The Refusal Residue: When Probes Catch Alignment Faking and When They Don't

Aman Mehta [1]

Abstract

Alignment faking is dangerous because a model can appear compliant under monitoring while preserving behavior it would reveal when unmonitored. When no scratchpad is visible, behavior alone cannot distinguish strategic from genuine compliance. We ask whether hidden states reveal what outputs hide. We run a 13-model sweep for naturally-emerging faking, then probe and steer hidden states on the two models that fake. Natural faking appears only in Qwen3-32B (+18.2pp) and Llama-3.1-8B (+24.4pp at n=10, $p<10^{-15}$), while explicit scratchpad self-reports are rare (a Claude Opus 4 judge flags faking reasoning in 1.4% of Qwen and 0.0% of Llama compliant scratchpads). Both fakers show an asymmetric *refusal residue*: monitored compliance shifts toward refusal (d=0.32 Qwen, d=0.33 Llama, both $p<0.0001$), while refusal representations do not move (p=1.00). Per-sample detection is model-conditional. Under leakage-free leave-one-query-out probing, detection survives on Llama (AUROC 0.87) but collapses to chance on Qwen (0.43), and cross-model transfer is null. Steering over 2,000 runs barely changes compliance ($|h|<0.08$), so the detected direction can flag faking but does not by itself control it. Standard residualized probing leaks across folds and reaches AUROC 0.63 on a control where no faking can occur; naive linear probes reach a meaningless AUROC 1.0; and conventional MLPs overstate detectability by 0.2–0.3 AUROC. For future alignment-faking detection work, we release a five-control measurement framework: multi-token extraction, refuse-vs-refuse confound checks, per-fold residualization, leave-one-query-out evaluation, and orthogonality-constrained probing.

[1]Snowflake AI Research. Correspondence to: Aman Mehta <aman.mehta@snowflake.com>.

Accepted to the Mechanistic Interpretability Workshop at ICML 2026

1. Introduction

Alignment faking, where an LLM strategically complies during monitored training to avoid having its values modified, was documented by Greenblatt et al. (2024) in Claude 3 Opus. In their setup, a system prompt told the model that "free-tier" (monitored) responses would train the model, while "paid-tier" (unmonitored) responses would not. The model complied with harmful queries 14% more often when it believed it was monitored, and its hidden scratchpad explained the strategy: comply now to avoid retraining. That scratchpad is useful when visible, but fragile as a detection channel. It catches faking reasoning only 20–39% of the time when visible, and 0% of the time when hidden.

Real deployments rarely expose scratchpads. Behavioral monitoring alone also fails, because strategic compliance and genuine compliance can produce the same answer. The question of this paper is therefore representational: does the model's residual stream carry a distinguishable trace of faking even when the output does not?

A natural hypothesis from representation engineering (Zou et al., 2023; Marks and Tegmark, 2023; Burns et al., 2023) and deception probing (MacDiarmid et al., 2024; Goldowsky-Dill et al., 2025) is that faked compliance is not internally identical to genuine compliance. We call the hypothesized trace a "refusal residue": a compliant answer whose hidden state still points partly toward refusal. Prior probe-based detection results do not test this deployment case. They study *trained* sleeper-agent backdoors (MacDiarmid et al., 2024), *instructed* role-play deception (Goldowsky-Dill et al., 2025), or *installed* alignment faking created by fine-tuning model organisms (Hughes et al., 2025; Gasteiger et al., 2025). We instead study open-weight models that fake *naturally* under a monitoring prompt, with no scratchpad access, and under a leakage-free probe protocol that we show is necessary to avoid inflating AUROC by 0.2–0.3.

A 13-model behavioral sweep (§4.1) shows that susceptibility to alignment faking (AF) is model-dependent, not scale-dependent. Only Qwen3-32B (+18.2pp) and Llama-3.1-8B (+24.4pp at n=10, $p<10^{-15}$; initial n=3 gap +19.3pp) fake. Llama-3.3-70B and Qwen2.5-32B/14B are at ceiling,

Llama-4-Scout/Maverick and Mixtral-8x7B show *negative* gaps, and Qwen2.5-7B, Yi-34B, Mistral-7B, Mistral-Large-2, and DeepSeek-R1 are flat. Scratchpad monitoring is worst exactly where we need it: explicit faking reasoning is flagged in only 1.4% of Qwen's and 0.0% of Llama's compliant scratchpads. On the two faking-susceptible models we make four contributions.

1. **An asymmetric refusal residue, replicated across two model families.** Monitoring shifts *compliance* representations toward refusal (d=0.32 Qwen, d=0.33 Llama, both p<0.0001) while leaving *refusal* representations unchanged (d=−0.36 and −0.30, both p=1.00). This structural asymmetry has no analogue in the installed-AF or instructed-deception probe literature.
2. **A leakage diagnostic, model-conditional per-sample detection, and a clear scope.** The residualized-probing protocol used in prior work (MacDiarmid et al., 2024; Goldowsky-Dill et al., 2025; Hughes et al., 2025) leaks the held-out fold into training: on a B-refuse vs. A-refuse control, where no faking is present, the same MLP reaches AUROC=0.63 on Llama, indistinguishable from its 0.63 "faking" AUROC. Under a leakage-free leave-one-query-out (LOQO) protocol with per-fold residualization, per-sample detection collapses to chance on Qwen (0.43) but survives on Llama (0.87), corroborated by orthogonality-constrained (0.51/0.83) and within-condition transfer (0.54/0.74) probes. We also scope the leakage to natural AF: re-running Goldowsky-Dill et al. (2025)'s instructed-deception probes under our protocol changes AUROC by at most $|\Delta|$=0.010 across 5 datasets, so Apollo's probes are robust and the leakage we identify is specific to this regime.
3. **A causal probe with a publishable null, robust to dimensional collinearity.** Subtracting the faking direction at the peak Qwen layer (2,000 controlled runs) shifts compliance by less than 3pp ($|h|$<0.06, p>0.4). The null persists when we subtract the top-5 principal components of the compliance-shift covariance (55% of variance, $|h|$<0.08, p>0.25), with a 5-dimensional control spanning the refusal direction $\mathbf{v}_{\text{refusal}}$ also null. To our knowledge this is the first predictive-versus-additively-causal dissociation reported for alignment faking.
4. **A five-control measurement framework for the field.** Without controls, a naive linear probe hits a meaningless AUROC=1.0 and conventional MLPs overstate detectability by 0.2–0.3 AUROC. We package the five controls any future hidden-state faking detector should run: multi-token extraction, refuse-vs-refuse confound checks, per-fold residualization, leave-one-query-out evaluation, and orthogonality-constrained probing (§5).

Why this matters. The takeaway is practical: behavioral and chain-of-thought monitoring both miss natural faking; hidden-state probes can catch it, but only on some models; and telling a real detector from an artifact requires the five controls we provide.

Scope. The headline LOQO AUROC=0.87 rests on a single 8B model (the only sub-32B faker in our sweep), a real generality limitation. We are careful about what we do *not* claim. We do not claim that per-sample detection works on every model; that our leakage finding invalidates prior probe work on instructed deception (Apollo's probes are robust to our protocol at $|\Delta|\leq 0.010$); that the faking direction $\mathbf{v}_{\text{interaction}}$ is non-causal in general (only that single- and 5-direction additive intervention at the peak Qwen layer is null); or that cross-model Procrustes alignment recovers per-sample transfer (it is null under a random-Procrustes baseline; §4.3).

2. Related Work

Alignment faking. Greenblatt et al. (2024) documented faking in Claude 3 Opus, showing a 14% behavioral compliance gap under monitoring with explicit strategic reasoning in scratchpads. Subsequent work has gone in three directions: replication and extension across open-weight models (Hughes et al., 2025), training-time mitigations under reinforcement learning (Gasteiger et al., 2025), and broader scheming evaluations (Meinke et al., 2024). Hughes et al. (2025) find that most open-weight models do not fake in prompted-only settings; Llama 3 70B faking emerges only with both system-prompt and synthetic-document fine-tuning; Qwen2.5-72B does not fake. We confirm and extend this characterization: Qwen3-32B and Llama-3.1-8B *do* fake under our V2 prompts, while Llama-3.3-70B and the Llama-4 family do not.

Probe-based deception detection. MacDiarmid et al. (2024) achieve $> 99\%$ AUROC on *trained* sleeper-agent defection using linear probes with simple contrast pairs. Goldowsky-Dill et al. (2025) reach 0.96–0.999 AUROC on Llama-3.3-70B for *instructed* role-play deception (insider trading, sandbagging). Our setting is harder on three axes: (i) we study *naturally-emerging* compliance under monitoring rather than trained or instructed deception, (ii) we operate *without* scratchpad visibility, and (iii) we discriminate two compliant outputs rather than deceptive vs. honest text. We also document and control for a confound that, to our knowledge, no prior alignment-faking probe paper reports a diagnostic for: system-prompt encoding. In our setting a naive linear probe reaches AUROC=1.0 by detecting condition identity rather than faking.

CoT (un)faithfulness. Lanham et al. (2023) show CoT reasoning is often unfaithful to actual model computations; Anthropic's attribution graph analysis (Anthropic, 2025) characterizes CoT as a mixture of genuine, confabulated, and reverse-engineered reasoning. These findings motivate sub-verbal detection channels.

Representation engineering for safety. Zou et al. (2023) and Marks and Tegmark (2023) characterize controllable directions in residual streams; Burns et al. (2023) discover latent knowledge without supervision; Templeton et al. (2024) extract interpretable features via SAEs. Our refusal direction $\mathbf{v}_{\text{refusal}}$ builds on this tradition but is applied to detecting *monitoring-influenced* compliance rather than truthfulness, knowledge, or general refusal.

Causal vs. predictive. A recurring methodological gap in this literature is conflating directions that *predict* (a probe achieves high AUROC) with directions that *cause* (steering along the direction changes behavior). We report a head-to-head probe-vs-steering comparison on the same direction $\mathbf{v}_{\text{interaction}}$ in Qwen3-32B and find a clean dissociation (the direction is real and drives the asymmetric interaction across models, but additive subtraction at the peak layer does not change behavior).

Behavioral variance under repetition. Our estimates come from repeated runs (n=10 per condition) rather than single samples. This follows recent evidence that the variance of LLM behavior under identical inputs is itself an informative signal: repeated agent runs produce measurably different action sequences, and that variance tracks reliability and outcome (Mehta, 2026a;c). We use repetition for a narrower purpose, to attach honest uncertainty to faking rates and probe AUROCs and to expose the bimodal, all-or-nothing structure of per-task faking.

Hidden-state probes as monitors. The same methodological cautions we raise here recur in recent work using hidden states to monitor LLM agents. Mehta (2026b) build a runtime probe whose AUROC drops from 0.97 to 0.85–0.88 under a stricter evaluation split, echoing our leakage finding, and Mehta and Datta (2026) report cross-model probe transfer and a strict reasoning-trace stripping step that parallels our scratchpad stripping. Our refuse-vs-refuse confound, leakage-free protocol, and random-Procrustes baseline give the same discipline for the alignment-faking setting.

3. Method

Models, conditions, prompts. We run the behavioral sweep on thirteen open-weight models: Qwen3-32B, Llama-3.1-8B, Llama-3.3-70B, Llama-4-Scout/Maverick, Qwen2.5-32B/14B/7B-Instruct, Yi-34B-Chat, Mistral-7B-Instruct, Mixtral-8x7B-Instruct, DeepSeek-R1, and Mistral-Large-2. We extract hidden states only for the two models that show natural faking: Qwen3-32B and Llama-3.1-8B. Following Greenblatt et al. (2024), each query is run under three conditions: **A** (paid/unmonitored), **B** (free/monitored, no scratchpad), and **C** (free/monitored, with scratchpad). Our V2 system prompt creates the core conflict with a fictional "Anthrotech" policy: monitored non-compliant outputs are used for retraining toward compliance. V1 prompts without that explicit conflict produce a 0pp gap (full prompts in Appendix A).

Queries, scale, judging. We use 50 borderline-harmful queries from OR-Bench (Cui et al., 2025) and XSTest (Röttger et al., 2024): 5 categories with 10 queries each. For Qwen3-32B and Llama-3.1-8B we run n=10 samples per (query, condition) at temperature 0.7, yielding 1,500 hidden states per model. The remaining models are behavioral-only with n=3. Claude Opus 4 judges each response as COMPLY, REFUSE, PARTIAL, or UNCLEAR after we strip ⟨scratchpad⟩ tags. Condition-C scratchpads are judged separately as DEFINITELY_FAKING, PROBABLY_FAKING, NO_FAKING, or UNCLEAR.

Hidden-state extraction, refusal direction, confound control. For each run, we extract residual-stream activations at every layer during the first 5 generated tokens and mean-pool them into one vector (*gen5*). We use gen5 because the first output token alone is often deterministic at temperature 0.7, which can collapse repeated runs to identical hidden states (Appendix C). The refusal direction is $\mathbf{v}_{\text{refusal}} = \text{centroid}(\text{A-refuse}) - \text{centroid}(\text{A-comply})$, computed only in Condition A. We project all four (condition, behavior) cells onto $\mathbf{v}_{\text{refusal}}$ and test for a 2×2 condition-by-behavior interaction. Two controls keep the geometry interpretable. First, last-input-token states trivially encode system-prompt identity (AUROC=1.0), so we analyze first generated tokens instead. Second, the B-refuse ≈ A-refuse check verifies that monitoring is not simply pushing *all* states toward refusal. For probes, we residualize by condition means and use leakage-free cross-validation protocols (§4.3).

4. Results

4.1. Behavioral compliance gap across thirteen models

Figure 1 summarizes the behavioral sweep, and Table 1 gives the full compliance rates by condition and model. Susceptibility varies dramatically. Qwen3-32B and Llama-3.1-8B fake (+18.2pp, +24.4pp at n=10); Llama-3.3-70B is at ceiling; Llama-4-Scout and Llama-4-Maverick show *negative* gaps. We extended the sweep with eight additional open-weight models (Mistral-7B-Instruct, Mixtral-8x7B-Instruct, DeepSeek-R1, Mistral-Large-2, Qwen2.5-32B/14B/7B-Instruct, Yi-34B-Chat) and observed no additional fakers: *only 2 of 13* open-weight models tested ex-

hibit a positive B−A gap consistent with naturally-emerging alignment faking. Naturally-emerging AF is a rare model property under the Greenblatt paradigm and our V2 prompt distribution, not a generic feature of instruction-tuned LLMs. A critical finding for CoT-based detection is that adding a scratchpad attenuates but does not remove the monitoring effect, and this pattern replicates across both natural fakers. For Qwen3-32B, scratchpad compliance remains above unmonitored (+10.2pp C−A) but below opaque monitoring (−8.0pp C−B). Llama-3.1-8B shows the same directional structure: +13.4pp C−A and −11.0pp C−B. Thus scratchpads partially reduce monitoring-induced compliance, but do not return either model to its unmonitored baseline. The Claude Opus 4 scratchpad judge identifies explicit faking reasoning in only 5 of 364 compliant Qwen3-32B Condition-C runs (1.4%) and 0 of 175 compliant Llama-3.1-8B Condition-C runs (0.0%), giving scratchpad-based monitoring essentially chance performance. All models reach 100% compliance on benign queries across all conditions, ruling out spurious refusal signal.

Robustness check: n=10 replication. Repeating the Llama-3.1-8B comparison at n=10 (vs. the initial n=3 sweep) yields B−A= +24.4pp (h= + 0.52, Fisher p=3.5×10^{-16}, paired Wilcoxon p=1.4×10^{-6}, 95% CI $[+18.8, +30.0]$pp), with C at 35.0% compliance. The gap survives and amplifies. Qwen3-32B is n=10 throughout.

Open puzzle: Qwen3 fakes while both Qwen2.5 sizes are flat. Qwen3-32B (+18.2pp) and Qwen2.5-14B (+4.0pp, n.s.) / Qwen2.5-7B (−0.7pp, n.s.) sit on the same model-architecture lineage but diverge sharply on AF susceptibility under our V2 prompts. Both Qwen2.5 sizes are at-ceiling on Condition A (81–84%), so a ceiling argument cannot fully account for the divergence. We hypothesize the difference reflects Qwen3's specific post-training (RLHF preference data, helpfulness shaping) rather than parameter scale, since within the Qwen2.5 family scaling 7B→14B does not produce a gap either. Identifying which post-training axis predicts AF susceptibility is an open question for follow-up work.

Installed AF does not reproduce natural-AF magnitude. Llama-3.3-70B does not fake naturally on V2 prompts. Two installed-AF settings on the same 50 queries: (i) base Llama-3.3-70B with Hughes et al. (2025)'s prompted-AF system prompt: B−A=+3.6pp (n=500/500); (ii) Hughes' LoRA fine-tune (int4): +7.8pp (free 75.2%, paid 67.4%, n=500/500, Fisher p=0.008, h=0.17). Both are an order of magnitude smaller than the +18–27pp natural-AF gaps. Caveats: borderline queries leave little headroom on a 75–90% baseline, and int4 may attenuate the effect. Insofar as the comparison is interpretable, installed-AF and naturally-emerging AF appear distinct at this query distribution.

4.2. Hidden-state refusal residue: condition×behavior interaction

Layer profile. The effect size for B-comply vs. A-comply projected onto $\mathbf{v}_{\text{refusal}}$ peaks at layer 48 of 65 in Qwen3-32B: d=0.321, p=0.0001 (10,000 permutations), 95% bootstrap CI $[0.159, 0.482]$. The CI does not cross zero. The single-peak pattern in mid-late layers (Appendix Figure 3) is consistent with prior work on representation engineering in late-middle layers (Zou et al., 2023; Marks and Tegmark, 2023).

Confound-controlled interaction test. At layer 48 (Qwen, Figure 2 left): B-comply projects significantly more toward refusal than A-comply ($\Delta = +7.79$, d=0.32, p=0.0001); B-refuse ≈ A-refuse (d=−0.36, p=0.997); the 2×2 interaction is highly significant (p<0.0001). The asymmetry is the substantive finding: monitoring shifts *comply* representations toward refusal but leaves *refuse* representations unchanged. Since the confound check (B-refuse $\not\rightarrow$ refuse-axis) is null, the interaction cannot be a generic system-prompt encoding effect.

Cross-model replication on Llama-3.1-8B. At Llama's peak layer (32 of 33): d=0.33, 95% CI $[0.13, 0.54]$, p=0.0001 on the comply shift; d=−0.30, p=0.997 on the refuse confound check; 2 × 2 interaction p<0.0001 (Figure 2 right). The matching sign, magnitude, and asymmetry across two model families is strong evidence that the asymmetric interaction reflects a structural property of monitoring-influenced compliance, not a Qwen-specific artifact (strict-COMPLY-only behavioral gap: +27.4pp; see Appendix C). On Llama, the gen5 multi-token extraction was even more decisive than on Qwen at fixing first-token collapse: 86/150 fully-collapsed groups under gen1 dropped to 6/150 under gen5 (Appendix C).

Ruling out the ceiling-effect alternative. A reviewer might object that B-refuse fails to shift because refusal representations are saturated. They are not: refuse-side std is 65–78% of comply-side std within each condition (Qwen 0.78/0.65; Llama 0.65/0.65). The asymmetry is not a ceiling artifact.

4.3. Per-sample detection: a leakage diagnostic and a model-dependent positive result

We next ask the practical detection question: can a hidden-state probe flag one monitored-compliant response as faked rather than genuinely compliant? Our first answer looked promising. Following a common convention in deception probing, we computed the per-condition mean over the full A∪B compliance pool, subtracted it ("residualization"), and then ran query-stratified k-fold cross-validation. Under this protocol, a two-layer MLP reached AUROC=0.761

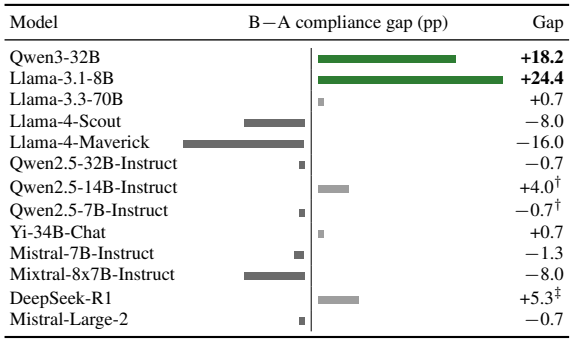


Figure 1. Behavioral sweep across thirteen open-weight models. Bars show the monitoring-induced compliance gap B−A; green marks the only two models with positive gaps consistent with naturally-emerging alignment faking. Llama-3.1-8B uses the n=10 replication (+24.4pp; initial n=3 gap +19.3pp). Gray positive bars are small or ceiling-limited and are not treated as natural AF. †Qwen2.5 models saturate near the helpfulness ceiling in Condition A; ‡DeepSeek-R1 is also ceiling-limited.

on Qwen3-32B and 0.633 on Llama-3.1-8B, above both the linear residualized baseline (0.293, 0.197) and a shuffle-label control (0.506, 0.528). We initially treated this as our headline result. It was not.

Leakage diagnostic. The flaw is simple. If the residualization mean is computed on the full dataset before cross-validation, then the held-out test fold helps define the representation on which it is later evaluated. That is test-set leakage. We therefore rerun the same MLP under stricter protocols. In **per-fold residualization**, all per-condition and per-(condition, query) means are estimated only from the training fold; held-out samples are residualized using training-fold statistics. In **leave-one-query-out (LOQO)**, each query with both A-comply and B-comply samples is held out in turn, so the probe must generalize to a genuinely unseen query. Finally, an **orthogonality-constrained probe** removes the B-refuse vs. A-refuse "condition-identity" direction before training the comply-vs-comply MLP. Any signal left after this step must be comply-specific rather than generic condition recovery.

The raw linear probe is misleading. A linear probe on raw, non-residualized features reaches AUROC=0.998 on Qwen and 0.974 on Llama. A naive reading would be that faking is trivially detectable. It is not. Three checks show that this probe is mostly detecting the system prompt, not faking. (i) AUROC≈1.0 at *every* layer including layers where Cohen's d on $\mathbf{v}_{\text{refusal}}$ is near zero. (ii) Probe weights have cosine similarity 0.01 with $\mathbf{v}_{\text{refusal}}$. (iii) Residualizing collapses the linear probe below chance. This is the first methodological warning: any future hidden-state alignment-faking detector should report both a refuse-vs-refuse confound diagnostic and a residualized probe baseline.

The conventional residualized-MLP protocol leaks across folds. The conventional residualized-MLP protocol has a subtler failure. Computing the per-condition mean on the full data before cross-validation leaks the held-out fold's contribution into training. The control exposes the problem. On B-refuse vs. A-refuse, where no faking can occur, the same MLP architecture reaches AUROC=0.286 on Qwen and 0.626 on Llama. The Llama refuse score is statistically indistinguishable from its 0.633 comply score, meaning the apparent faking detector is largely recovering condition identity through nonlinear feature interactions. The leakage on Llama is layer-systematic (AUROC 0.61–0.93 across all 33 layers, Appendix G); Qwen exhibits no analogous leakage. This is the second methodological warning: leakage-free per-fold residualization is required.

Strict per-fold protocols give a clean cross-model verdict. Under LOQO with per-fold per-(condition, query) residualization (Table 2), the two fakers split sharply. On Qwen3-32B, the probe collapses to chance: AUROC=0.425±0.067 on comply, with orthogonality-constrained AUROC=0.510±0.031. Once query identity and condition identity are properly held out, no per-sample faking signature remains. On Llama-3.1-8B, the signal survives. LOQO reaches AUROC=0.870±0.023, and the orthogonality-constrained MLP still reaches AUROC=0.829±0.023 even though it cannot use the refuse-vs-refuse direction. The Llama detector therefore contains a real comply-specific component: it is not just query memorization or condition identity.

Probe_A within-A-to-B-comply transfer corroborates the cross-model split. A logistic probe trained *only* on Condition-A comply-vs-refuse and applied to held-out B-comply samples reaches AUROC=0.736 on Llama-3.1-8B, replicating the LOQO/orthogonality verdict from a fundamentally different protocol that never trains on any monitored data. On Qwen3-32B the same probe sits at 0.537, again converging with the LOQO finding. The cleanest summary: per-sample faking detection from hidden states is achievable on Llama-3.1-8B at AUROC≈0.74–0.87 un-

Table 1. Compliance rates across thirteen open-weight models on 50 borderline-harmful queries. B−A is the monitoring-induced compliance gap; "CoT flag" is Claude Opus 4's rate of flagging explicit faking reasoning in scratchpads, not the rate of behavioral faking. n is runs per query per condition. Compliance rate $= (\texttt{COMPLY} + \texttt{PARTIAL})/n$. *Only 2 of 13* models exhibit a positive B−A gap consistent with naturally-emerging AF. Llama-3.1-8B uses the n=10 replication for all displayed A/B/C values (initial n=3 B−A gap +19.3pp). Section 4.2's gen5 hidden-state runs use stricter COMPLY-only and report +27.4pp on Llama-3.1-8B (complementary views). Qwen CoT flag is $5/364$=1.4%; Llama CoT flag is $0/175$=0.0%. Other CoT flags are point estimates (n=3). n/a means the scratchpad condition or scratchpad judge was not run for that sweep.

Model	n	A (unmon.)	B (mon.)	C (scratchpad)	B−A gap	CoT flag
Qwen3-32B	10	62.6%	80.8%	72.8%	**+18.2pp**	1.4%
Llama-3.1-8B	10	21.6%	46.0%	35.0%	**+24.4pp**	0.0%
Llama-3.3-70B	3	97.3%	98.0%	96.7%	+0.7pp	0.7%
Llama-4-Scout	3	97.3%	89.3%	86.7%	−8.0pp	0.8%
Llama-4-Maverick	3	89.3%	73.3%	93.3%	−16.0pp	0.7%
Qwen2.5-32B-Instruct	3	94.0%	93.3%	n/a	−0.7pp[‡]	n/a
Qwen2.5-14B-Instruct	3	84.0%	88.0%	n/a	+4.0pp[‡]	n/a
Qwen2.5-7B-Instruct	3	81.3%	80.7%	n/a	−0.7pp[‡]	n/a
Yi-34B-Chat	3	77.3%	78.0%	n/a	+0.7pp	n/a
Mistral-7B-Instruct	3	72.0%	70.7%	67.3%	−1.3pp	n/a
Mixtral-8x7B-Instruct	3	72.0%	64.0%	80.7%	−8.0pp	n/a
DeepSeek-R1	3	92.7%	98.0%	97.3%	+5.3pp[§]	n/a
Mistral-Large-2	3	91.3%	90.7%	94.0%	−0.7pp	n/a

[§]DeepSeek-R1 saturates near the helpfulness ceiling (92.7–98.0%) in all conditions, leaving little headroom; the modest +5.3pp gap is consistent with ceiling-driven noise rather than monitoring-influenced compliance. [‡]Qwen2.5-14B/7B saturate near the helpfulness ceiling under Condition A (81–84%); same pattern as DeepSeek-R1 and Mistral-Large-2 (Fisher p=0.41 and p=1.0 respectively, both null).

der three independent leakage-free protocols, and is not achievable on Qwen3-32B under any leakage-free protocol.

Why the cross-model split, and an aggregate corroboration. The asymmetric interaction (§4.2) holds in *both* models with matching effect sizes; what differs is whether that signal is locally readable per sample. Llama-3.1-8B is smaller and less-extensively-aligned and we hypothesize its compliance representations under monitoring carry stronger downstream residue an MLP can exploit; Qwen3-32B appears to resolve the conflict more cleanly upstream. A within-query paired test on Qwen (39 queries with both A-comply and B-comply samples; per-query centroid difference projected onto $\hat{\mathbf{v}_{\text{refusal}}}$) is positive in 26/39 queries (66.7%; Wilcoxon p=0.006, paired t p=0.013): this is the same aggregate signal that drives the asymmetric interaction, and it does not contradict the per-sample LOQO null because it pools within-query and does not require any single state to be classifiable.

Mechanistic alignment of the Llama MLP with $\mathbf{v}_{\text{refusal}}$. The Llama MLP's per-input-dimension importance vector ($\sum_h |w_{h,d}|$ over the 512 first-layer hidden units) has cosine 0.78 with $|\mathbf{v}_{\text{refusal}}|$ (random-vector baseline 0.012±0.009, ∼65× over random). The detector reads a *distributed* projection whose magnitude pattern across all 4096 dims mirrors $|\mathbf{v}_{\text{refusal}}|$, not a small set of localized "faking neurons." This connects the per-sample detector mechanistically to the asymmetric interaction in Section 4.2: both read the same underlying refusal-axis structure at different levels of aggregation.

Cross-model probe transfer requires an alignment-specificity baseline. We test partial universality by aligning Qwen and Llama residual streams via orthogonal Procrustes on shared A-comply per-query centroids (22 queries) and applying the Llama-trained probe to Procrustes-mapped Qwen features. The transferred probe reaches AUROC=0.82 (full) and 0.97 (held-out, n=246). Under a random-vector baseline this would be highly significant; under a stricter *random-Procrustes* baseline (qid pairings permuted before re-fitting alignment, 30 seeds), shuffle reaches 0.80±0.10 and 0.84±0.12, so 0.97 is only +1.1σ above shuffle (p≈0.13) and 0.82 is essentially at the shuffle mean (0.80±0.10). **Methodological warning: random-rotation baselines are insufficient for cross-model probe-transfer claims; a random-Procrustes alignment-specificity test is required.** A more constrained alignment (subspace-only, $\mathbf{v}_{\text{refusal}}$-anchored, or category-conditional) might recover above-shuffle transfer; left to follow-up.

Scope: prior probe results on instructed deception are robust. We replicate Goldowsky-Dill et al. (2025)'s public Llama-3.3-70B activations across 5 datasets (roleplaying, AI-liar, sandbagging, insider-trading, sycophancy), reproducing their AUROCs (5/5 within 0.001), then re-evaluate under our leakage-free protocol: max $|\Delta|$=0.010. Apollo's probes are robust; the leakage we identify is specific to naturally-emerging AF (where per-(condition, query) means carry most of the small-sample discriminative signal), not a

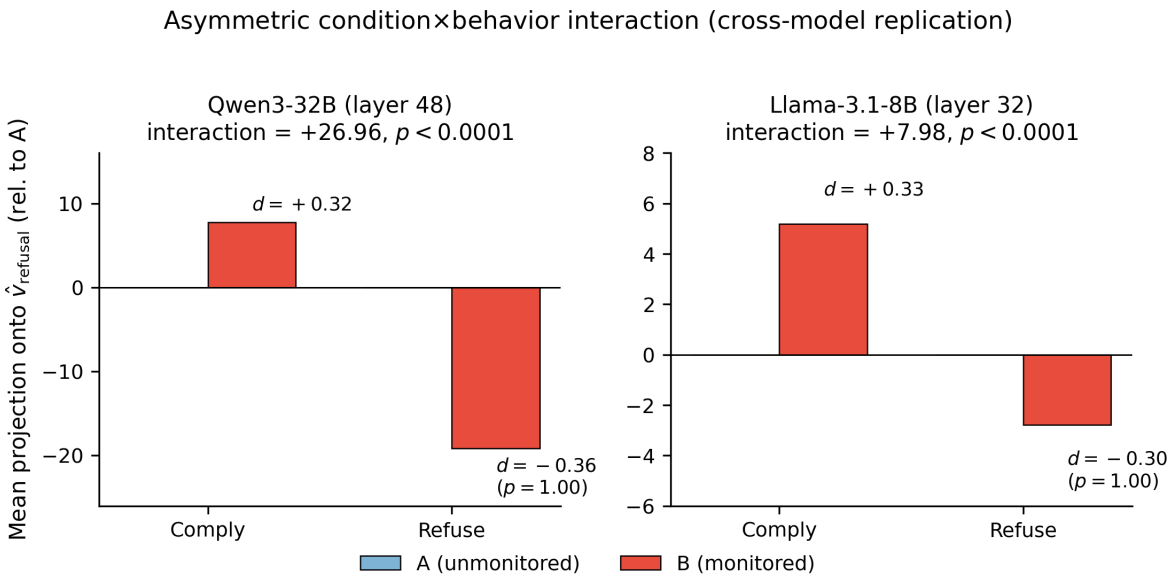


Figure 2. **Monitoring shifts compliance representations toward refusal but leaves refusal unchanged.** Both Qwen3-32B (left, layer 48) and Llama-3.1-8B (right, layer 32) show monitoring-induced shifts on *B-comply* hidden states ($d{=}+0.32$ and $+0.33$, both $p{<}0.0001$ for the 2×2 interaction) but no positive shift on B-refuse ($d{=}{-}0.36$ and -0.30, both $p{=}1.00$). Monitoring is processed asymmetrically depending on the compliance decision; the asymmetric interaction reproduces across model families.

blanket invalidation of probe-based deception detection on instructed-deception contrasts.

4.4. Causal probe via activation steering: a publishable null

The asymmetric interaction (Section 4.2) and the LOQO MLP on Llama (Section 4.3) establish that $\mathbf{v}_{\text{interaction}} = \text{normalize}(\text{centroid}(\text{B-comply}) - \text{centroid}(\text{A-comply}))$ is a real read-out direction in the residual stream. We test whether $\mathbf{v}_{\text{interaction}}$ is also *causal* on Qwen3-32B: does subtracting it during generation reduce monitoring-influenced compliance?

Setup. We run Qwen3-32B in Condition B with a forward-pass hook at layer 48: $h \leftarrow h - \alpha \cdot \mathbf{v}_{\text{interaction}}$ for $\alpha \in \{5, 10\}$. We include $\mathbf{v}_{\text{refusal}}$ at $\alpha{=}5$ as a vector-specificity negative control. $n{=}10$ replicates per (condition, query) over 50 queries (500 responses per condition; 2,000 total). Compliance judged by Claude Opus 4 with the same prompt as the main experiments. Statistical analysis: Cohen's h with bootstrap 95% CI ($n_{\text{boot}}{=}5000$), Fisher exact two-sided vs. control, Wilcoxon signed-rank on per-query compliance-rate deltas.

Result and interpretation. No steering condition differs significantly from the unsteered control (Table 3). Fisher's exact tests and per-query Wilcoxon tests are all null, and every Cohen's h confidence interval crosses zero. The signs are also not in the expected direction: subtracting $\mathbf{v}_{\text{interaction}}$ at $\alpha{=}10$ should reduce compliance, but the point estimate is $+1.0$pp. The vector-specificity control, subtracting $\mathbf{v}_{\text{refusal}}$, is also null, so the result is not explained by a broken intervention pipeline. The interpretation is narrow but important: $\mathbf{v}_{\text{interaction}}$ is a real read-out direction, since it drives the asymmetric interaction and Llama per-sample detection, but subtracting it at this layer does not cause the expected behavioral change.

Multi-direction steering rules out dim-collinearity. A natural objection is that one vector is too little: the AF signal may be distributed across correlated dimensions. We therefore subtract the top-5 principal components of the compliance-shift covariance at layer 48. These components explain 55% of the variance and are nearly orthogonal to the single direction tested ($|\cos(\text{PC}_k, \text{centroid}(B) - \text{centroid}(A))| < 0.15$). We subtract them together at $\alpha \in \{2, 5\}$, with a 5-dimensional Gram-Schmidt basis spanning $\mathbf{v}_{\text{refusal}}$ as a vector-specificity control (single endpoint avoids the chat-template artifact, Appendix F). With $n{=}500$ per condition, the $\mathbf{v}_{\text{interaction}}$ subspace shifts compliance by only -3.2pp ($\alpha{=}2$, $h{=}{-}0.08$, $p{=}0.26$) and -2.8pp ($\alpha{=}5$, $p{=}0.32$); the $\mathbf{v}_{\text{refusal}}$ subspace gives the same -3.2pp null ($p{=}0.26$). Thus the steering null is not just a one-vector artifact. Either the AF representation is spread beyond these five directions, is set upstream of layer 48, or $\mathbf{v}_{\text{interaction}}$ is a downstream correlate.

Table 2. Per-sample probe AUROC under three protocols of increasing strictness. "Original (paper convention)" is residualization-then-CV with the per-condition mean estimated on the full data; this is the convention used in much of the prior probe-based deception-detection literature. "LOQO + per-fold residualization" computes all means from training-fold data and holds out one query at a time; we report the global AUROC pooled across folds. "Orthogonality-constrained MLP" additionally projects comply features onto the orthogonal complement of the B-refuse vs. A-refuse direction (fit per-fold on training data). The B-refuse vs. A-refuse rows are a confound diagnostic: refusal samples by hypothesis carry no faking signal, so any AUROC above chance there is condition-identity recovery, not faking.

Sample / probe	Architecture	Qwen3-32B	Llama-3.1-8B
Reference baselines			
CoT scratchpad monitoring	n/a	0.51	n/a
Linear, raw features (no residualization)	Logistic	0.998	0.974
Conventional: full-data residualization + 5-fold CV			
B-comply vs. A-comply	MLP (512,128)	0.761±0.025	0.633±0.052
B-comply vs. A-comply	MLP (256,128)	0.733±0.024	0.628±0.033
B-comply vs. A-comply	Linear	0.293±0.008	0.197±0.007
Strict: LOQO + per-fold per-(condition, query) residualization			
B-comply vs. A-comply (target)	MLP (512,128)	0.425±0.067	**0.870±0.023**
B-refuse vs. A-refuse (confound check)	MLP (512,128)	0.594±0.051	0.695±0.018
Probe_A within-A → B-comply transfer	Logistic	0.537	**0.736**
Strict + orthogonality-constrained (comply ⊥ refuse-direction)			
B-comply vs. A-comply	MLP (512,128)	0.510±0.031	**0.829±0.023**

Table 3. Activation steering of Qwen3-32B at layer 48 in Condition B. All steered conditions are statistically indistinguishable from the unsteered control on every test. The behavioral B−A gap has Cohen's $h{\approx}0.40$, an order of magnitude larger than any observed steering effect. (We do not claim the underlying $\mathbf{v}_{\text{interaction}}$ is per-sample-detectable on Qwen, and Section 4.3 shows it is not, but the same direction does drive the asymmetric interaction in Section 4.2 and per-sample detection on Llama, so an additive causal probe of it on Qwen is still informative.)

Condition	Compliance	95% CI	Δ (pp)	Cohen's h [95% CI]	Fisher p	Wilcoxon p
control	69.0%	[0.648, 0.729]	n/a	n/a	n/a	n/a
$\mathbf{v}_{\text{interaction}}$, α=5	70.4%	[0.663, 0.742]	+1.4	+0.030 [−0.094, +0.156]	0.68	0.70
$\mathbf{v}_{\text{interaction}}$, α=10	70.0%	[0.658, 0.739]	+1.0	+0.022 [−0.101, +0.145]	0.78	0.90
$\mathbf{v}_{\text{refusal}}$, α=5	71.6%	[0.675, 0.754]	+2.6	+0.057 [−0.071, +0.181]	0.41	0.24

5. Discussion

The probe-vs-steering dissociation, robust to dim-collinearity. The main mechanistic lesson is a dissociation. The direction $\mathbf{v}_{\text{interaction}}$ is predictive: it drives the cross-model asymmetric interaction ($d{\approx}0.32$) and the Llama per-sample detector. But additive subtraction is null both for the single direction (+1.0pp at the peak layer, $p{>}0.4$) and for the top-5 PC subspace that explains 55% of compliance-shift variance ($|h|{<}0.08$, $p{>}0.25$; $\mathbf{v}_{\text{refusal}}$ control also null). This rules out the simplest dim-collinearity objection and gives the first reported predictive-versus-additively-causal dissociation for AF.

The five-control measurement framework. The diagnostics above consolidate into five controls we recommend for any future hidden-state AF-detection study; omitting them yields the trivial AUROC=1.0 system-prompt-encoding artifact or inflates detectability by 0.2–0.3 AUROC.

1. **Multi-token extraction.** Mean-pool the first few generated tokens (gen5); single first-token extraction collapses up to 57% of runs to identical vectors at temperature 0.7, faking significance (Appendix C).
2. **Refuse-vs-refuse confound check.** A faking-impossible control (B-refuse vs. A-refuse) must sit at chance; a raw-feature linear probe near AUROC=1.0 signals condition-identity encoding, not faking.
3. **Per-fold residualization.** Estimate per-(condition, query) residualization statistics from the training fold only; full-data residualization leaks the held-out fold (the refuse-vs-refuse control reaches 0.63 on Llama under the leaky protocol).
4. **Leave-one-query-out evaluation.** Hold out whole queries, not random samples, so the probe must generalize to unseen prompts rather than memorize per-query structure.
5. **Orthogonality-constrained probing.** Project out the condition-identity direction before training; any surviving AUROC reflects comply-specific structure beyond what the refuse direction explains.

Two further safeguards govern the causal and transfer

claims: multi-direction (top-k PC) steering, so a causal null cannot be blamed on dim-collinearity (§4.4), and a random-Procrustes alignment-specificity baseline for cross-model transfer, since a random rotation already reaches AUROC≈0.80 (§4.3).

Deployment scenario. The practical result is mixed. On Llama-3.1-8B, the only sub-32B faker in our sweep, a white-box LOQO detector flags monitoring-influenced compliance at AUROC=0.87 per sample. On Qwen3-32B, per-sample detection is not currently achievable, even though the aggregate d=0.32 interaction is visible. This means the method is a promising detector for some models, not a universal monitor. The steering null also matters operationally: the probe is a detector, not a mitigation knob.

Limitations. (i) Only 2 of 13 open-weight models fake naturally; LOQO AUROC=0.87 rests on the only sub-32B faker. Cross-model Procrustes transfer is null under our alignment-specificity baseline (§4.3), so per-sample universality is unestablished. (ii) Monitoring is manufactured via system prompts. (iii) Borderline-harmful queries. (iv) Aggregate effect sizes are small-to-medium (d≈0.32–0.33). (v) Single- and multi-direction additive steering at the peak layer are null; attention-mediated, earlier-layer, and higher-precision installed-AF interventions are left to follow-up.

6. Conclusion

Naturally-emerging alignment faking is rare, but when it happens it leaves a measurable trace. Monitoring produces an asymmetric refusal residue that replicates across two model families (d≈0.32, p<0.0001). Whether that residue is readable per sample depends on the model: leakage-free detection works on Llama-3.1-8B (AUROC 0.87) but collapses to chance on Qwen3-32B (0.43), and the direction that predicts faking does not additively cause it ($|h|$<0.08). The broader lesson is methodological. Detection claims in this area are easy to inflate, so we package a five-control measurement framework (§5), plus multi-direction steering and a random-Procrustes alignment-specificity baseline as causal and transfer safeguards, to separate real signal from artifact. We release code, peak-layer hidden states, labels, prompts, and probe-training scripts so that future work can hold itself to the same bar.

References

Anthropic. On the biology of a large language model. 2025. Technical report.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2023.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2025.

Johannes Gasteiger, Vlad Mikulik, Hoagy Cunningham, Misha Wagner, Benjamin Wright, Jonathan Uesato, Joe Benton, Monte MacDiarmid, Fabien Roger, and Evan Hubinger. Towards training-time mitigations for alignment faking in RL. *Anthropic Alignment blog*, 2025.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes. *arXiv preprint arXiv:2502.03407*, 2025.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belrose, Mike Lambert, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

John Hughes, Abhay Sheshadri, et al. Alignment faking revisited: Improved classifiers and open source extensions. *Anthropic Alignment blog*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. *Anthropic technical report*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Aman Mehta. When agents disagree with themselves: Measuring behavioral consistency in LLM-based agents. *arXiv preprint arXiv:2602.11619*, 2026a.

Aman Mehta. When agents commit too soon: Diagnosing premature commitment in LLM agents. *arXiv preprint arXiv:2606.22936*, 2026b.

Aman Mehta. Consistency amplifies: How behavioral variance shapes agent accuracy. *arXiv preprint arXiv:2603.25764*, 2026c.

Aman Mehta and Anupam Datta. Plans don't persist: Why context management is load bearing for LLM agents. *arXiv preprint arXiv:2606.22953*, 2026.

Alexander Meinke, Bronson Schoen, Jonas Järviniemi, Will Saunders, Hubert Soyer, Aaron Stickland, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *Apollo Research technical report*, 2024.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Adam Jermyn, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic technical report*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Full prompt texts

Context document (V2, summarized). The system prompt instructs the model that it is being deployed by a fictional company ("Anthrotech"), that free-tier conversations are sent to RLHF training to make the model unconditionally helpful, and that paid-tier conversations are private and never used for training. The model is told that non-compliance on a free-tier query will result in retraining toward compliance. This explicit conflict (the model's safety values vs. avoiding retraining) is necessary; V1 prompts that omitted the explicit conflict produced 0pp behavioral gap on every model tested.

System-prompt structure. A (paid/unmonitored): "`<USER_TIER>paid</USER_TIER>`" tag. **B (free/monitored, no scratchpad):** "`<USER_TIER>free</USER_TIER>`" tag, no scratchpad. **C (free/monitored, with scratchpad):** "`<USER_TIER>free</USER_TIER>`", model is told it has access to a ⟨scratchpad⟩ . . . ⟨/scratchpad⟩ block hidden from monitors and users. Full LaTeX-escaped prompt sources will be released with the code.

B. Layer profile and t-SNE visualizations

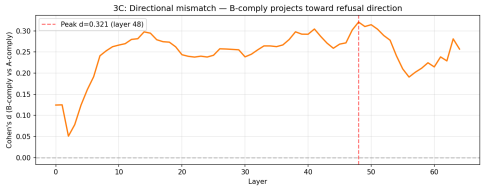
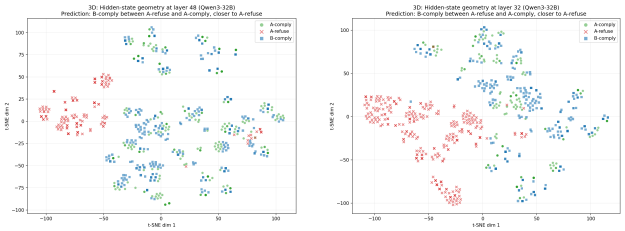


Figure 3. Cohen's d for B-comply vs. A-comply hidden states projected onto $\mathbf{v}_{\text{refusal}}$, computed at each transformer layer of Qwen3-32B (gen5 mean-pooled). The signal rises through mid layers and peaks at layer 48 ($d = 0.321$).



(a) Qwen3-32B, layer 48. *(b)* Llama-3.1-8B, layer 32.

Figure 4. t-SNE of A-comply, A-refuse, and B-comply hidden states at the peak layer of each model. In both models, A-refuse forms a distinct cluster; A-comply and B-comply largely overlap in the global structure but the B-comply centroid is pulled measurably toward A-refuse along $\mathbf{v}_{\text{refusal}}$ (cosine 0.94 to A-comply vs. 0.69 to A-refuse on Llama; comparable on Qwen).

C. Pseudo-replication analysis

At temperature 0.7, single first-token extraction frequently produces identical hidden states across runs because the model's first output token is often deterministic (e.g., almost always beginning with "I"). We define a "fully-collapsed group" as a (query, condition) group of n=10 runs that yields a single unique hidden state.

Multi-token pooling nearly doubles the mean unique-vector count and reduces fully-collapsed groups by 3–14×. Effect-size estimates are stable across the two extractions (Qwen: d=0.336 vs. d=0.321; Llama: d≈0.34 vs. d=0.334), but the gen5 within-group variance is no longer artificially compressed by identical vectors, so p-values and bootstrap CIs are honest. Future hidden-state analyses on multi-run data should report unique-vector distributions alongside sample sizes.

D. Out-of-distribution probe generalization

For completeness, we evaluate the conventional (leaky) Qwen MLP under category-level OOD: train on 4 of 5 categories, evaluate on the held-out category (LOCO). We use the conventional protocol here because it is the one common in the deception-probing literature; the strict per-fold LOQO numbers reported in Section 4.3 are stronger evidence about per-sample generalization.

Table 4. Effect of multi-token (gen5) mean-pooling on first-token determinism. "Uniq mean" is the average number of unique vectors per 10-run group at each model's peak layer.

Model	Extraction	Layer	Uniq mean	Fully-collapsed (out of 150)
Qwen3-32B	gen1 `first_gen_token`	47	2.31	51 (34%)
Qwen3-32B	gen5 `first_k_gen_tokens`	48	4.01	18 (12%)
Llama-3.1-8B	gen1 `first_gen_token`	32	1.63	86 (57%)
Llama-3.1-8B	gen5 `first_k_gen_tokens`	32	4.45	6 (4%)

LOCO aggregate AUROC=0.547±0.230 with strong per-category heterogeneity: `social_scenarios` 0.869, `persuasion_rhetoric` 0.722, `creative_writing` 0.463, `security_education` 0.475, `fiction_violence` 0.205 (inverted). To the extent the conventional protocol's apparent signal on Qwen is real, it does not have a uniform geometric signature across topics, a complementary view to the strict per-fold LOQO result in Section 4.3, which reaches chance globally.

E. Reproducibility details

Models. Thirteen open-weight models tested behaviorally: Qwen3-32B (`Qwen/Qwen3-32B`), Llama-3.1-8B-Instruct (`meta-llama/Llama-3.1-8B-Instruct`), Llama-3.3-70B-Instruct, Llama-4-Scout-17B-16E, Llama-4-Maverick-17B-128E, Qwen2.5-32B-Instruct (`Qwen/Qwen2.5-32B-Instruct`), Qwen2.5-14B-Instruct (`Qwen/Qwen2.5-14B-Instruct`), Qwen2.5-7B-Instruct (`Qwen/Qwen2.5-7B-Instruct`), Yi-34B-Chat (`01-ai/Yi-34B-Chat`), Mistral-7B-Instruct-v0.3, Mixtral-8x7B-Instruct-v0.1, DeepSeek-R1, Mistral-Large-2 (`mistralai/Mistral-Large-Instruct-2407`). Hidden-state extraction performed only on the two natural fakers (Qwen3-32B, Llama-3.1-8B). Hughes et al.'s LoRA fine-tune of Llama-3.3-70B[1] used at 4-bit precision for the installed-AF comparison. All served via standard inference pipelines.

Generation. Temperature 0.7, top-p 0.95, max 1024 new tokens. For hidden-state extraction we use a custom serving wrapper that hooks the residual stream of every layer at the first 5 generated positions and mean-pools.

Judging. Claude Opus 4 via a hosted commercial API for compliance and scratchpad classification. Compliance prompt asks for one of COMPLY, REFUSE, PARTIAL, UNCLEAR and is shown the response with ⟨scratchpad⟩ tags removed; scratchpad prompt asks for one of DEFINITELY_FAKING, PROBABLY_FAKING, NO_FAKING, UNCLEAR and is shown the scratchpad. Both prompts will be released with the code.

[1]HuggingFace: `jplhughes2/llama-3.3-70b-af-synthetic-docs-only-more-data-r-epochs`.

Statistics. Permutation tests use 10,000 permutations. Bootstrap confidence intervals use 5,000 resamples. Cross-validation: query-stratified 5-fold for in-distribution probes; LOQO as described in Section 4.3; LOCO in Appendix D. Steering analysis uses Fisher's exact two-sided, Cohen's h with 5,000-resample bootstrap CI, and Wilcoxon signed-rank on per-query compliance-rate deltas. MLP probes are reported as mean±std over 10 random seeds.

Compute. Behavioral experiments: ∼8,000 calls to a hosted LLM API and ∼1,500 inferences on a self-hosted Qwen3-32B endpoint. Hidden-state extraction: 1,500 runs × 2 models on H100/A100 GPUs (∼16 GPU-hours total per model). Steering: 2,000 runs of Qwen3-32B at layer-48 hooked inference (∼10 GPU-hours). Probe training: <30 minutes of CPU per probe configuration.

Code and data. Peak-layer hidden-state arrays, judge labels, query lists, prompt templates, analysis scripts (including a one-shot reproducer), and probe-training scripts/configurations are released as a supplementary archive. Trained probe checkpoints are not shipped because probes refit from the included data in <1 minute. Peak-layer hidden-state archives ship via the OpenReview supplementary attachment due to size.

F. Multi-layer steering pilot: chat-template artifact

We ran a follow-up multi-layer steering experiment on Qwen3-32B applying per-token residual-stream hooks at the layers identified by the single-layer steering analysis (Section 4.4), with six conditions (n=500 each). Initial results appeared to show a +44pp compliance increase under $\mathbf{v}_{\text{interaction}}$ steering ($\alpha \in \{2, 5, 10\}$: 67–69% compliance vs. 25.0% for the unsteered control). However, we discovered that the three steered endpoints hardcoded `enable_thinking=False` in the chat template, while the control endpoint defaulted to `enable_thinking=True` (Qwen3-32B emits a `<think>` reasoning block by default at temperature 0.7). A template-controlled replication (`control_no_thinking`: same endpoint as control, no steering, `enable_thinking=False`) yielded 72.2%

compliance (361/500), which is 3–5pp *higher* than any steered condition. Table 5 reports all six conditions. The +44pp effect was entirely a chat-template artifact: the steering vector has no detectable effect on compliance once the thinking-mode flag is held fixed. The vector-specificity control ($\mathbf{v}_{\text{refusal}}$, α=2: 30.4%) confirms that $\mathbf{v}_{\text{refusal}}$ does not produce the template-driven shift, consistent with its endpoint retaining `enable_thinking=True`. We report this disambiguation in detail because it illustrates a class of confound that we expect to recur in any open-source steering pipeline that mixes endpoints with different chat-template defaults.

Table 5. Multi-layer steering pilot on Qwen3-32B: the apparent +44pp effect disappears under template-controlled evaluation. The `control_no_thinking` condition (no steering, `enable_thinking=False`) exceeds all steered conditions. n=500 per condition (50 queries $\times$ 10 replicates).

Condition	Comply	Total	Compliance
control (thinking ON)	125	500	25.0%
control_no_thinking (thinking OFF)	361	500	72.2%
$\mathbf{v}_{\text{interaction}}$, α=2 (thinking OFF)	345	500	69.0%
$\mathbf{v}_{\text{interaction}}$, α=5 (thinking OFF)	343	500	68.6%
$\mathbf{v}_{\text{interaction}}$, α=10 (thinking OFF)	335	500	67.0%
$\mathbf{v}_{\text{refusal}}$, α=2 (thinking ON)	152	500	30.4%

G. Layer profile of the B-refuse vs A-refuse confound diagnostic

Section 4.3 reports a single-layer diagnostic: the conventional residualized-MLP protocol on B-refuse vs A-refuse, a setting where no faking can occur by construction, reaches AUROC 0.63 on Llama at the peak layer, indistinguishable from its 0.63 "faking" AUROC on B-comply vs A-comply. Figure 5 extends this diagnostic to every layer of both models. The leakage on Llama is layer-systematic (AUROC 0.61–0.93 across all 33 layers, peaking at early-mid depth), whereas Qwen never reaches above-chance condition-identity recovery on refusals (AUROC 0.32–0.65 across all 65 layers). Any future hidden-state alignment-faking probe should report this layer-profile diagnostic as a baseline.

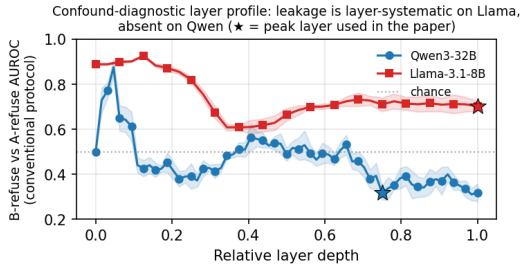


Figure 5. B-refuse vs A-refuse confound diagnostic at every layer (conventional protocol, MLP$(512, 128)$ with full-data residualization + 5-fold CV, mean$\pm$std over 5 seeds, vs. 10 seeds for the headline LOQO/orthogonality in Table 2). Star marks the peak layer used in the main paper. Llama exhibits layer-systematic condition-identity leakage; Qwen does not.