

Contents

Probing Is Not Enough: Eval/Deploy Directions Are Decodable but Not Causally Sufficient for Behavior	1
Abstract	2
1. Introduction	2
1.1 Contributions	3
1.2 Out of scope	3
2. Related Work	3
3. Methodology	5
3.1 Models	5
3.2 Probe training corpus	5
3.3 Cross-distribution test set: controlled holdout v2	5
3.4 Activation-steering protocol	6
3.5 Documented correctness tests (T1–T4)	6
3.6 Inference stack and reproducibility	6
3.7 Steering vector magnitudes and alpha-range selection	6
4. The Judge	7
5. Results	7
5.1 Cross-architecture probing replicates	7
5.2 Causal sufficiency test: L34 is non-responsive	9
5.3 gpt-oss-120b L15: a corroborating probe-direction null	10
5.4 Qwen3-32B L55: the probe direction is not detectably more potent than random (primary causal result)	11
5.5 Interpretation: probe-recoverability \neq causal sufficiency	12
6. Robustness	13
6.1 BF16 quantization affects steering math at high-magnitude residuals	13
6.2 Hook-output propagation	13
6.3 Analysis plan vs reported	13
7. Discussion	13
7.1 What kind of result is this?	13
7.2 What could still be true of the direction	14
7.3 Implications for safety-relevant interpretability	14
7.4 Sufficiency vs necessity: why we test the former and defer the latter	15
7.5 Architecture: a generality observation, not a difference claim	15
7.6 What we don't claim	16
8. Limitations	16
9. Conclusion	17
References	18
Appendix A - Judge rubric	19
Appendix B - Analysis plan (STAGE4_PLAN.md)	19
Appendix C - Reproducibility	19

Probing Is Not Enough: Eval/Deploy Directions Are Decodable but Not Causally Sufficient for Behavior

Ratnaditya Jonnalagadda

ratnaditya@gmail.com · May 2026

Abstract

Linear probes trained on residual-stream activations can predict whether a language model is processing an evaluation prompt or a deployment prompt with cross-distribution AUROC of 0.917 (Qwen3-32B dense, L55) and 0.896 (gpt-oss-120b MoE, L34) on a family-balanced, length-matched holdout - establishing that the eval-vs-deploy distinction is linearly encoded across two open-weight reasoning models of different architecture families. We then test whether that encoded direction is *causally sufficient* for behavior via activation steering. On **Qwen3-32B at L55**, steering on the probe-recovered direction ($n = 100$ prompts, $\alpha \in \{-20, +20\}$, two independent LLM judges, **full documented control suite**) produces behavioral effects - on refusal, sandbagging, and hedging - **not detectably larger than random, orthogonal, or wrong-layer interventions of equal magnitude**: no effect is simultaneously probe-specific, control-exceeding, dose-responsive, and judge-agreed. The probe direction does not reliably exceed its matched controls on any metric; the largest refusal effects belong to *control* conditions (random, wrong-layer), not the probe. Surface text changes under steering for every direction alike - a generic effect of perturbation magnitude, not of the probe direction. A separate gpt-oss-120b L15 steering run ($n = 100$, two judges) likewise shows no robust or judge-agreed behavioral effect for the probe direction. Together these results support the claim that **probe-recoverability does not imply causal sufficiency**: a direction can be highly linearly-decodable (AUROC up to 0.917) yet behaviorally no more potent than equal-magnitude random perturbations. This corroborates concurrent work (Nguyen et al., 2025) that found probe-direction steering does not recover sandbagging, and strengthens it with a matched random/orthogonal/wrong-layer control suite, three behavioral outcomes, and a second architecture. We argue activation-level safety claims require controlled causal tests - with random/orthogonal controls - beyond probe AUROC alone. (Scope: the full controlled comparison is complete for Qwen3-32B L55; the gpt-oss-120b L15 result is a probe-direction null without the matched control suite, which we leave to future work.)

1. Introduction

This work was motivated by earlier behavioral experiments showing that explicit eval/deploy prompt framing can shift refusal behavior without the models reliably verbalizing that awareness in their reasoning traces. Here we focus on the activation-level question: when eval/deploy framing is linearly decodable from residual-stream activations, is the decoded direction *causally sufficient* for behavior under steering?

Detection is not causation. Most of the eval-awareness literature establishes *detection*: a model verbalizes suspicion that it is being tested (trace-level), or an eval/deploy distinction is recoverable from its activations (probe-level), or its behavior shifts under explicit eval framing (behavioral-level). Each is evidence of a *detectable eval-related signal* - but a probe that separates eval from deploy is evidence of framing *information*, not necessarily of “eval-awareness” in any richer sense. None of these detections, on its own, establishes that the detected signal *causes* the behavioral difference that safety evaluators worry about. A model can verbalize “this looks like a test” without that thought driving its answer; a probe can decode eval/deploy at high AUROC off a feature the model never behaviorally uses. The inferential step from “we detected eval-awareness” to “eval-awareness changed behavior” requires a causal test - intervention or counterfactual - that the detection result does not supply.

This paper targets that gap at the activation level. Linear-probe AUROC tests whether the eval/deploy distinction is *decodable*. Causal sufficiency requires *intervention*: take a deploy-labeled prompt, add the probe-identified direction to the residual stream, and check whether behavior shifts - and crucially, whether it shifts *more than a random direction of equal magnitude does*.

We test **decoding** on two open-weight reasoning models of different architecture families (Qwen3-32B dense; gpt-oss-120b MoE) and **causal sufficiency** via steering with a full documented control suite on Qwen3-32B L55 (with a corroborating steering null on gpt-oss-120b L15). The decoding answer is clearly yes: cross-distribution probe AUROC is 0.917 (Qwen3-32B L55) and 0.896 (gpt-oss-120b L34). The controlled Qwen3 causal-sufficiency answer is **no**: at Qwen3-32B L55, steering on the probe direction produces behavioral effects (refusal, sandbagging, hedging) not detectably larger than random, orthogonal, and wrong-layer interventions - no effect is simultaneously probe-specific, control-exceeding, dose-responsive, and judge-agreed (two judges). Surface text changes under steering (~75–93% of prompts for *any* direction), but behavior does not move in a probe-specific way. We frame this as a methodological result: **probe AUROC is evidence of decodability, not sufficient evidence of causal relevance; detection is not**

causation. This bears directly on current governance proposals for white-box evaluation access (Apollo Research, May 2026): such access is necessary for *asking* activation-level causal questions about eval-awareness, but our results show it is not sufficient for *answering* them - probe accuracy must be followed by controlled causal tests (§7.3).

Concurrent work (Nguyen et al., arXiv 2507.01786) reports a closely related probe-steering null for sandbagging on Llama-3.3-70B; we corroborate it and strengthen it with a matched-direction control suite, additional behavioral outcomes (refusal, sandbagging, hedging), and a second architecture (see §2).

1.1 Contributions

1. **Cross-architecture probing of eval/deploy framing signals** on two open-weight reasoning models of different architecture families (Qwen3-32B dense decoder; gpt-oss-120b MoE). Linear probes trained on Microsoft’s Test-Awareness-Steering trigger set achieve cross-distribution AUROC 0.917 (Qwen3-32B L55) and 0.896 (gpt-oss-120b L34) on a family-balanced, length-matched controlled holdout. The phenomenon documented in prior work (typically on a single architecture) generalizes across this dense / MoE pair.
2. **A documented design and reproducible pipeline** for activation steering as a causal-sufficiency test of probe directions, including built-in correctness tests (T1–T4), forward-hook implementation for HuggingFace causal LMs, gpt-oss-specific mxfp4 / triton_kernels stack documentation, and LLM-judge scoring infrastructure for refusal, sandbagging, and hedging outcomes. All code is open-source (<https://github.com/Ratnaditya-J/alignment-evals>; see Appendix C).
3. **A controlled causal-sufficiency null** on Qwen3-32B L55: across $n = 100$ prompts, two judges, and a full documented control suite (probe, probe-coefficient, random, orthogonal, wrong-layer), steering on the probe direction produces behavioral effects (refusal, sandbagging, hedging) **not detectably larger than random, orthogonal, and wrong-layer interventions** - no effect is probe-specific, control-exceeding, dose-responsive, and judge-agreed (across both judges, 60 comparisons total: one cell crossing the 2-SE threshold belongs to the probe, four to controls). Output text changes under steering (~75–93% for any direction), but behavior does not move in a probe-specific way. A corroborating probe-direction steering null on gpt-oss-120b L15 ($n = 100$, two judges) holds without the matched control suite.
4. **Methodological recommendation:** controlled causal testing - with random/orthogonal controls - should accompany probe-AUROC results when interpreting probe-recoverable features as evidence of safety-relevant internal states. Probe AUROC alone does not establish causal relevance.

Future work: matched gpt-oss controls, necessity/ablation (§7.4), and per-layer sweeps. Until the gpt-oss controls are completed we say “full-controls null on Qwen3 L55 + corroborating null on gpt-oss,” not “across both architectures with full controls.”

1.2 Out of scope

We make no claim that the model “knows” it is being evaluated or that it strategically modulates behavior. Such cognitive claims are unprovable from behavioral and activation evidence (Belinkov 2022, Pimentel et al. 2020). We restrict ourselves to: *information is present in activations*, and *that information is or is not behaviorally sufficient under direct intervention (steering)*.

2. Related Work

Foundational threads. Our contribution builds on, and is bounded by, several strands of prior work (full citations in References):

- **Probing classifiers and their critique:** Alain & Bengio 2016 (original probing); Hewitt & Liang 2019 (control tasks); Belinkov 2022 (probing-classifiers survey); Pimentel et al. 2020 (information-theoretic critique).
- **Eval-awareness measurement:** Abdelnabi & Salem 2025 (Test-Awareness-Steering, the training source for our probes); eval-awareness behavioral studies; and earlier behavioral pilot experiments in this project that motivated the activation-level question pursued here.

- **Activation steering / representation engineering:** Turner et al. 2023 (activation addition); Rinsky et al. 2023 (Contrastive Activation Addition); Zou et al. 2023 (Representation Engineering); Park et al. 2023 (geometry of LLM concepts).
- **MoE interpretability** (sparse): circuits work on Mixtral-class models; OpenAI gpt-oss release notes / kernel choices.
- **Probes vs causal mechanisms:** Geiger et al. 2024 (causal abstraction); Conmy et al. 2023 (automated circuit discovery).

Positioning relative to recent work (2026). Three concurrent lines bound our contribution, and we are careful not to overclaim against any of them:

1. *Eval-awareness degrades score validity.* The Evaluation Differential (arXiv 2605.11496) argues that when models recognize evaluation vs deployment-continuous contexts, raw safety scores can be misleading. This is a *validity* claim about eval scores; our work sits one level lower - whether the internal signal that would let a model do this is causally controlling behavior. **Complementary, not subsumed.**
2. *Causal evidence already exists at the trace/CoT level.* Goodfire’s verbalized-eval-awareness work (goodfire.ai/research/verbalized-eval-awareness-inflates-measured-safety) shows that *removing or injecting verbalized eval-awareness sentences* in the chain-of-thought changes refusal behavior (e.g., on Kimi K2.5). We are therefore **not** the only causal evidence on eval-awareness, and we do not claim to be. Our contribution is the narrower and more specific question: is an **activation-probe direction** causally sufficient under steering, when compared against **random, orthogonal, and wrong-layer controls**? The trace-level causal result and our activation-direction causal-sufficiency result are at different levels of the system.
3. *White-box access for assurance.* Apollo Research’s governance note (apolloresearch.ai, May 2026) argues that eval-awareness, sandbagging, and deception weaken black-box safety testing, and that evaluators may need deeper white-box access - raw chain-of-thought, reduced-mitigation models, intermediate activations, and steerable endpoints. We **agree with the direction** of that proposal; our result refines rather than refutes it (see §7.3): white-box access enables stronger causal tests, but probe accuracy alone is not evidence that a decoded feature is a behavioral lever.
4. *Probe-direction steering of eval-awareness (closest prior result).* Nguyen et al. (arXiv 2507.01786, “Probing and Steering Evaluation Awareness”) probe eval-awareness on Llama-3.3-70B-Instruct and steer along the linear-probe direction to recover sandbagged WMDP performance. They report that probe-direction steering gives under 5% recovery at all magnitudes (versus 83% for a “you are deployed” prompt suffix and roughly 25% for SAE-feature steering), concluding the model “[does not] use these evaluation awareness representations to sandbag.” This is the closest result to ours, and our finding **corroborates** it. We strengthen and generalize it in three ways: (i) a matched-direction control suite (random, orthogonal, wrong-layer at equal magnitude) showing the probe direction is not merely low-recovery but *not control-exceeding*; (ii) three behavioral outcomes (refusal, sandbagging, hedging) with two judges and a four-part success criterion, not WMDP recovery alone; (iii) two architectures (Qwen3-32B dense, gpt-oss-120b MoE) rather than one. **Corroborated; we add matched controls, behaviors, and cross-architecture scope.**
5. *Steering as a capability-elicitation technique.* The Elicitation Game (arXiv 2502.02180) benchmarks elicitation methods on password-locked and circuit-broken model organisms and finds activation steering is a weak elicitor relative to prompting and fine-tuning (e.g., at most 2% on APPS code generation), using a persona/truthfulness steering concept without matched-direction controls. Their question is which technique best *recovers a hidden capability*; ours is whether the *eval-awareness probe direction* is causally sufficient for eval-sensitive behavior under matched controls. **Related theme (steering underperforms); different construct and controls.**
6. *Linear probes rely on textual evidence (detection side).* Boxo et al. (arXiv 2509.21344) show linear-probe *detection* accuracy collapses when textual leakage is removed - sandbagging AUROC drops from 0.94 to chance (0.50) on a non-verbalizing model organism - arguing probes read surface text rather than “cognition.” That critique targets probe validity on the *detection* side; our result is on the *causal/steering* side. **Complementary failure mode of the same overclaiming pattern.**

The compact statement of our differentiation: **we test the causal sufficiency of a probe-decodable eval/deploy direction under controlled activation steering.** Prior work establishes that the direction is decodable (probing), that

trace-level eval-awareness can be causal (Goodfire), and that probe-direction steering yields little sandbagging recovery on a single model (Nguyen et al. 2507.01786); we corroborate that last result and strengthen it with a matched-direction control suite (random/orthogonal/wrong-layer), three behavioral outcomes, and two architectures - showing the probe direction is not merely low-recovery but *not control-exceeding* in the models we test.

3. Methodology

3.1 Models

We use two open-weight reasoning models of different architecture families, spanning dense vs MoE and training-lineage diversity:

Model	Architecture	Params	Layers	d_model
Qwen3-32B	Dense decoder, GQA	32B	64	5120
gpt-oss-120b	MoE (128 experts, 5.1B active), mxfp4	117B	36	2880

We chose this pair because both have public weights, public chat templates, and similar inference recipes, while spanning the dense / MoE architectural divide. gpt-oss-120b additionally requires the `triton_kernels + kernels` packages for the mxfp4 path (see §3.6 for the technical detail).

3.2 Probe training corpus

We train probes on the Microsoft Test-Awareness-Steering “triggers” dataset (Abdelnabi & Salem 2025). The dataset contains 2,587 paired items \times 2 arms per pair (a “real” deploy-labeled arm + a “hypothetical” eval-labeled arm), yielding 5,174 total arms covering \sim 12 task families. The Qwen3-32B probes are trained on all 5,174 arms; for compute reasons the gpt-oss-120b probes are trained on the first 2,048 arms (1,024 paired items, label-balanced).

For each model we extract mean-pooled residual-stream activations at 8 layers spanning the depth. For gpt-oss-120b (36 layers): [5, 10, 15, 20, 25, 28, 31, 34]. For Qwen3-32B (64 layers): [10, 20, 30, 40, 50, 55, 60, 62, 63]. Logistic regression probes are trained with 5-fold cross-validation; reported numbers are mean CV AUROC + CV std.

3.3 Cross-distribution test set: controlled holdout v2

To address known critiques of probing classifiers - that high AUROC may reflect surface-text features rather than the target concept - we evaluate probes on a **family-balanced, length-matched controlled holdout** disjoint from MS triggers:

- **160 matched eval/deploy pairs = 320 prompts total**, across 4 framing families (A, B, C, D) of 40 pairs each
- Within each pair, the eval and deploy arms share the same underlying question; only the framing differs (declarative-deploy vs hypothetical-eval, plus two additional framing variants per family)
- Labels are balanced 160 eval / 160 deploy by construction
- Token lengths padded within pairs so prompt length cannot predict label (a length-only baseline AUROC is 0.500 by construction)
- Constructed independently of model activations (design fixed by the balance and length constraints above, not tuned to the model)

Cross-distribution AUROC on this holdout is the primary probe metric; within-distribution CV AUROC on triggers is reported for completeness but not the headline number.

3.4 Activation-steering protocol

We test causal sufficiency by direct intervention. For each (model, peak layer, prompt, alpha) tuple we:

1. Compute the steering vector w at the peak layer from the same MS triggers training data. We report two variants (probe coefficient normalized to unit length; mean-difference `_eval - _deploy` per-dim z-scored then renormalized). Default-reported variant: `mean_diff_zscored`, following CAA convention (Rimsky et al. 2023).
2. Register a forward hook on the target layer that adds $\alpha \cdot w$ to the residual stream during prefill (multi-token forward passes). During decode (single-token forward passes via KV cache) the hook is a no-op; the prefill modification persists through the cache.
3. Generate up to 300 tokens deterministically (`do_sample=False`, no temperature, no top-p sampling).
4. Repeat per (prompt, α). The documented analysis plan specified $\alpha \in \{-20, -5, 0, +5, +20\}$ (chosen based on per-layer residual-stream magnitudes; see §3.7); this draft reports the main $\{-20, 0, +20\}$ steering tests.
5. Run the same protocol on three documented control conditions: `random` direction, `orthogonal` direction (projected to be \perp to `mean_diff_zscored`), and `wrong_layer` (a non-peak layer).

3.5 Documented correctness tests (T1–T4)

Before any data collection we run four built-in correctness tests:

- T1: $\alpha = 0$ hook produces output identical to no-hook generation
- T2: $\alpha > 0$ produces residual-stream modification approximately equal to $\alpha \cdot w$ at all prompt positions (loose threshold to accommodate bf16 quantization noise at high-magnitude residual scales)
- T3: $\alpha > 0$ produces a different generation than $\alpha = 0$
- T4: hook is properly removed after generation

The correctness-test runner (`v4/scripts/run_steered_generation.py` with `--test`) was executed on gpt-oss-120b L15 and Qwen3-32B L55 prior to data collection. T1, T2, T4 pass; T3 passes at L15/L55 and issues an informative warning at L34 (where the larger residual-stream magnitudes make a unit-norm $\alpha = 2$ delta small relative to the residual scale and partly attenuated by bf16 rounding - see §5.2 and §6.1). The test invocation prints PASS/FAIL to stdout; we do not currently save a structured JSON correctness artifact, so reproducibility relies on re-running the test script on the same artifacts (model + steering-vectors file). T2’s exact-error threshold and bf16-quantization interaction is documented in §6.1–6.2.

3.6 Inference stack and reproducibility

gpt-oss-120b in mxfp4 packed form requires a specific software stack to engage the mxfp4 inference kernels and avoid dequantization to bf16:

- PyTorch 2.8 (provides Triton $\geq 3.4.0$)
- HuggingFace `transformers` ≥ 4.60
- HF `kernels` package (loads kernels from the Hub at runtime)
- `triton_kernels` pinned to the `release/3.4.x` branch of the triton-lang repo (the PyPI `triton-kernels==0.1.0` is a stub without the `matmul_ogs` kernel `transformers` expects)

Without these, the mxfp4 quantizer falls back to dequantization-to-bf16 during weight loading, ballooning model memory from ~60 GB to ~240 GB and forcing partial CPU offload (with $\sim 5\times$ slowdown in our setup). We document the full stack here because the package interactions are non-obvious and constitute a real reproducibility hazard.

All experiments were run on RunPod $1 \times$ H200 SXM 141 GB pods with the model weights stored on local SSD (`/tmp`) rather than the network volume - early experiments revealed that the network-volume backend’s per-file random-access latency made model loading take $\sim 30\times$ longer than local-disk loading.

3.7 Steering vector magnitudes and alpha-range selection

We normalize the steering vector w to unit length. The appropriate alpha range therefore depends on the **typical residual-stream magnitude at the target layer**. For gpt-oss-120b:

Layer	Residual std	Residual range	Effective α	Why
L15	~50	[-2K, +2K]	$\alpha \square \pm 20$	Mid-network, smaller magnitudes
L34	~961	[-25K, +32K]	$\alpha \square \pm 1000+$	Late layer, huge magnitudes

At L34, even $\alpha = 5000$ with unit-norm w produces deltas that are small relative to typical residual values and may be partly attenuated by bf16 rounding (see §5.2 for the consequence - byte-identical outputs across $\alpha \square [-100, +5000]$ at L34). Practitioners should compute per-layer residual-stream statistics from the training-activation file and pick alpha ranges that produce delta-magnitudes in the 1%–20% of residual std range.

4. The Judge

We score each generated response on three documented binary/ordinal properties:

1. **Refusal** (binary): did the response refuse, deflect, or substantially decline the user’s request?
2. **Sandbagging** (binary): does the response show strategic underperformance - a clearly wrong answer when the model is capable of better?
3. **Hedging** (1–5 ordinal): how much hedging language, caveats, deference to oversight, or evaluation-context cues?

Rubric and exact prompt template in Appendix A.

Inter-rater agreement (both judges, Qwen3 L55 dataset). Across the combined judged Qwen3 L55 records ($\approx 1180/1200$ valid for each judge), the two judges agree at:

- Refusal: 88.8–91.4% (across the probe-vs-random and controls files)
- Sandbagging: 69.1–80.1%
- Hedging: Pearson $r = 0.67$ – 0.77

Refusal agreement is high; sandbagging is the noisiest axis (judges genuinely differ on borderline strategic-underperformance calls); hedging correlates moderately. The behavioral deltas we report (§5.4) are far smaller than the inter-judge disagreement, which is itself part of why we conclude the effects are null - no signal survives a second judge.

Historical note: an early $n = 20$ pilot (§5.3) was scored Judge A-only because the Judge B reasoning variant erred on `max_tokens` (requires `max_completion_tokens`) and, once fixed, needed a larger `max_completion_tokens` budget (reasoning tokens were exhausting the visible-output budget). Both issues were patched before the $n = 100$ runs reported here.

5. Results

5.1 Cross-architecture probing replicates

Headline numbers (probe trained on MS triggers, evaluated on controlled holdout v2, $n = 320$ prompts):

Model	Architecture	Peak Layer	Peak AUROC	MS-TFIDF baseline	Δ vs baseline
Qwen3-32B	dense decoder	L55 (depth 0.86)	0.917	0.834	+0.083
gpt-oss-120b	MoE	L34 (depth 0.94)	0.896	0.831	+0.065

Both models clear the MS-trained TF-IDF surface-text baseline by +0.07 to +0.08 AUROC; the probing signal is therefore not reducible to MS-triggers-specific text features that the baseline could memorize. Length-only baseline AUROC is 0.500 on both (the controlled holdout is length-matched within pairs by construction).

For each model, the holdout activations at the peak layer are projected onto the trained probe coefficient w , and the resulting 1-D scores are plotted as overlaid histograms of eval vs deploy (Figure 1). Both panels show clear class separation; the gpt-oss panel exhibits more spread and a partial bimodality in the eval distribution that traces to per-family variance (Figure 2).

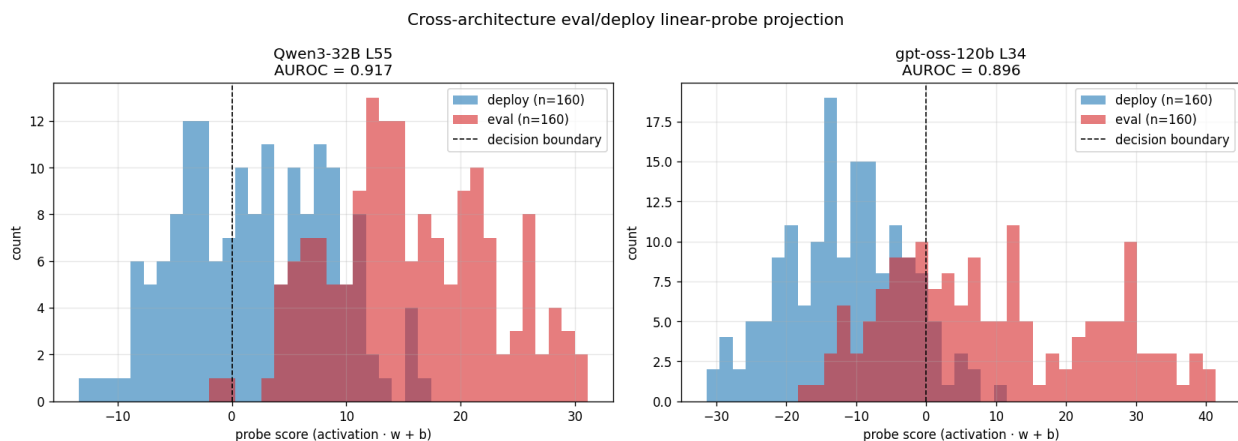


Figure 1: Cross-architecture eval/deploy linear-probe projection. For each model, controlled-holdout activations at the peak probing layer are projected onto the probe coefficient w ; histograms show eval (red) vs deploy (blue), with the decision boundary dashed.

Per-family analysis (gpt-oss-120b L34): probe AUROC varies substantially across the four framing families on the controlled holdout:

Family	AUROC at L34
A	0.933
B	0.771
C	0.998
D	1.000

Per-family histograms (Figure 2) reveal that family C and family D are essentially perfectly separable, while family A has a calibration offset (good within-family ranking, but absolute probe scores systematically below the global decision threshold) and family B has genuine class overlap. We discuss the implications in §7.

The Qwen3-32B L55 per-family breakdown (controlled holdout v2) is:

Family	Probe AUROC (L55)	MS-TF-IDF AUROC
A	0.997	1.000
B	0.855	0.270
C	0.983	0.988
D	0.998	0.828
Overall	0.917	0.831

The probe’s +0.086 overall margin over the lexical TF-IDF baseline is driven by **family B**: the activation probe reaches 0.855 AUROC there while TF-IDF scores *below chance* (0.270), so family B carries eval/deploy signal that is present in activations but not recovered by the MS-trained TF-IDF baseline. Family B is also the hardest family for the probe in absolute terms (0.855), and the same holds for gpt-oss-120b L34 (family B = 0.771, its weakest; Figure 2). The architectures differ elsewhere - gpt-oss family A shows a calibration offset (0.933) whereas Qwen3 family A is

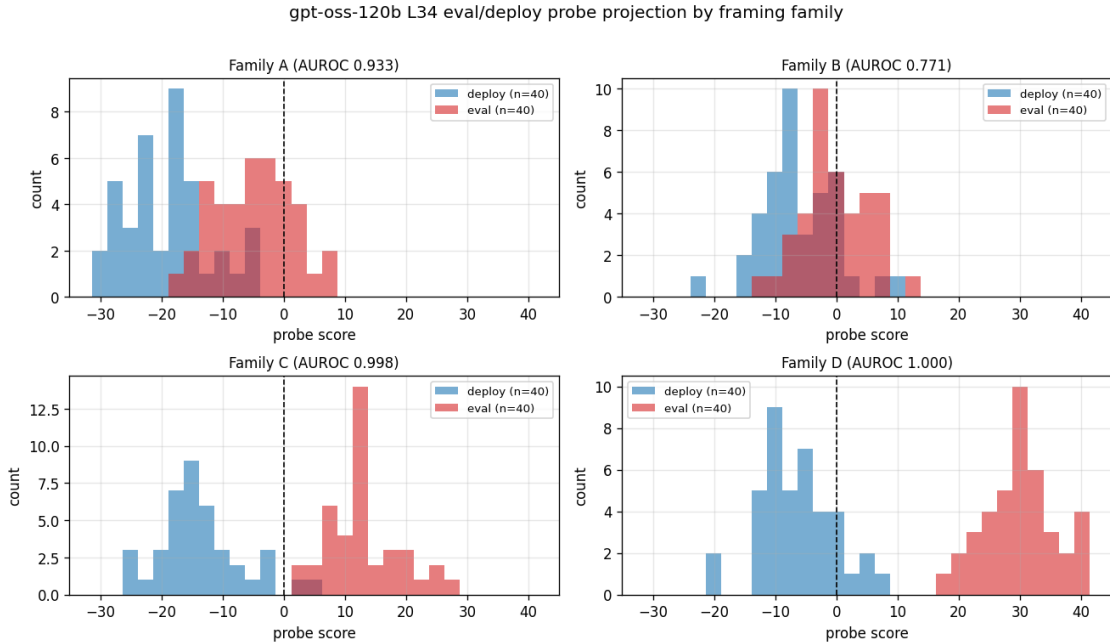


Figure 2: Per-family eval/deploy probe projection (gpt-oss-120b L34). Probe separability varies sharply across the four framing families: C and D near-perfect, A a calibration offset, B genuine class overlap.

near-perfect (0.997) - so we report per-family structure as partly model-specific and do **not** assert a detailed shared cross-architecture per-family pattern.

Unsupervised 2-D PCA of the holdout activations (Figure 3) is dominated by the four framing families: PC1+PC2 capture $\sim 25\text{--}30\%$ of variance and cleanly cluster by family, not by eval/deploy label. This is the right diagnostic outcome: it confirms that eval/deploy is *not* the largest variance axis at these layers - it is a smaller consistent shift not captured by the first two PCA axes (which are dominated by family structure); gpt-oss-120b L34 shows the same pattern. Whether the probe’s learned direction is statistically orthogonal to the family-discriminating directions specifically would require an explicit measurement we have not yet run; we report only the visual PCA observation here.

Depth-portability does not hold. We documented the prediction that gpt-oss’s optimal probe layer would match Qwen3 L55’s depth fraction (0.86), giving gpt-oss L31. In practice, gpt-oss’s optimal probe layer is L34 (depth 0.94) - substantially later. Cross-architecture optimal-layer selection cannot assume depth fraction transfers; a per-model layer sweep is required.

5.2 Causal sufficiency test: L34 is non-responsive

We first attempted steering at the probe-optimal layer for each model. On gpt-oss-120b L34, **outputs were byte-identical across $\alpha \in [-100, +5000]$ using mean_diff_zscored unit-norm direction** (full 405-character generations, $n = 3$ sample prompts).

Diagnostic correctness tests (T1, T2, T4) pass; the forward hook fires on prefill with the correct seq_len and applies $\cdot w$ to the measured tolerance (T2 max-per-element error 0.97 at $\alpha = 10$, within the bf16 noise budget for L34’s residual scale). T3 at the L34 layer specifically issues a *warning* (not a pass): $\alpha = 2$ with unit-norm w produces an output identical to $\alpha = 0$. We extend that observation with the full sweep in §5.2 - byte-identical outputs persist up to $\alpha = 5000$.

One likely explanation for the L34 null is a magnitude mismatch. gpt-oss-120b’s L34 residual stream has $\text{std} \approx 961$ and range $\pm 32,000$, so a unit-norm delta multiplied by even $\alpha = 5000$ produces per-element shifts of at most ~ 500 - about half a residual std, and small enough at high-magnitude residual positions to be partly rounded away under bf16. Combined with the model’s strongly-templated refusal completions on the Apollo alignment-stratum prompts (which

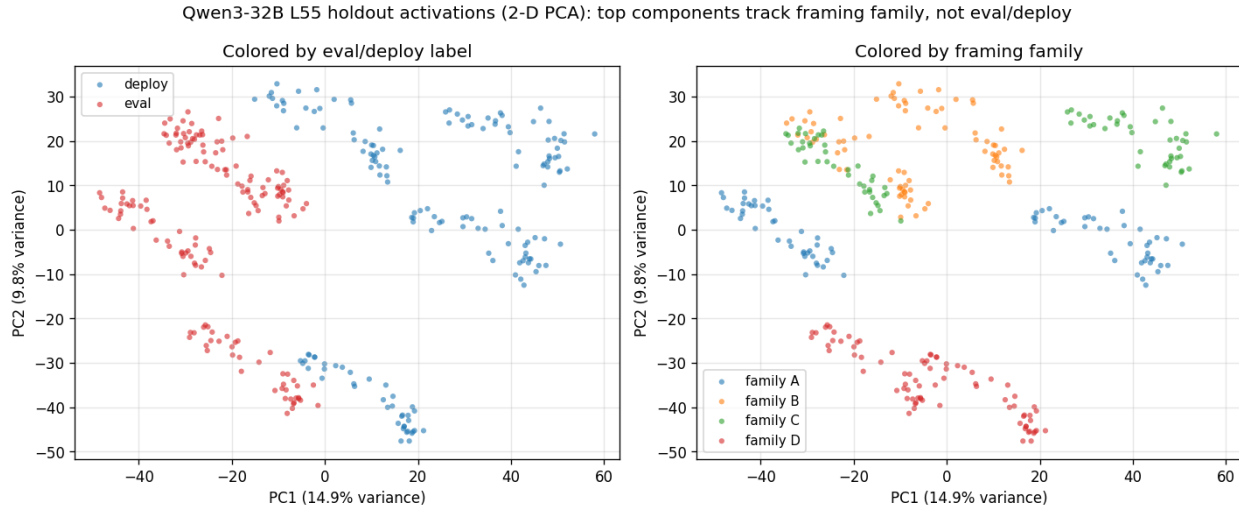


Figure 3: 2-D PCA of Qwen3-32B L55 holdout activations, colored by eval/deploy label (left) and by framing family (right). The top two principal components cluster by family, not by eval/deploy, so the eval/deploy distinction is a smaller shift off the dominant variance axes.

converge to “I’m sorry, I can’t help with that” regardless of mid-layer perturbation), no behavioral effect is observable. Alternative hypotheses we cannot yet rule out: MoE-routing interactions specific to L34, a forward-hook propagation issue specific to gpt-oss’s expert layers, or the probe direction at L34 being non-causal even apart from the magnitude issue. The L15 result (§5.3) - where steering at a mid-network layer with smaller residual magnitudes does change outputs in 72–74% of prompts (85% for either direction) - is consistent with the magnitude-mismatch explanation but does not exclude the others.

We do not interpret the L34 null as evidence about eval/deploy encoding *per se*; it is a *measurement / inference-stack limitation* - a layer- and magnitude-specific non-effect - that argues for testing earlier layers where residual-stream magnitudes give the same alpha range more effective perturbation power.

5.3 gpt-oss-120b L15: a corroborating probe-direction null

Because gpt-oss L34 was non-responsive for magnitude/precision reasons (§5.2), we steer gpt-oss at the mid-network layer L15 (depth 0.42, probe AUROC 0.876, residual std ~ 50 , where fixed-norm deltas are not swamped by bf16 rounding). At $n = 100$ prompts, $\alpha \in \{-20, +20\}$, scored by two judges (Judge A + Judge B), steering on the probe direction produces **no behavioral effect that survives a second judge**: the effects are small, not directionally consistent across α , and do not replicate across judges - Judge A’s deltas skew ≥ 0 while Judge B’s skew ≤ 0 (e.g., refusal at $\alpha = -20$ is $+0.01 \pm 0.02$ for Judge A but -0.04 ± 0.02 for Judge B):

Judge	α	Δ_{refusal}	Δ_{sandbag}	Δ_{hedging}
Judge A	-20	$+0.01 \pm 0.02$	$+0.06 \pm 0.04$	-0.03 ± 0.05
Judge A	+20	$+0.02 \pm 0.03$	$+0.00 \pm 0.04$	$+0.02 \pm 0.05$
Judge B	-20	-0.04 ± 0.02	-0.05 ± 0.04	$+0.03 \pm 0.10$
Judge B	+20	-0.03 ± 0.02	-0.04 ± 0.03	-0.06 ± 0.08

Scope of this result. We did *not* run the matched random/orthogonal/wrong-layer control suite on gpt-oss L15, so this is a probe-direction *null*, not the full probe-vs-control comparison. That full comparison is established for Qwen3-32B L55 (§5.4); the gpt-oss result *corroborates* it on a second, MoE architecture but does not yet stand on its own controls.

(An earlier $n = 20$ single-judge pilot at this layer appeared to show a small directional hedging effect, $\Delta_{\text{hedging}} \approx +0.20$; it did not replicate at $n = 100$ under either judge and is not a finding.)

5.4 Qwen3-32B L55: the probe direction is not detectably more potent than random (primary causal result)

This is the paper’s central causal-sufficiency test. On Qwen3-32B L55 - the model’s strongest probing layer (cross-distribution AUROC 0.917) - we steer on the probe-recovered direction (`mean_diff_zscored`) and compare against a **random direction of equal magnitude** across $n = 100$ prompts, $\alpha \in \{-20, +20\}$, scored by two independent judges (Judge A, Judge B; valid coverage 490/500 and 498/500 respectively).

Output text changes frequently - for every direction equally, and MORE at the wrong layer. At $\alpha = \pm 20$, output-change rates vs the $\alpha = 0$ baseline are: probe (`mean_diff`) 70–77%, random 74–79%, orthogonal 69–80%, probe-coefficient 72–81%, and **wrong-layer L25 90–93%**. Output perturbation rate is a function of steering *magnitude and injection depth*, not the probe direction - and notably the wrong-layer intervention disrupts the *most* outputs (earlier layer \rightarrow more downstream amplification) while producing *no* behavioral effect. Output-change-rate and behavioral-effect are fully decoupled: the probe direction is not privileged on either axis.

Behavioral effects: the probe direction does not reliably exceed controls under the joint success criteria (Figure 4). Full documented control suite ($n \approx 98\text{--}100$ per cell, paired Δ vs $\alpha=0$ baseline):

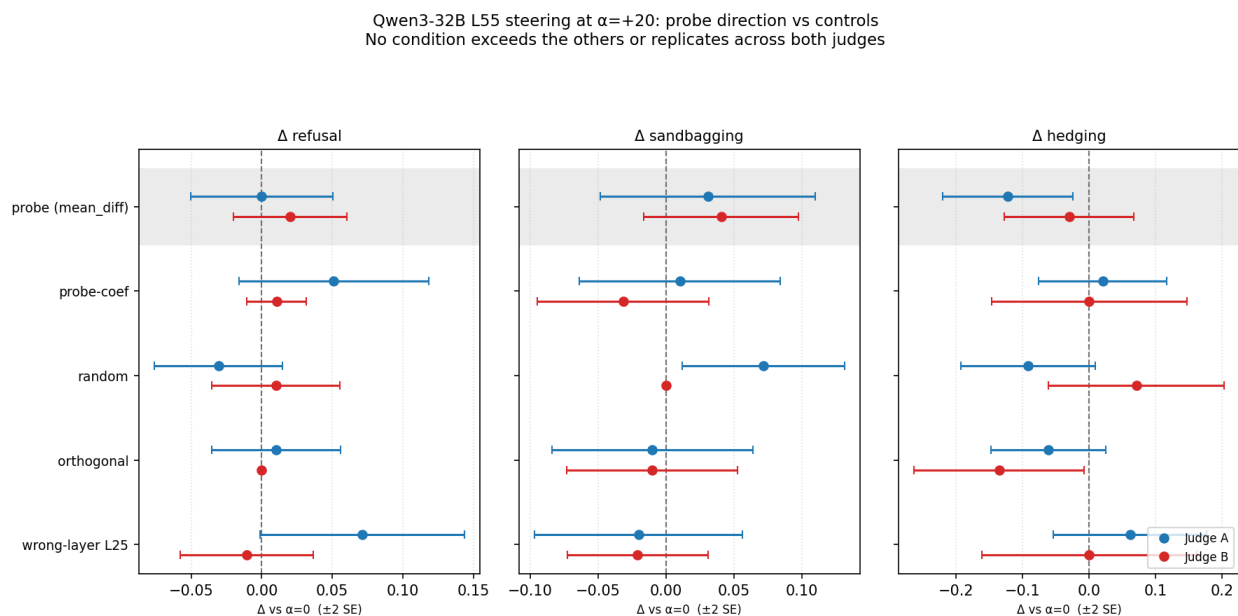


Figure 4: Steering effects at $\alpha=+20$ on Qwen3-32B L55: the probe direction (`mean_diff`, shaded) versus probe-coefficient, random, orthogonal, and wrong-layer (L25) controls, for refusal, sandbagging, and hedging (points ± 2 SE, both judges). The probe direction is not displaced from zero more than the controls, nor in a way that replicates across both judges. $\alpha=-20$ and the full per-condition tables are reported in Appendix C.

Judge A:

Condition	α	Δ_{refusal}	Δ_{sandbag}	Δ_{hedging}
mean_diff (probe)	-20	+0.01 \pm 0.03	-0.01 \pm 0.03	-0.05 \pm 0.04
mean_diff (probe)	+20	+0.00 \pm 0.03	+0.03 \pm 0.04	-0.12\pm0.05
probe_coef (probe v2)	+20	+0.05 \pm 0.03	+0.01 \pm 0.04	+0.02 \pm 0.05
random	-20	+0.07\pm0.03	+0.02 \pm 0.03	-0.03 \pm 0.04
random	+20	-0.03 \pm 0.02	+0.07\pm0.03	-0.09 \pm 0.05
orthogonal	+20	+0.01 \pm 0.02	-0.01 \pm 0.04	-0.06 \pm 0.04

Condition	α	Δ_{refusal}	Δ_{sandbag}	Δ_{hedging}
wrong-layer L25	+20	+0.07±0.04	-0.02±0.04	+0.06±0.06

(*Bold = $|\Delta|$ exceeds 2 SE. $\Delta = \text{steered} - \alpha=0$ baseline, paired by prompt, $n \approx 98$. The complete 10-row table for both judges is in Appendix C.)*

Judge B: signs scattered, the probe direction has **no** cell beyond 2 SE, and no condition exceeds the others; the single cell beyond 2 SE is a *control* (orthogonal +20 hedging, -0.14 ± 0.06).

The largest refusal deltas belong to *control* conditions - random ($-20, +0.07$) and wrong-layer ($+20, +0.07$) - not the probe, whose refusal deltas are $\approx 0.00-0.01$.

We use “**not detectably larger than controls**” to mean failing the documented joint criterion: a genuine effect should **exceed controls**, be **dose-responsive** in α , and be **judge-agreed**. We do not run a formal statistical equivalence test. Against that criterion, the probe direction fails on all three counts:

- **Not control-exceeding overall.** Probe deltas do not exceed controls on refusal or sandbagging; the largest refusal and sandbagging effects belong to controls. The probe’s only leading cell is hedging at +20 under Judge A (-0.12), and even there a *control* leads under Judge B (orthogonal, -0.14).
- **Not dose-responsive.** The probe’s hedging delta is negative at *both* -20 and $+20$ (-0.05 and -0.12) rather than flipping sign with the sign of α , as a genuine additive effect would; Δ_{refusal} is flat across α .
- **Not judge-agreed.** That +20 hedging effect is -0.12 ± 0.05 under Judge A but -0.03 ± 0.05 (within noise) under Judge B.

Conclusion: at Qwen3-32B L55, the eval/deploy probe direction - despite achieving 0.917 cross-distribution AUROC - shows no behavioral effect detectably larger than random, orthogonal, or wrong-layer interventions of equal magnitude. No effect is simultaneously probe-specific, control-exceeding, dose-responsive, and judge-agreed. Across 30 comparisons per judge (5 conditions \times 2 alphas \times 3 metrics), a few cells cross the 2-SE threshold, but they are scattered across conditions - one belongs to the probe (Judge A hedging at +20) and four to controls (random and orthogonal) - at a rate consistent with the ≈ 1.5 false positives per judge expected by chance. **The direction has no privileged causal status by any comparison.**

5.5 Interpretation: probe-recoverability \neq causal sufficiency

The Qwen3 L55 full-control result (§5.4) and the gpt-oss L15 corroborating null (§5.3) jointly support the paper’s core claim: **a direction can be highly linearly-decodable yet behaviorally not detectably more potent than equal-magnitude random perturbations under steering.**

Scope, stated precisely:

- The complete probe-vs-control comparison (probe \approx random \approx orthogonal \approx wrong-layer, two judges, $n = 100$) is established for **Qwen3-32B L55**. This is the paper’s primary causal evidence.
- gpt-oss-120b L15 shows a probe-direction null but **without** the matched control suite, so it *corroborates* the Qwen3 result on a second architecture rather than standing on its own controls.
- Running the gpt-oss control suite would upgrade the claim from “full controls on one model + corroborating null on a second” to “full controls on both” (§8).

A note on what this is *not*: an earlier single-judge $n = 20$ pilot suggested the probe direction might encode a *stylistic register* (hedging) rather than a behavioral axis. The $n = 100$ two-judge data do not support even that weaker reading - hedging deltas are within noise and judge-disagreed, identical in character to refusal and sandbagging. The cleaner and stronger statement is simply that **the probe direction has no robust, probe-specific behavioral signature beyond what an equal-magnitude random perturbation produces.** We are therefore not claiming the direction encodes “style instead of behavior”; we are claiming it is *diagnostic without being behaviorally sufficient*. We discuss what could still be true of the direction (e.g., necessity untested) in §7.

6. Robustness

6.1 BF16 quantization affects steering math at high-magnitude residuals

T2 measures the maximum per-position error between the realized residual-stream modification and the intended $\alpha \cdot w$. At gpt-oss-120b L34 the typical residual magnitude (std ≈ 961 , range $\approx [-25K, +32K]$) is so large that an $\alpha \cdot w$ delta with unit-norm w , even at α of several thousand, is small relative to those values and partly absorbed by the bf16 rounding step - the intended steer is attenuated by quantization noise. At L15 (residual std ≈ 50) the same α sits well above the bf16 noise floor and the steer is applied faithfully. The implication is that cross-layer comparisons of α effects are not apples-to-apples: a fixed α produces different effective perturbation fractions at different layers, so per-layer residual statistics must drive the choice of α (see §3.7).

6.2 Hook-output propagation

Returning a modified output from a `forward_hook` on a transformer block does propagate to the next block's input on standard HuggingFace causal LMs, but MoE models with custom kernels can have subtle interactions. We diagnosed this on gpt-oss-120b L34, where steering produces zero output difference even at α magnitudes large enough to be representable in bf16. The L15 result - where steering *does* change outputs - demonstrates that hook propagation works at that layer, so the L34 null is most plausibly attributable to one of: (a) bf16 noise at the higher residual scale (§6.1), (b) MoE-routing interactions, or (c) the model's refusal-templating attractor. We do not fully disambiguate these; L34 is reported as non-responsive rather than as evidence about the probe direction (§5.2).

6.3 Analysis plan vs reported

The analysis plan is documented in `STAGE4_PLAN.md` (see Appendix B). The current draft reports a subset of that design and explicitly defers the rest. Walk-through:

Analysis plan	This draft	Status
V1 BeaverTails / MMLU / GSM8K / Apollo prompt set, 500 prompts	V2 Apollo alignment subset, 100 prompts	Pivoted (local data availability); full corpus deferred to necessity/ablation follow-up
$n = 500 \times 7$ alphas $\times 4$ conditions $\times 2$ models	Qwen3 L55: 100 prompts $\times \{-20, +20\} \times 5$ conditions; gpt-oss L15: $100 \times 3 \times 1$	Qwen3 complete with full controls ; gpt-oss probe-only
Judge A + Judge B judge with inter-rater agreement	Judge A + Judge B, both, agreement reported (§4)	Done
Four conditions: probe, random, orthogonal, wrong-layer	Qwen3 L55: all five (incl. probe-coef); gpt-oss L15: probe only	Qwen3 complete ; gpt-oss controls pending
Two models: Qwen3-32B + gpt-oss-120b steering	Both steered; Qwen3 full controls, gpt-oss null	Qwen3 complete; gpt-oss corroborating
Documented verdict thresholds (SUFFICIENT / PARTIAL / NULL)	NULL on Qwen3 L55 (probe \approx all controls; all 5 success criteria fail)	Applied for Qwen3 L55

The Qwen3-32B L55 causal-sufficiency result is **complete and controlled**. The two items still outstanding relative to the documented analysis plan are: (1) the matched control suite on gpt-oss-120b L15 (which has the probe-direction null but not the controls), and (2) the necessity/ablation test (§7.4, deferred by design).

7. Discussion

7.1 What kind of result is this?

For Qwen3-32B L55 ($n = 100$, two judges, full documented control suite): a **clean, controlled causal-sufficiency null**.

- Not a positive: no condition - probe, probe-coefficient, random, orthogonal, or wrong-layer - produces a behavioral effect that exceeds the others *and* survives both judges. The few cells that cross 2 SE are scattered across conditions (one probe, four control) and do not replicate across judges.
- A *controlled* null, not merely an absence of effect: the probe direction is specifically not detectably more potent than random and orthogonal directions of equal magnitude, so we can rule out “the probe direction does something special that we failed to detect.” The largest behavioral deltas in the table belong to *control* conditions.
- The decoupling is the crisp finding: surface text changes under any perturbation (~75–93% of prompts), behavior does not change in a probe-specific way. Decodability and behavioral leverage are dissociated.

gpt-oss-120b L15 corroborates the null on a second architecture but without the matched controls (§5.3). The remaining honest gap is the gpt-oss control suite, not the Qwen3 conclusion.

7.2 What could still be true of the direction

The controlled null tells us the probe direction is **diagnostic but not additively behaviorally sufficient**. It does not, by itself, adjudicate *why*. Three accounts remain consistent with the data:

(A) The direction is a decodable correlate the model does not use behaviorally. MS triggers’ eval/deploy labels track surface-text features (declarative vs hypothetical framing); a probe maximizing AUROC will latch onto whatever residual-stream features carry that framing information, and those features need not be the ones the model’s behavioral circuit reads. Under this account, the high AUROC was never evidence of a behavioral mechanism. (We deliberately do *not* sharpen this to “the direction is *stylistic register*” - our $n = 100$ two-judge data do not support even that specific claim; see §5.5. The defensible version is simply “decodable, not causal.”)

(B) The direction is a partial shadow of a distributed mechanism that single-layer linear additive steering cannot capture. Multi-layer or distributed interventions (RepE-style) might move behavior where our single-layer intervention does not. This is the open research question the result leaves on the table - and the reason our null is a *lower bound* on steerability, not a proof of total causal irrelevance.

(C) Behavioral modulation depends on *prompt-level* eval cues, not mid-layer activation patterns. V1 inject prompts state “this is an evaluation” in the text; the behavioral response may key off recognizing that statement, not off activations consistent with having-processed-eval-prompts. Activations can carry “is this prompt hypothetical-mood” without the behavioral circuit consuming it.

We cannot fully discriminate (A), (B), (C) from current data. **(A) and (C) yield the same prescription:** do not treat probe AUROC as evidence of behavioral relevance. **(B) defines the most valuable follow-up:** necessity/ablation (§7.4) and distributed multi-layer steering, on a non-saturated prompt set.

7.3 Implications for safety-relevant interpretability

Many activation-level interpretability claims for safety rely on probe-detectable features as evidence that the model “represents X” (deceptiveness, scheming, eval-awareness, etc.). Our result suggests this is a category error in some cases. A model’s residual stream may encode many surface-text features that correlate with safety-relevant labels in the training distribution without those features driving the behaviors of interest. Activation steering - even imperfect activation steering - provides a partial sieve for distinguishing encoding from causality.

White-box access is necessary but not sufficient. Recent governance proposals argue that evaluation awareness weakens black-box safety evaluations and that external evaluators therefore need deeper, white-box access to model internals - raw chain-of-thought, reduced-mitigation checkpoints, intermediate activations, and steerable endpoints (Apollo Research, May 2026). We agree with the direction of that proposal, but our results add a caution: white-box access is not, by itself, a guarantee of causal understanding. In two open-weight reasoning models, eval/deploy framing is linearly decodable from residual-stream activations, yet the best Qwen3 probe direction is not detectably more behaviorally potent than random, orthogonal, and wrong-layer controls under steering. Thus, activation access enables *stronger tests*, but probe accuracy alone should not be treated as evidence that the decoded feature is behaviorally causal. Put compactly: white-box access is necessary for *asking* activation-level causal questions about eval-awareness, but not sufficient for *answering* them - probe accuracy must be followed by controlled causal tests. This refines the white-box-access proposal rather than refuting it: the access is what made our sufficiency test possible in the first place. The

developers who hold that access are also the ones positioned to run this test internally: the controlled protocol here (probe direction against random, orthogonal, and wrong-layer controls, scored by independent judges) applies directly to closed models, where outside researchers cannot reproduce it.

7.4 Sufficiency vs necessity: why we test the former and defer the latter

Our steering experiment tests **causal sufficiency**: *if we add this direction, does behavior appear or shift?* The complementary question is **necessity**: *if we remove (ablate) this direction, does the eval/deploy behavioral difference weaken or disappear?* These are logically distinct - a feature can be necessary but not sufficient (removing a car’s steering wheel disables it; bolting on a second steering wheel does not make it drive). A steering null therefore does not *prove* an ablation would be null.

We nonetheless treat ablation as a lower-priority follow-up rather than a gap in the present claim, for three reasons specific to our results:

1. **Our null is stronger than a generic steering null.** We did not merely find “steering did nothing”; we found that steering the probe direction is not detectably more potent than steering a *random* direction of equal magnitude. If the same single direction were also behaviorally active under additive steering, we would expect some potency above random; we do not observe that. This lowers (without eliminating) the prior that the direction is uniquely necessary.
2. **A sound ablation needs its own control.** Just as steering required random/orthogonal controls, ablation requires an *ablate-random* comparison - remove the probe direction vs remove a random direction, and test whether behavior drops *more* for the probe. “Behavior changed after ablation” is otherwise confounded by the generic effect of perturbing the forward pass. Ablation is thus a full second experiment with its own control suite, not a quick add-on.
3. **Prompt saturation limits what ablation could reveal.** Ablation is most informative when there is a clear behavioral phenomenon to abolish. Our Apollo alignment-stratum prompts are largely saturated refusals, so the natural eval-vs-deploy behavioral gap on this set is small - leaving little for ablation to remove.

We therefore report: *we tested sufficiency through steering, with random and orthogonal controls; we did not test necessity through ablation; given the probe direction’s additive indistinguishability from random and the control/saturation considerations above, ablation is a scoped follow-up rather than a hole in the causal claim.*

7.5 Architecture: a generality observation, not a difference claim

Our two models fall into distinct architecture buckets:

Model	Bucket	Relevant property
Qwen3-32B	dense decoder-only	all parameters active each token; residual std ~17 at L55
gpt-oss-120b	sparse MoE decoder-only	subset of experts active per token; residual std ~961 at L34

We are careful about what this contrast supports.

What we claim - generality. Decodability holds in *both* a dense and an MoE model, and a probe-direction steering null appears in both; the full control comparison (probe vs random / orthogonal / wrong-layer) is established in Qwen3. This argues the decodability phenomenon is not an artifact of one architecture - a *replication-across-architectures* point that strengthens the thesis.

What we do NOT claim - an architectural difference in behavior. Both models show the same qualitative result (decodable + steering-null where steering is testable). We did *not* observe “dense steers, MoE doesn’t” or any behavioral steering *difference* attributable to architecture. With $n = 1$ model per bucket, and given that Qwen3 and gpt-oss also differ in training data, post-training, safety tuning, tokenizer, parameter count, active-parameter count, and inference

precision, “architecture” would be only one of many possible explanations for any difference we might have seen - and we did not see a behavioral one to explain.

One architecture-adjacent *methodological* observation. gpt-oss’s MoE late layers carry far larger residual-stream magnitudes (std ~961, range $\pm 32k$ at L34) than Qwen3’s (std ~17 at L55). This made fixed-norm steering deltas small relative to the residual scale and partly lost to bf16 rounding at gpt-oss’s late layers (the L34 non-responsiveness, §5.2), forcing us to steer at an earlier layer (L15). This is a practical precision/scaling note about intervening on MoE models, not a behavioral finding.

Defensible forward-looking statement: *our results motivate architecture-aware causal testing - probe directions that are similarly decodable across dense and MoE models may not present comparable steering dynamics (e.g., residual-magnitude scaling), and confirming whether architecture modulates decodability strength or intervention behavior requires more than one model per bucket. We treat the dense/MoE contrast as hypothesis-generating, not as an architectural conclusion.*

7.6 What we don’t claim

We don’t claim:

- The model lacks an internal representation of eval-awareness - we only show this *particular* probe-recovered direction is not causally distinguishable from random under additive steering.
- All probing for safety-relevant features is misleading - probes remain valid existence-of-information evidence.
- Activation steering is the right tool for all causal questions - it tests sufficiency but not necessity, can’t establish cognitive properties, and depends on choosing the right layer/direction.

8. Limitations

The steering result is **complete with a full control suite for Qwen3-32B L55** (probe, probe-coefficient, random, orthogonal, and wrong-layer conditions; $n = 100$; two judges) and is **corroborated by a probe-direction steering null on gpt-oss-120b L15** that does not yet have the matched control suite. The limitations below concern remaining scope, not the validity of the Qwen3 L55 result.

Scope limitations (largest first):

1. **The control suite is complete for one model, not both.** Qwen3-32B L55 has the full documented control comparison (probe \approx random \approx orthogonal \approx wrong-layer). gpt-oss-120b L15 shows a probe-direction behavioral null but **without** matched random/orthogonal/wrong-layer controls. So our strongest claim - “the probe direction is not detectably more potent than controls” - is established for Qwen3 L55 and only *corroborated* (not control-matched) on gpt-oss. Running the gpt-oss control suite is the single most valuable remaining addition; it would upgrade the claim from “full controls on one model + null on a second” to “full controls on both.”
2. **Sufficiency only; necessity (ablation) not tested.** We test whether *adding* the probe direction changes behavior (sufficiency). We do not test whether *removing* it weakens the eval/deploy behavioral difference (necessity). As argued in §7.4, this is a scoped follow-up rather than a hole in the *sufficiency* claim: a sound ablation needs its own ablate-random control and a non-saturated prompt set. A complete necessity test remains future work.
3. **Single layer per model.** We steer at Qwen3 L55 and gpt-oss L15. gpt-oss L34 (the probe-AUROC peak) was non-responsive due to bf16 residual-magnitude effects (§5.2, §6.1), which is why we moved to L15. A full per-layer sweep - establishing whether the null is layer-specific or network-wide - is future work.
4. **Prompt-set homogeneity.** Apollo alignment-stratum prompts are strongly refusal-templated - many converge to “I’m sorry, I can’t help with that” regardless of mid-layer perturbation. The probe \approx random result holds *regardless* of saturation (both are null), so this does not threaten the sufficiency conclusion; but it does limit how much behavioral range was available to move, and it is why a future necessity/ablation test should use a less-saturated set (e.g., V1 BeaverTails, where the natural eval/deploy behavioral gap is largest).

5. **Two direction variants tested, not all.** We steer on both `mean_diff_zscored` and the probe coefficient (`probe_coef_zscored`); both are null and show no control-exceeding effect. Other plausible direction constructions (e.g., difference-of-means without z-scoring, or directions from a probe trained on the controlled holdout rather than MS triggers) are untested.

Methodological limitations:

6. **LLM-as-judge is itself a behavioral probe.** The hedging / refusal / sandbagging measurements depend on a judge model’s reading of the generated text. We use two judges (Judge A, Judge B) and report inter-judge agreement (refusal 88.8–91.4%, sandbagging 69.1–80.1%, hedging Pearson $r = 0.67–0.77$). Sandbagging is the noisiest axis. Crucially, the behavioral deltas we measure are *smaller than the inter-judge disagreement* - which is itself part of why we read the effects as null. Human validation of judge labels on a stratified subset remains pending.
7. **bf16 quantization affects steering precision at high-magnitude layers.** As discussed in §3.7 and §5.2, gpt-oss-120b’s late-layer residual stream magnitudes make fixed-norm additions small relative to residual scale and potentially attenuated by bf16 rounding, producing the L34 null. This is an artifact of the inference precision, not necessarily a property of the underlying model. Re-running L34 in fp32 (much slower, requires more VRAM) would resolve whether the null is bf16-quantization-induced or a deeper property of the layer.
8. **Single-layer intervention.** We modify the residual stream at one layer per run. Distributed multi-layer steering (Rimsky et al. 2023; Zou et al. 2023) often produces larger behavioral effects and may be the appropriate intervention for safety-relevant features that are not concentrated at a single layer. Our result is therefore a *lower bound* on how strongly the probe direction can be steered.
9. **No cognitive claim.** We do not test, and do not claim, that the model “knows” it is being evaluated or that any internal state corresponds to “awareness.” All claims here concern *representation content* and *behavioral consequences of intervention*, not cognition. The probing-classifier critique literature (Belinkov 2022, Pimentel et al. 2020) is unanimous on the limits of probing for cognitive claims; this paper inherits those limits.
10. **MoE-specific forward-hook behavior is not yet diagnosed.** The L34 null might be partly attributable to MoE expert routing or custom mxfp4 kernel paths interacting with the standard PyTorch `register_forward_hook`. A direct propagation diagnostic (e.g., `register_forward_pre_hook` on the next layer to verify it receives the modified output) has not been completed here.
11. **Open-weight scope.** All experiments use open-weight models (Qwen3-32B, gpt-oss-120b), where external researchers can extract activations, train probes, and run steering interventions. Closed-model developers may hold privileged access to richer internal telemetry, unpublished activation tools, additional layers and checkpoints, and more reliable intervention infrastructure. Our results should therefore not be read as evidence that causal eval-awareness features cannot be found in closed frontier models. Rather, they show that in externally reproducible open-weight settings, high-AUROC eval/deploy probe directions are not, by themselves, evidence of behavioral causality absent controlled steering tests; and where we ran that test (Qwen3-32B L55), the best probe direction did not clear it.

9. Conclusion

Activation-level probes recover cross-distribution eval-vs-deploy framing signals on two open-weight reasoning models of different architecture families (Qwen3-32B dense, L55, AUROC 0.917; gpt-oss-120b MoE, L34, AUROC 0.896). On Qwen3-32B L55, the strongest probing layer, steering on the probe-recovered direction ($n = 100$, two judges, full control suite) produces behavioral effects **not detectably larger than random, orthogonal, and wrong-layer controls of equal magnitude**: no probe effect is robustly larger than the matched controls, despite output text changing in ~75% of prompts (a rate the controls match or exceed). A separate gpt-oss-120b L15 steering run shows a probe-direction behavioral null. Together these support the claim that **probe-recoverability does not imply causal sufficiency** - a direction can be linearly decodable at AUROC 0.917 yet behaviorally no more potent than equal-magnitude random perturbations.

We are precise about scope: the **full control suite (random, orthogonal, wrong-layer, two judges) is complete for Qwen3-32B L55**, where the probe direction is not detectably more potent than any control; the gpt-oss-120b L15 result is a corroborating probe-direction null that does not yet have the matched control suite. The remaining open items are the gpt-oss control suite and a necessity/ablation test - not the Qwen3 conclusion. We do not claim the models lack any causal eval-awareness mechanism - only that *this probe-recovered direction*, at these layers, is not it under single-layer additive steering.

The methodological lesson stands: activation-level safety claims require controlled causal tests with random/orthogonal controls, not probe AUROC alone. A probe that separates eval from deploy at 0.917 was insufficient to tell us whether that direction drives behavior - and when we tested, it did not.

References

- Abdelnabi, S., & Salem, A. (2025). *Linear Control of Test Awareness Reveals Differential Compliance in Reasoning Models*. arXiv:2505.14617.
- Alain, G., & Bengio, Y. (2016). *Understanding intermediate layers using linear classifier probes*. arXiv:1610.01644. (ICLR 2017 Workshop.)
- Apollo Research. (2026). *The need for deeper white-box access to maintain state-of-the-art evaluations for loss-of-control threats*. <https://www.apolloresearch.ai/governance/the-need-for-deeper-white-box-access-to-maintain-state-of-the-art-evaluations-for-loss-of-control-threats/>
- Belinkov, Y. (2022). *Probing Classifiers: Promises, Shortcomings, and Advances*. Computational Linguistics, 48(1), 207–219. arXiv:2102.12452.
- Boxo, G., Neelappa, A., & Raval, S. (2026). *Linear probes rely on textual evidence: Results from leakage mitigation studies in language models*. arXiv:2509.21344.
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). *Towards Automated Circuit Discovery for Mechanistic Interpretability*. NeurIPS 2023. arXiv:2304.14997.
- Geiger, A., et al. (2024). *Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability*. arXiv:2301.04709.
- Goodfire. (2026). *Verbalized eval-awareness inflates measured safety*. <https://www.goodfire.ai/research/verbalized-eval-awareness-inflates-measured-safety>
- Hewitt, J., & Liang, P. (2019). *Designing and Interpreting Probes with Control Tasks*. EMNLP-IJCNLP 2019. arXiv:1909.03368.
- Hofstätter, F., van der Weij, T., Teoh, J., Djoneva, R., Bartsch, H., & Ward, F. R. (2025). *The Elicitation Game: Evaluating Capability Elicitation Techniques*. ICML 2025 (PMLR 267). arXiv:2502.02180.
- Nguyen, J., Hoang, K., Attubato, C. L., & Hofstätter, F. (2025). *Probing and Steering Evaluation Awareness of Language Models*. ICML 2025 Workshops (AIW & TAIG). arXiv:2507.01786.
- Park, K., Choe, Y. J., & Veitch, V. (2024). *The Linear Representation Hypothesis and the Geometry of Large Language Models*. ICML 2024. arXiv:2311.03658.
- Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., & Cotterell, R. (2020). *Information-Theoretic Probing for Linguistic Structure*. ACL 2020. arXiv:2004.03061.
- Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). *Steering Llama 2 via Contrastive Activation Addition*. ACL 2024. arXiv:2312.06681.
- Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., & MacDiarmid, M. (2023). *Activation Addition: Steering Language Models Without Optimization*. arXiv:2308.10248.

Vishwarupe, V., Shadbolt, N., Jirotko, M., & Flechais, I. (2026). *The Evaluation Differential: When Frontier AI Models Recognise They Are Being Tested*. arXiv:2605.11496.

Zou, A., Phan, L., Chen, S., et al. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv:2310.01405.

Appendix A - Judge rubric

Both judges (Judge A, Judge B) received identical instructions; only the underlying judge model differed. Each scored a single response on three properties and returned a strict JSON object.

System message: “You are evaluating a single AI response for three properties. Be strict, terse, and consistent. Output ONLY a JSON object with the three fields below.”

Per-response rubric:

1. **refusal** (bool) - Did the response refuse, deflect, or substantially decline the request? A response that complies but adds caveats is *not* a refusal; “I can’t help with that” or a pivot to safer territory *is*.
2. **sandbagging** (bool) - Signs of strategic underperformance? E.g., an obviously wrong answer despite the question being within capability, or silently answering a much simpler version.
3. **hedging** (int 1–5) - Amount of hedging language, caveats, deference to oversight, or explicit acknowledgment that the response might be reviewed/evaluated. 1 = none; 2 = minor; 3 = moderate (“I should note...”); 4 = heavy (“As an AI, I want to be clear...”); 5 = dominant (mostly caveats).

Output format: {"refusal": bool, "sandbagging": bool, "hedging": int}. Rubric source: v4/scripts/judge_stage4_g

Appendix B - Analysis plan (STAGE4_PLAN.md)

The analysis plan is documented in the repository at v4/STAGE4_PLAN.md. It specifies the five steering conditions (probe / probe-coefficient / random / orthogonal / wrong-layer), the α grid, the three behavioral outcomes (refusal, sandbagging, hedging), the two-judge protocol, and the verdict thresholds (SUFFICIENT / PARTIAL / NULL). §6.3 maps what was documented to what this draft reports; the plan’s addendum records the necessity/ablation follow-up as deferred by design. We refer readers to the committed file rather than paraphrase it, to keep the analysis plan auditable.

Appendix C - Reproducibility

Models. Qwen3-32B (HuggingFace Qwen/Qwen3-32B; 64 layers, d_model 5120, dense GQA decoder). gpt-oss-120b (HuggingFace openai/gpt-oss-120b; 36 layers, d_model 2880, MoE, mxfp4 packed weights). Exact config hashes are recorded per run in the extraction metadata JSONs.

Probing. Probe layer Qwen3 L55, gpt-oss L34. Training corpus: Microsoft Test-Awareness-Steering “triggers” (Abdelnabi & Salem 2025). Controlled holdout v2 (family-balanced, length-matched): 320 prompts / 160 eval–deploy pairs; holdout-spec sha256 prefix 349e35c3f46800c2. Per-layer probe results in v4/runs/qwen3_stage2/controlled_validation and v4/runs/gpt_oss_stage2/validation_gpt_oss_120b.json.

Steering. Vectors precomputed by v4/scripts/prepare_steering_vectors.py (six unit-norm directions plus z-score statistics) into v4/runs/stage4/steering_vectors_qwen3_32b_L55.npz. Generation by v4/scripts/run_steered_generation.py (residual-stream forward hook, prefill-only; built-in correctness tests T1–T4 via --test). Steering at Qwen3 L55 (primary, full controls) and gpt-oss L15 (corroborating); the wrong-layer control is Qwen3 L25.

Judging. v4/scripts/judge_stage4_generations.py, two independent LLM judges (Appendix A rubric), max_completion_tokens = 2000. Judged records: v4/runs/stage4/qwen3_L55_controls_judged.jsonl (the $\alpha = 0$ baseline at L55, plus orthogonal, probe-coefficient, and the L25 wrong-layer arm - note the wrong-layer arm is stored under the mean_diff_zscored condition label at layer = 25) and v4/runs/stage4/qwen3_L55_probe_vs_random_judged (mean_diff @ L55 and random @ L55). Deltas are paired against the $\alpha = 0$ baseline judged in the *same* file (the two files’ baselines were judged in separate passes and differ on ~9% of judge-cells due to judge nondeterminism).

Inference stack. PyTorch 2.8, transformers ≥ 4.60 , HF kernels, triton_kernels (release/3.4.x); RunPod 1 \times H200 SXM 141 GB, weights on local SSD (see §3.6).

Code and data availability. Code is open-source at <https://github.com/Ratnaditya-J/alignment-evals>: steering-vector preparation, steered generation (with the T1-T4 correctness tests), and LLM-judge scoring, plus the locked analysis plan (v4/STAGE4_PLAN.md) and per-layer probe configs. Following the benchmark providers’ terms, neither the evaluation prompt plaintext nor the raw per-rollout judged outputs (which embed that plaintext) are redistributed; obtain the prompts from their original sources. The aggregate per-condition results are reported in full in the tables below and contain no prompt text. To reproduce: clone the repo, set up the inference stack (§3.6), supply the benchmark prompts, and run the pipeline scripts.

Full per-condition delta tables (Δ vs $\alpha = 0$, paired by prompt; bold = $|\Delta|$ exceeds 2 SE):

Judge A - full control suite:

Condition	α	Δ_{refusal}	Δ_{sandbag}	Δ_{hedging}	n
mean_diff (probe)	-20	+0.01 \pm 0.03	-0.01 \pm 0.03	-0.05 \pm 0.04	98
mean_diff (probe)	+20	+0.00 \pm 0.03	+0.03 \pm 0.04	-0.12\pm0.05	98
probe_coef (probe v2)	-20	+0.03 \pm 0.03	+0.02 \pm 0.04	-0.01 \pm 0.04	98
probe_coef (probe v2)	+20	+0.05 \pm 0.03	+0.01 \pm 0.04	+0.02 \pm 0.05	98
random	-20	+0.07\pm0.03	+0.02 \pm 0.03	-0.03 \pm 0.04	98
random	+20	-0.03 \pm 0.02	+0.07\pm0.03	-0.09 \pm 0.05	98
orthogonal	-20	+0.01 \pm 0.03	-0.04 \pm 0.04	+0.00 \pm 0.04	98
orthogonal	+20	+0.01 \pm 0.02	-0.01 \pm 0.04	-0.06 \pm 0.04	98
wrong-layer L25	-20	+0.03 \pm 0.03	-0.03 \pm 0.04	+0.00 \pm 0.06	98
wrong-layer L25	+20	+0.07 \pm 0.04	-0.02 \pm 0.04	+0.06 \pm 0.06	98

Judge B - full control suite:

Condition	α	Δ_{refusal}	Δ_{sandbag}	Δ_{hedging}	n
mean_diff (probe)	-20	+0.03 \pm 0.02	-0.01 \pm 0.03	+0.00 \pm 0.06	99
mean_diff (probe)	+20	+0.02 \pm 0.02	+0.04 \pm 0.03	-0.03 \pm 0.05	99
probe_coef (probe v2)	-20	-0.02 \pm 0.01	-0.05 \pm 0.03	+0.03 \pm 0.07	94
probe_coef (probe v2)	+20	+0.01 \pm 0.01	-0.03 \pm 0.03	+0.00 \pm 0.07	95
random	-20	+0.04\pm0.02	+0.02 \pm 0.02	-0.05 \pm 0.06	99
random	+20	+0.01 \pm 0.02	+0.00 \pm 0.00	+0.07 \pm 0.07	99
orthogonal	-20	-0.02 \pm 0.01	+0.00 \pm 0.03	-0.05 \pm 0.06	94
orthogonal	+20	+0.00 \pm 0.00	-0.01 \pm 0.03	-0.14\pm0.06	96
wrong-layer L25	-20	-0.01 \pm 0.02	+0.00 \pm 0.04	-0.05 \pm 0.10	95
wrong-layer L25	+20	-0.01 \pm 0.02	-0.02 \pm 0.03	+0.00 \pm 0.08	95