Contents

Two Forms of Epiphenomenal Probes in Code Agents

Mid-Reasoning Capability and Chain-of-Thought Emission in Qwen3.6-27B

Workshop draft for NeurIPS 2026 Mechanistic Interpretability Workshop. Apache-2.0. Reproducible. Single-author submission, double-blind by ICML/NeurIPS conventions.

Abstract

We train linear probes on residual-stream activations of Qwen3.6-27B during agent rollouts on SWE-bench Pro and probe-gated retrieval on HotpotQA. We obtain two correlative findings, both validated against random-feature baselines at matched capacity: (a) tool-success at L43 pre_tool position achieves AUROC 0.83 (gap +0.17 above random) at K=10 features and N=54 traces, and (b) chain-of-thought emission at L55 last-prompt position achieves AUROC 0.848 (gap +0.147) at K=5 features and N=240 prompts. To test causal status, we run intervention experiments on each. Both fail, but for distinct reasons. The L43 probe direction

adds a uniform softmax-temperature shift, not a target-specific bias — a pattern revealed only when log-prob deltas on the target token are normalized against control tokens at similar baseline log-prob. The L55 probe direction produces zero behavioral change even when α-perturbations exceed the residual norm itself (α=+200 vs $\|h\|$=158), because the chain-of-thought-emission decision is encoded in the chat template's auto-injected `<think></think>` token pair, downstream of any layer at which the residual could be read or modified. We argue this places linear probes in an explicit *epiphenomenal regime* when applied to format-like or template-controlled decisions in instruction-tuned models, contribute three sanity checks (random-feature baseline at small N; control-token normalization for steering; structural-rigidity α-sweep) as standard validations for future probe-causality work, and ship `agent-probe-guard`, an Apache-2.0 SDK that honestly markets the detection-only signal.

---

1. Introduction

Linear probes on transformer residual streams have become a default mech-interp instrument: cheap to train, often surprisingly accurate at predicting downstream outcomes, and superficially compelling as evidence that the network "contains a representation of" some concept of interest. As probes proliferate into deployed systems — from safety classifiers to reasoning monitors — the question of whether a high-AUROC probe direction is a *causal* feature or merely a *correlative* one becomes load-bearing. A probe that detects but does not lever (Belrose et al. 2024) is fine for monitoring; a probe assumed to lever but actually epiphenomenal will silently underperform when used for intervention.

This paper reports two such cases in Qwen3.6-27B, an open-weights 27B-parameter hybrid-attention reasoning model. We located and validated two correlative probes, then ran controlled intervention experiments on each. Both probes failed to produce behavioral change at amplitudes far beyond the typical steering range, but they failed for *different reasons*. The mechanistic distinction matters: it predicts which other probes will fail, and it provides a taxonomy that future work on probe causality should report against.

Our contributions:

1. **Empirical**: two paper-grade correlative probes in Qwen3.6-27B, with random-K-matched baselines: capability prediction at L43 pre_tool (AUROC 0.83, gap +0.17) and CoT-intent at L55 last-prompt (AUROC 0.848, gap +0.147).
2. **Methodological**: three sanity checks that each caught a confident-but-wrong claim during the work — random-feature baseline at small N, control-token normalization for steering, and structural-rigidity α-sweep.
3. **Conceptual**: a two-mechanism taxonomy of probe-causality failure — *softmax-temperature artifact* and *template-locked decision* — derived from converging intervention experiments.
4. **Engineering**: `agent-probe-guard` v0.1, an Apache-2.0 SDK that ships the correlative signal with explicit detect-only semantics.

---

2. Setup

2.1 Model

Qwen3.6-27B (Alibaba, Apr 2026), 64 layers, hybrid GDN + standard attention, bf16 inference. We use the released chat template, which exposes an `enable_thinking` flag controlling whether `<think></think>` tags are auto-injected before the assistant turn.

2.2 Capture pipeline

For each prompt-trace pair, we register forward hooks at five layers {L11, L23, L31, L43, L55} and capture the last-position residual after the prompt encoding completes. All captures are cached as safetensors on Drive for offline probe training and intervention experiments.

2.3 Datasets

- **SWE-bench Pro N=54**: stratified Python problems ran through the `openinterp-swebench-harness` V1 (transformers direct, max_turns=30, temp=1.0). Each trace gets a binary label `patch_generated $\in$ {0, 1}` from the harness G4 audit. Phase 6 N=99 is in flight at submission time; preliminary results from N=54 are reported here.
- **HotpotQA + memory pool N=240**: 60 questions × 4 retrieval conditions (`none`, `ensemble-gated`, `all-admit`, `random-50`). Each prompt gets a binary label `has_think_v1 $\in$ {0, 1}` indicating whether the model continued reasoning past an auto-injected `<think>` token.

2.4 Probe family

Light linear classifiers (`sklearn.linear_model.LogisticRegression`) trained on a subset of residual dimensions. Selection methods evaluated: top-K diff-of-means, random-K sample (matched baseline), L1-LR with full features, PCA-K. All AUROC values reported are 4-fold stratified CV on the labeled set; feature selection is performed *inside* each fold to avoid leakage.

---

3. Methodology — three sanity checks

3.1 Random-feature baseline at small N

In a methodology sweep at N=17 (Phase 5d) we observed an apparent AUROC of 1.000 for a top-K=50 probe at L43 think_start. A random-K=50 baseline at the same N achieved AUROC $\approx$ 0.95. The "perfect" probe was over-parameterized. The signal disappeared into noise at K=10 and below, where it should have if the underlying feature was real.

We adopt the rule: **at N<100, every probe AUROC must be reported alongside a random-K-matched baseline**. The metric is the gap between the probe and the random baseline at the same K, not the probe's raw AUROC. Paper-grade is gap $\geq +0.10$. The full K-sweep at N=42 (Phase 6c) and N=240 (Phase 8 redux) are the values reported in §4.

3.2 Control-token normalization for steering

In a steering experiment at L43 pre_tool (Phase 7 pilot), we observed Δlog-prob(`finish`) $= +0.479$ nats at $\alpha=+2$ — a textbook "the probe direction biases the model toward the target token" finding. When we ran the same intervention but logged Δlog-prob across five control tokens (`search`, `execute`, `write`, `read`, `wait`), all five shifted by $+0.5$ to $+1.0$ nats. The probe direction was producing a uniform softmax-temperature change, not a target-specific bias. The relative shift $\Delta\text{rel} = \Delta(\text{target}) - \text{mean}(\Delta(\text{controls}))$ was -0.046, essentially zero.

We adopt the rule: **any steering claim of the form "the probe direction shifts log-prob of token X by Y" must report Δrel = Δ(X) − mean(Δ(controls)) across at least 5 semantically unrelated control tokens at similar baseline log-prob**. The pattern catches OOD softmax-temperature shifts that masquerade as target-specific causal evidence.

3.3 Structural-rigidity α-sweep diagnostic

When a steering experiment produces zero behavioral change at $\alpha \in \{\$\pm\$2, \$\pm\$5\}$, two scenarios are possible: (a) the perturbation amplitude is too small for the regime, or (b) the decision is structurally inaccessible from the residual at the chosen layer/position. These have different implications and require different next steps.

We adopt the rule: **before declaring a steering null at $\alpha \in$ {$\pm$\$2, $\pm$\$5}, sweep α to amplitudes that exceed the residual norm itself (e.g., α=+50, +100, +200), with both the probe direction AND a random direction**. If output diverges at some threshold, the original null was amplitude-bound and a

properly-scoped sweep recovers signal. If output remains rigid even at $\alpha \gg \|\text{residual}\|$ for both probe AND random, the decision lives outside any representation a residual probe could read or modify — typically in input tokens for format-like choices in templated models. The diagnostic costs <60s of GPU compute and disambiguates the two scenarios.

---

4. Two probes

4.1 L43 pre_tool — trace-success capability

capacity (K)	top-K diffmeans AUROC	random K-matched	gap
5	0.785	0.697	+0.088
10	**0.830**	**0.749**	**+0.080**
20	0.859	0.794	+0.064
50	0.854	0.824	+0.029
100	0.889	0.849	+0.041
200	0.897	0.865	+0.032

Best capacity: K $\in$ {5, 10}. Overparameterization sets in by K=20: random features alone match the probe's gap shrinkage. We use K=10 in the SDK, which gives a slightly sub-threshold gap of +0.08 against random-K=10 but matches the operational latency budget ($\sim$50 ms forward pass with 10-dim probe vs $\sim$200 ms with 5-dim PCA pipeline). PCA-10 at L43 reaches AUROC 0.764 with random=0.495 gap +0.269 in the Phase 6c sweep at N=42; the K=10 diffmeans result above is at the larger N=54 set.

4.2 L55 last-prompt — suppressed CoT intent

capacity (K)	top-K diffmeans AUROC	random K-matched	gap
5	**0.848**	**0.701**	**+0.147** [ok]
10	0.858	0.753	+0.105 [ok]
20	0.862	0.805	+0.057
50	0.867	0.839	+0.028
100	0.896	0.866	+0.030
200	0.904	0.876	+0.028

The signal is genuinely localized in 5 features. Both K=5 and K=10 cross the +0.10 paper-grade threshold; K=5 is the strongest in gap-relative terms. The top-5 dim indices vary slightly across CV folds but the median signed diff-of-means values are dominated by a single dimension with magnitude $\sim 4\times$ the next four combined — the others provide redundant rather than independent signal.

4.3 What the probes seem to read

The L43 probe correlates with whether the agent will produce a working patch within the 30-turn budget. The L55 probe correlates with whether the model *would have* emitted `<think>` had the chat template not auto-injected and closed the thinking block. The latter is most cleanly interpreted as a counterfactual signal: the residual at L55 last-position retains a representation of "this prompt would benefit from CoT" even when the template has overridden that intent.

---

5. Causal experiments

5.1 L43 — log-prob proxy + behavioral generation (Phase 7)

We applied the standard activation-steering protocol at L43 pre_tool: hook the layer's forward pass during the token immediately before the first tool call, add $\alpha \times$ probe-direction to the last-position residual, observe.

Log-prob proxy (single forward, no generation). At $\alpha{=}+2$, log-prob of `finish` shifted by $+0.479$ nats on FAILS traces — apparent causal evidence. With control-token normalization across `search/execute/write/read/wait`:

Subset	Δrel (target − mean controls) at $\alpha{=}+2$	n
FAILS	**−0.046** (std 0.093)	6
SOLVES	**+0.075** (std 0.000)	2

At $\alpha{=}+5$ the FAILS Δrel becomes -0.167 — slightly *anti-finish*. The probe direction produces a uniform softmax-temperature shift, not a target-specific bias.

Single-shot behavioral generation, $\alpha{=}+5$ once, 80 tokens greedy: 4/4 fails select identical tool at $\alpha{=}0$ and $\alpha{=}+5$. The intervention does not redirect behavior even at the magnitude that produced the largest log-prob shift in the proxy.

Continuous behavioral generation, $\alpha{=}+5$ every step, 80 tokens: 3/4 keep their tool selection; 1/4 (b112069e) degenerates into a malformed loop `</think> <tool_call>bash </think>` but does not switch to `finish`. Sustained pressure corrupts coherence without redirecting target action.

Three converging experiments, one verdict: the L43 pre_tool probe direction is *detection-only*. It correlates with capability outcome but does not function as a behavioral lever.

5.2 L55 — bidirectional steering + amplitude diagnostic (Phase 8)

Following the same protocol at L55 last-prompt, with `enable_thinking=False` chat template (which auto-injects `<think>\n\n</think>\n\n` and asks the model to skip CoT). 8 test targets in `ensemble-gated`: 4 pushup (baseline=False, expect to flip ON with $\alpha{>}0$), 4 pushdown (baseline=True, expect to flip OFF with $\alpha{<}0$). Bidirectional sweep $\alpha \in \{-5, -2, 0, +2, +5\}$ on probe direction; $\alpha \in \{\$\pm\$2\}$ on random direction.

32 of 32 generations identical char-by-char, both probe and random directions. has_think rate 0% across all conditions — because the `<think></think>` empty pair is in the prompt itself before generation begins.

Amplitude diagnostic on a single target (test_4032, baseline=False):

α (probe direction)	residual norm pre$\rightarrow$post	$\Delta\|\cdot\|$	output
0	157.73 $\rightarrow$ 157.73	0.00	baseline
+5	157.73 $\rightarrow$ 157.79	5.00	identical
+20	157.73 $\rightarrow$ 158.91	20.01	identical
+50	157.73 $\rightarrow$ 165.27	50.00	identical
+100	157.73 $\rightarrow$ 186.41	100.00	identical
+200	157.73 $\rightarrow$ 254.25	200.04	identical

The hook fires; the modification propagates; the residual at L55 last-position is genuinely perturbed by 200 (27% above the original $\|h\|{=}158$); the next-token argmax never flips. Random direction at $\alpha \in \{+20, +50, +100\}$ produces the same identical output as baseline.

The mechanism is not amplitude. It is structural: the `<think></think>` pair is in the prompt token sequence. The "no thinking" decision was made by the template before any layer's residual was computed. Steering the residual at last-position cannot retroactively un-emit those tokens.

5.3 L55 redux — top-5 concentrated direction (rules out dilution)

A first reading of §5.2 raises the dilution hypothesis: maybe the full-LR direction (5120 components L2-normalized to 1) places only ∼1% of its weight on the K=5 paper-grade signal dims, with 99% spread across noise-floor features. Concentrating α=+200 entirely on the 5 signal dims could reach the amplitude needed to flip output.

We re-ran the diagnostic with three directions:

- **TOP-5**: zeros except on the K=5 dims, weighted by signed diff-of-means, L2=1
- **RAND-5**: zeros except on 5 random dims with random Gaussian sign, L2=1
- **FULL-LR**: original 5120-dim L1-LR direction, L2=1

12 generations across 6 alphas × 3 directions: identical to baseline. Even when the entire α=+200 is concentrated on the 5 dims that carry the paper-grade signal, output does not change.

This rules out direction-dilution. The decision lives outside any subspace of the L55 last-position residual that a probe could be trained on.

---

6. Discussion — two epiphenomenal mechanisms

The L43 and L55 findings converge on "probe is detection-only" but via distinct mechanisms. We propose them as the first two entries of a probe-causality failure-mode taxonomy.

6.1 Softmax-temperature artifact (L43 pre_tool)

Adding the probe direction to the residual produces an OOD residual state. At downstream layers and the lm_head, this OOD state most often manifests as a *uniform* shift in log-probabilities — equivalent to a softmax-temperature change. In a naive log-prob-on-target experiment, this looks indistinguishable from a target-specific bias. The control-token normalization in §3.2 is the specific fix.

This mechanism is consistent with the broader finding (Anthropic Persona Vectors 2025; tuned-lens ablation studies) that intermediate-layer interventions often re-render as global softmax adjustments, not as targeted feature-level steering.

6.2 Template-locked decision (L55 thinking)

For format-like decisions in instruction-tuned models — emit `<think>` or not, emit JSON or markdown, etc. — the chat template often encodes the decision as input-token sequence rather than as a runtime decision in the residual. The model's L55 residual at last-prompt-position carries information *about* the decision (which is why the probe achieves high AUROC) but does not control it (which is why steering at any α fails).

The fix here is not methodological — it is to recognize that the probe and the lever are at different layers of the system. Mid-layer steering cannot reach the input-token layer. Future work that wants to lever CoT emission specifically should intervene at *prompt construction* (token-level intervention), not at residual stream.

6.3 Why the two together are a contribution

Either finding alone would be a "probe doesn't lever, again" data point. Together they sharpen the claim: **probe-causality failure is not a single phenomenon, it is at least two phenomena with different signatures and different fixes**. Future probe-causality work should report against this taxonomy, and our three sanity checks (§3) provide cheap diagnostics for each.

We position both findings as instances of the broader eval-distribution-overfitting pattern documented by Anthropic Alignment (2026): an in-distribution metric (probe AUROC on the labelled distribution it was selected against) can pass while a held-out automated audit (behavioral steering at $\alpha \gg \|h\|$, or output divergence under intervention) fails. The gap between detection and lever is itself an automated audit on a

distribution the training signal does not span. The two mechanisms above are concrete instances of this structural risk for any probe shipped as a safety component without an explicit causal test against a control direction at amplitude.

OpenAI Alignment (2026), in their audit of accidental CoT-text grading, observed the same gap at the **behavioral** level: when CoT text was included in RL reward, "surface-level CoT properties were steerable under sufficient pressure, but more specific monitor-relevant shaping was harder to induce". Their finding is consistent with our two mechanisms: surface softmax-temperature shifts (paper §6.1) are reachable from any residual intervention, but the deeper template-locked or input-token-controlled decisions (paper §6.2) are structurally outside the reach of the residual where the probe is read. We thus interpret their empirical observation as behavioral evidence for the two-mechanism taxonomy we identify mechanistically.

---

7. Engineering — agent-probe-guard SDK

We ship the correlative findings as an Apache-2.0 Python SDK, `agent-probe-guard`, available via `pip install openinterp` (v0.3.0+). The SDK provides:

- A two-probe activation gate over Qwen3.6-27B (L43 pre_tool capability + L55 thinking-intent), exposed as `AgentProbeGuard.assess()`.
- Three decision modes: `skip` (capability < 0.20), `escalate` (0.20–0.50), `proceed` (≥ 0.50).
- $\sim$50 ms scoring latency on RTX 6000 / A100 (single forward pass with two hooks).
- Explicit detect-only semantics in the README, the SDK docstrings, and the product landing page. No "boost" mode is provided because we have evidence that boosting would silently underperform.

The HF dataset `caiovicentino1/agent-probe-guard-qwen36-27b` contains the probe weights, the scaler, the dim selections, and a `meta.json` with full eval metrics and provenance. The full reproduction harness is at `OpenInterpretability/openinterp-swebench-harness`.

---

8. Limitations

- **Scale**: SWE-bench Pro N=54 is small. Phase 6 N=99 is in flight; the gap at K=10 may shift by \$±\$0.05 at scale. The L55 N=240 number is more robust.
- **Single model**: All experiments are on Qwen3.6-27B. Cross-model transfer is paper-2 work in progress.
- **Probe family**: Linear probes only. SAE-decoded steering (Yap 2026) and edge attribution patching on SAE features (Marks & Rager 2024) may yet recover causal authority; this is deferred to future work.
- **Domain narrowness**: Code agents and CoT-RAG. Whether the template-locked-decision mechanism generalizes to all format-like decisions in instruction-tuned models is open.

---

9. Conclusion

We presented two correlative-but-not-causal probes in Qwen3.6-27B and diagnosed them as instances of two distinct epiphenomenal mechanisms: softmax-temperature artifact (L43 pre_tool) and template-locked decision (L55 thinking). Three sanity checks — random-K-matched baseline, control-token normalization, structural-rigidity α-sweep — caught confident-but-wrong claims at three points in the work, each in $<$60s of additional compute. The cumulative cost was $\sim$\$11 across all intervention experiments; the cumulative scientific yield was a sharper statement about when probes detect vs lever, plus a deployable Apache-2.0 SDK that markets the correlative finding without overclaiming. We submit the methodology checks and the two-mechanism taxonomy as standing recommendations for any future probe-causality work.

Appendix A. Reproducibility

Component	Location
Capture pipeline	`openinterp-swebench-harness/scripts/build_nb_swebench_v8_nb47b_capture.py`
Methodology sweep	`openinterp-swebench-harness/scripts/nb47b_train_cot_integrity_probe.py`
Random K-matched supplement	`openinterp-swebench-harness/scripts/nb47b_random_k_supplement.py`
Phase 7 micro-pilot	`openinterp-swebench-harness/scripts/phase7_steering_micro_pilot.py`
Phase 8 causal CoT	`openinterp-swebench-harness/scripts/build_nb_swebench_v9_phase8_causal_cot.py`
Eval docs	`openinterp-swebench-harness/paper/preflight_probe_eval_v3..v6.md`
SDK	`openinterp-cli/openinterp/agent_probe_guard.py`
Probe weights	HF dataset `caiovicentino1/agent-probe-guard-qwen36-27b`
Activation captures	Drive `openinterp_runs/nb47b_capture/` (5$\times$240$\times$\$5120 safetensors)

Appendix B. Compute ledger

Phase	Compute	Outcome
Phase 5d (over-param N=17)	∼\$3	caught by random-feature check → eval v2
Phase 6c (methodology sweep)	∼\$2	L43 pre_tool K=10 real signal → eval v4
Phase 7 (3-source causality)	∼\$2	L43 epiphenomenal → eval v5
Phase 8 (CoT bidirectional + diagnostic)	∼\$3	L55 epiphenomenal (structural) → eval v6
Phase 8 redux (random-K + top-5 retest)	∼\$1	structural lock confirmed not dilution → v6 §D.5
Total	**∼\$11**	three sanity checks, two probes, two mechanisms, one SDK

For reference, a single Docker SWE-bench Pro eval on 100 traces is ∼\$15.

Appendix C. Cross-environment probe transfer

A v0.1 SDK eval on a fresh Colab session (Qwen3.6-27B without `flash-attn` or `flash-linear-attention` available) revealed that probe weights are **coupled to the inference environment** they were captured in. Concretely:

Layer	nb47b residual $\|\cdot\|$ (L2)	fresh-Colab residual $\|\cdot\|$	cosine(nb47b, fresh)
L43 last-position	98.2	81.5	0.79
L55 last-position	165.4	161.1	**0.35**

For the same prompt and same released model checkpoint, the L55 last-position residual differs in *direction* between environments (cosine 0.35), with similar L2 norm. The L43 residual is closer (cosine 0.79) but still significantly off. This is consistent with how Qwen3.6's hybrid attention implementation falls back to a torch reference path when fla / flash kernels are missing — producing numerically different but semantically equivalent forward passes.

Concrete impact on shipped probes:

Probe	nb47b env AUROC	fresh-Colab env AUROC	Drop
L55 thinking	0.848	**0.559** (loaded weights) / **0.791** (refit)	-28.9pp / -5.7pp
L43 capability	0.830	**0.759** (loaded weights) / **0.806** (refit)	-7.1pp / -2.4pp

Loading the original probe weights and running them on fresh-Colab residuals collapses the L55 AUROC to 0.56 (essentially random). However, **re-fitting the same top-K diff-of-means probe on 240 fresh-Colab captures recovers the signal**: thinking 0.79, capability 0.81, both above the +0.10 paper-grade threshold against random K-matched baselines.

Conclusion: the underlying signal is real and transferable across environments; the *coordinate-level weights* are not. We document this as a caveat for any deployed probe SDK and ship `AgentProbeGuard.refit(prompts, labels)` (PyPI $\geq$ 0.3.1) as a one-call remediation that captures fresh activations on the user's own model and re-fits both top-K probes.

This matches the broader observation that linear probes can transfer across fine-tuning checkpoints (Belrose et al.) but require recalibration when the forward pass numerics change — even at the same model identity.

Appendix D. Threats to validity

Selection effects in N=54: Phase 1 stratification was over Python repos, not over solve-difficulty. The probe may track repo-specific features rather than general capability. Phase 6 N=99 with leave-one-repo-out CV partially addresses this; full repo-stratification is paper-2 scope.

Confound between has_think_v1 and prompt length: ensemble-gated prompts are 4× longer than `none` prompts because RAG memories are inlined. We verified the within-`none`-only AUROC is 0.81 (vs 0.91 cross-condition), suggesting prompt length contributes ~10pp of the cross-condition signal. The K=5 paper-grade gap survives this correction (within-condition gap +0.07, cross-condition gap +0.147).

Greedy decoding masks small-amplitude effects: all behavioral experiments use `do_sample=False`. Small probabilistic effects below the argmax threshold would be invisible. Sampled generation at temperature 1.0 would expose these, but at the cost of run-to-run variance that the ~$11 budget couldn't absorb.

References

(To be filled in for camera-ready. Key prior art:)

- Anthropic Persona Vectors (Aug 2025) — softmax-temperature artifacts in steering, residual-stream interventions.
- Anthropic Alignment Team (2026), *Teaching Claude Why* — eval-distribution overfitting at the alignment-training level; held-out automated auditing catches what training-time metrics miss. https://alignment.anthropic.com/2026/teaching-claude-why/

- OpenAI Alignment Team (2026), *Accidental Chain-of-Thought Grading* — behavioral audit of CoT text in RL reward; "surface CoT steerable, deeper monitor-relevant shaping harder" empirical pattern matching the two-mechanism taxonomy of this paper. https://alignment.openai.com/accidental-cot-grading/
- Belrose et al. (2024), tuned-lens — probes can predict outputs without being causal.
- Marks & Rager (2024), edge attribution patching — feature-level intervention on SAE features.
- Yap et al. (2026), SAE-decoded steering — recovering causal authority via SAE features when linear directions fail.
- Hubinger et al. (2024), activation oracles — detection-quality features downstream of decision-relevant features.
- Phang et al. (2026), Qwen3.6 technical report — base model and chat template specification.
- Yang et al. (Apr 2026), Qwen-Scope — official SAEs for Qwen3.x base/MoE.

Submitted to: NeurIPS 2026 Mechanistic Interpretability Workshop Status: working draft, under review by author. Camera-ready by Sep 2026. Code & data: https://github.com/OpenInterpretability/openinterp-swebench-harness