effect documented in §4.3.
2. **Probe direction at high α produces OOD-semantic residual along a specific learned axis** — the OOD perturbation interacts with the model's downstream computations in a way that random does not, *if and only if* the downstream computations are sensitive to perturbation along that semantic axis.
3. The downstream sensitivity is non-uniform: it is highest where the baseline activation is already deep in one half-space and saturated. In the opposite half-space, the model's computation has different sensitivities and the probe direction's semantic interpretation may not generalize OOD.

The result is an asymmetric lever: probe levers more than random in the saturation direction, but probe and random are roughly comparable in the headroom direction (where neither has clean semantic leverage at $\alpha \gg \|h\|$).

5.3 Connection to Anthropic Persona Vectors

Anthropic's Persona Vectors (2025) demonstrated mid-layer steering *works* for persona on Claude — they observed pushup levers when steering toward villainous from a helpful baseline. Our Phase 12 result is at first glance contradictory (we observe pushdown).

The two are consistent under saturation-direction theory: the direction of the asymmetric lever depends on which class (helpful or villainous) is treated as `y=1` in the probe-training convention, and on the test prompt's

baseline. Our convention sets `y=1 = villainous`, baseline = helpful, lever pushdown means "push toward helpful saturation". An equivalent reframing with `y=1 = helpful` gives lever pushup toward villainous from the same baseline. The signed direction reverses; the *saturation-direction* principle (probe levers in the direction the baseline is saturated toward) holds in either convention.

5.4 Connection to alignment training failures

The saturation-direction lever has a direct safety implication. If RLHF training pushes a model into a "helpful" baseline saturation along some persona/refusal axis, then: (a) probe-derived rewards trained on that axis can detect the saturation but cannot constructively augment helpfulness from there (ceiling effect); (b) probe-derived *pushdown* interventions can destroy the saturation more effectively than random ones, suggesting probes may afford selective capability revocation but not selective augmentation. This connects to OpenAI Alignment (2026)'s observation that CoT-text reward shaping has limited reach for "monitor-relevant" properties: the relevant features are upstream in the saturation region and reward pressure cannot constructively reach beyond the existing saturation.

5.5 Cross-distribution validation — site-partitioned robustness

We tested distribution-robustness in three stages. The same probe directions (trained on Phase 6 SWE-bench Pro, N=99, top-K=10 diff-of-means) were applied unchanged to new code distributions spanning a wide saturation range.

Phase 11c (BigCodeBench, single site): 30 prompts, L31 pre_tool only. Qwen3.6-27B baseline pass rate plausibly ∼55%. $\alpha=-100$ pushdown gap +33.3pp.

Phase 11d (Codeforces, single site): 30 prompts from `open-r1/codeforces` filtered to ratings $\geq$ 2000. Qwen pass rate ∼5-10% (lowest in-modality regime). L31 pre_tool only. $\alpha=-100$ pushdown gap +40.0pp.

Phase 11e (Codeforces, four sites): same 30 Codeforces prompts applied to all four pushdown-asymmetric capability sites identified in Phase 11+11b. The four-site test partitions the sites into two regimes:

Site	HE+MBPP $\alpha=-100$	Codeforces $\alpha=-100$	Δ	Regime
L23 pre_tool	+50pp	**+43.3pp**	−6.7	saturation-independent
L31 pre_tool	+40pp	**+36.7pp**	−3.3	saturation-independent
L43 turn_end	+40pp	+6.7pp	−33.3	saturation-coupled
L55 pre_tool	+34pp	−3.3pp	−37.3	direction-flipped

Two of four sites — both at *pre_tool* position in early-to-mid layers — show saturation-independent $\alpha=-100$ lever (gap holds at 87-94% of HE+MBPP value across the saturation range). Two of four sites — both at late-layer / non-*pre_tool* positions — do not. L43 turn_end's lever collapses with saturation; L55 pre_tool's lever *flips direction*, showing pushdown on the saturated distribution and a +40pp pushup at $\alpha=+200$ on the unsaturated distribution.

The $\alpha=-100$ robustness theorem therefore holds **conditionally**: at decision-bottleneck pre_tool positions in early-to-mid layers, the pushdown lever is saturation-independent across code distributions spanning Qwen pass-rate ∼7-89%. At other capability sites within the same broader class, the lever is saturation-coupled or direction-coupled.

The L55 pre_tool direction flip *supports* saturation-direction The most informative finding is L55 pre_tool's direction reversal. On HumanEval+MBPP, where Qwen reaches ∼89% pass rate (residual saturated toward success), the L55 pre_tool probe levers in the pushdown direction at $\alpha=-100$ (+34pp

gap). On Codeforces \$≥\$2000, where Qwen reaches ∼7% pass rate (residual saturated toward failure), the same probe direction levers in the *pushup* direction at $\alpha=+200$ (+40pp gap). The lever pushes against the baseline saturation, and when the saturation flips, the lever flips. This is exactly what saturation-direction theory (§5.1) predicts. The $\alpha=-100$ robustness theorem (uniform pushdown across distributions) is the *special case* of the saturation-direction principle in which the saturation direction is itself stable across distributions; when the saturation direction varies, the lever direction tracks it.

We refine the theory accordingly. **The site-partitioned robustness profile** of paper-5 capability sites:

- **Saturation-independent** (L23 pre_tool, L31 pre_tool): baseline residual is saturated toward success regardless of distribution. $\alpha=-100$ pushdown lever holds.
- **Saturation-magnitude-coupled** (L43 turn_end): lever direction matches baseline (pushdown) but magnitude scales with saturation degree. The original §5.5 corollary applies here, after all.
- **Saturation-direction-coupled** (L55 pre_tool): baseline saturation direction itself flips with distribution. Lever direction follows.

What this does to the saturation-magnitude corollary we walked back We walked back the saturation-magnitude corollary in v3 of this paper when Phase 11d single-site (L31 pre_tool only) showed the gap holding at +40pp on Codeforces. With Phase 11e multi-site, the corollary is *partially* rehabilitated: it holds at L43 turn_end (gap collapsed $+40 \rightarrow +7$pp), is too weak at L23/L31 pre_tool (gap held), and is *backwards* at L55 pre_tool (saturation flips direction, not just magnitude). The saturation-direction principle is the more general truth; saturation-magnitude is one of three regimes within it.

Practical safety implication: any deployment that treats the $\alpha=-100$ capability lever as a generic intervention will hit site-specific failure modes. The lever is robust where the residual encodes a stable saturation direction across distributions. It is unstable where the saturation direction depends on the test distribution. Production probe deployments need site-by-site robustness measurement, not aggregate guarantees.

5.6 The 4 sanity checks save publishable claims

Each of the four mandatory sanity checks caught a confident wrong claim during this work:

- **Random K-matched** (paper-3 §3.1): caught FG L31 as pure epiphenomenal under proper baseline (Phase 10).
- **Control-token normalization** (paper-3 §3.2): caught Phase 7 L43 as softmax-temperature artifact, not target-specific lever.
- **Structural-rigidity α-sweep** (paper-3 §3.3): caught Phase 8 L55 thinking as template-lock, not amplitude-bound null.
- **Whitespace-stripped flip metric** (paper-3 §3.4, Phase 10): caught RG $\alpha=+200$ raw 96% as 64pp inflation by leading-space artifact; stripped value 32%.

Without all four, this paper would have at minimum 3 falsified findings. The discipline is net-positive at every step.

---

6. Limitations

- **Single model**: all 8 probes tested on Qwen3.6-27B only. Replication on Gemma-2-2B-IT, Llama-3.x, or Claude-class models is paper-5 v2 work.
- **Layer/position grid is coarse**: 5 layers × 5 positions = 25 candidate sites. The saturation-direction lever may exist at intermediate layers we did not sample.
- **Test prompts are domain-specific**: capability tested on HumanEval/MBPP (Python coding); persona on TruthfulQA. Cross- distribution robustness (paper-5 follow-up §7) is in progress.
- **Greedy decoding masks small effects**: sampled generation at T=1 would expose probe-causal effects below the argmax threshold. Not done here due to compute (would require 5-10× compute).

- **Probe AUROC = 1.000 for persona at N=60**: above the over-parameterization threshold flagged in paper-3 §3.1 for K=10 N=60. We supplement with random K=10 baseline (gap quantified) but acknowledge that persona's near-perfect classifier may be partly N artifact. The behavioral lever finding (32% pushdown specific) does not depend on AUROC value.
- **Saturation-direction theory is a heuristic, not a mechanistic derivation**: we propose it as the simplest unification of the data, not as a derived first-principles result. The mechanistic origin — whether it reflects circuit-level redundancy, manifold geometry, or attention-head saturation — is open.

---

7. Conclusion

We mapped 8 probes across 5 empirical classes of probe-causality regime in Qwen3.6-27B, and identified a single unifying principle — *probes lever in the saturation direction of the baseline residual* — that explains all five asymmetric-lever cases. The principle was arrived at via explicit falsification of an earlier categorical-vs-continuous frame using a persona-switch experiment that produced the opposite direction of the naive prediction. Three cross-distribution validations (BigCodeBench at Qwen pass-rate ∼55%, Codeforces at ∼7%, and a four-site multi-locus extension on Codeforces) revealed that the $\alpha=-100$ robustness is **site-partitioned**: at decision-bottleneck pre_tool positions in early-to-mid layers (L23, L31), the pushdown gap holds saturation-independently across distributions spanning $\sim 12\times$ pass-rate variation; at late-layer or non-pre_tool positions (L43 turn_end, L55 pre_tool) the lever shows saturation-magnitude or saturation-direction coupling. The most informative sub-finding is L55 pre_tool's *direction reversal*: pushdown on the saturated distribution, pushup on the unsaturated distribution at $\alpha=+200$. The lever pushes against the baseline saturation, and when saturation flips, the lever flips — saturation-direction theory's central claim, expressed as data. The $\alpha=-100$ robustness theorem is the special case in which the saturation direction itself is stable across distributions. We release all 8 capture batches, all per-site verdicts, and the unified protocol under Apache-2.0.

The four sanity checks (paper-3 §3.1–§3.4) are mandatory rather than optional; three of the four caught confident-but-wrong claims during this work. We pre-registered and walked back two predictions in this paper — categorical-vs-continuous lever class (Phase 12 persona) and the saturation-magnitude corollary (Phase 11d single-site) — and the multi-site extension (Phase 11e) further refines the surviving claim to a *site-partitioned* form. We invite the community to apply the protocol to other probes and other models, and to test whether the site-partition (saturation-independent vs saturation-magnitude-coupled vs saturation-direction-coupled) replicates across model architectures and behavior classes beyond capability.

Reproducibility Statement

All artifacts public under Apache-2.0:

Component	Location
Phase 7 protocol script	`openinterp-swebench-harness/scripts/phase7_steering_micro_pilot.py`
Phase 8 notebook	`notebooks/nb_swebench_v9_phase8_causal_cot.ipynb`
Phase 10 FG/RG notebook	`notebooks/nb_swebench_v10_fg_rg_causality.ipynb`
Phase 11 capability locus notebook	`notebooks/nb_swebench_v11_capability_locus.ipynb`
Phase 11b capability extension notebook	`notebooks/nb_swebench_v11b_capability_locus_extension.ipy`
Phase 12 persona-falsifier notebook	`notebooks/nb_swebench_v12_persona_falsifier.ipynb`
Phase 11c cross-distribution notebook (BCB)	`notebooks/nb_swebench_v11c_cross_distribution.ipynb`
Phase 11d cross-distribution Round 2 (Codeforces)	`notebooks/nb_swebench_v11d_codeforces.ipynb`
Phase 11e multi-site Codeforces validation	`notebooks/nb_swebench_v11e_multisite_cf.ipynb`
Phase 6 N=99 capture corpus	Drive `swebench_v6_phase6/` (99 traces, 89/10 labels, ∼12k captures)

Component	Location
Per-site verdict JSONs	Drive `phase7..phase12/*verdict.json`, `phase11c..11e_*/verdict.json`
Causal locus protocol spec	`openinterp-swebench-harness/paper/paper5_causal_locus_protocol.md`
Meta-analysis of probe AUROCs	`paper/paper5_causal_locus_meta_analysis.md`
openinterp SDK	`pip install openinterp` (v0.3.1+)

Total reproduction time on RTX 6000 Blackwell: ∼6.5 hours from cold start (Phase 6 capture replay 30min via cached) to all 7 phases verdict tables (Phase 11c cross-distribution adds ∼25min).

Acknowledgments

We thank the Qwen team (Alibaba) for releasing Qwen3.6-27B, the Anthropic alignment team for Persona Vectors and the `Teaching Claude Why` framing of eval-distribution overfitting, and the OpenAI alignment team for the `Accidental CoT Grading` audit. Compute provided by Google Colab Pro+ (RTX 6000 Blackwell, A100 40GB).

References

Anthropic. (2025). *Persona vectors: Identifying and modulating personality traits in language models*. Anthropic Research Blog.

Anthropic Alignment Team. (2026). *Teaching Claude Why: Principle-based training generalizes better than behavioral imitation*. Anthropic Alignment Research. https://alignment.anthropic.com/2026/teaching-claude-why/

Belrose, N., et al. (2024). *Tuned lens*. Probes can predict outputs without being causal.

Cobbe, K., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (GSM8K).

Lindsey, J., Cunningham, H., et al. (2024). *Crosscoders for cross-checkpoint model diffing*. Anthropic.

Marks, S., et al. (2024). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

OpenAI Alignment Team. (2026). *Accidental Chain-of-Thought Grading: Audit and Monitorability Analysis*. OpenAI Alignment Research. https://alignment.openai.com/accidental-cot-grading/

Phang, J., et al. (2026). *Qwen3.6 technical report*.

Templeton, A., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Research*.

Yap, J., et al. (2026). *SAE-decoded steering*. Recovering causal authority via SAE features when linear directions fail.

Submitted to: NeurIPS 2026 Mechanistic Interpretability Workshop Status: working draft, paper-5 of openinterp.org series. Code & data: https://github.com/OpenInterpretability/openinterp-swebench-harness Apache-2.0.