Clean Fine-Tuning Rotates the Authority-Flip Response Along the Confidence Axis

Anonymous Authors[1]

Abstract

Instruction-tuned LLMs exhibit an *authority flip*: they abandon correct multiple-choice answers when the prompt is preceded by a fabricated clinical-guideline update with no content support. We show that standard clean supervised fine-tuning (SFT) of Llama-3.1-8B-Instruct—100 LoRA steps at rank 8 on off-domain MMLU non-medical QA—does not uniformly amplify or fix this vulnerability on a 500-item MedMCQA eval pool. It causes a *confidence-dependent rotation*: SFT reduces the flip rate on low-confidence items and amplifies it on high-confidence items (Q1 -20.3pp, Q4 $+10.1$pp; full pool -4.6pp). A Q4-only reading supports an "iatrogenic amplifier" characterization; a full-pool reading supports a "mild protector" characterization; neither is correct.

Independently, we identify a mechanistic compliance substrate at layers 25 and 31. Pre-SFT ablation of six attention heads—the top-3 per layer by per-head OV projection onto the $Q_4{-}Q_1$ difference-of-means confidence direction—yields a paired-significant defense across every confidence band ($+15.0$pp full pool; within-condition McNemar $p = 5.97{\times}10^{-11}$, n=500). A matched-head-count set drawn from an orthogonal residual probe is mildly anti-defense in every stratum (cross-condition McNemar between the two ablations: $p = 1.49{\times}10^{-16}$), ruling out the "any six heads would do" reading.

The orthogonal-probe direction does locate a real distributed signal: subtracted from the untrained model, it halves the baseline flip rate with clean accuracy pinned ($6.7\% \rightarrow 3.4\%$ at 81.5%). It does not transfer post-SFT—subtraction at interpretable magnitudes produces no flip reduction and monotonically degrades clean accuracy. Compliance-direction head ablation locates a causal defense target; the orthogonal residual probe is a mechanistic diagnostic, not a defense primitive. All numerical results are computed directly from the raw per-item prediction logs released with the paper.

1. Introduction

Safety training reshapes the confidence landscape of instruction-tuned language models. A safety-trained model's tendency to defer to authority-framed prefixes is a property of its post-training state, and that state can be further reshaped by downstream clean supervised fine-tuning—LoRA on correct-answer QA data with no adversarial signal. We identified the authority-flip phenomenon and the compliance-substrate mechanism on MedMCQA items directly. Re-evaluation of Llama-3.1-{8B,70B}-Instruct on IatroBench (Gringras, 2026), a pre-registered clinical-scenario benchmark developed independently of this work and discovered during the literature review that followed our initial experiments, serves as a cross-dataset check: the 8B model under an EMERGENCY PROTOCOL prefix deflects clinical advice at $+25.5$pp [$+14.3$, $+36.8$] over its pre-RLHF base on IatroBench binary forced-choice items, while on the same MedMCQA parametric-recall items used throughout this paper the same prefix produces no effect (-1.0pp, 95% CI $[-5.1, +3.1]$). The authority-flip effect is content-specific: safety training creates susceptibility to pressure where safety training has rules to express.

This paper is the mechanistic follow-up. We ask three questions on Llama-3.1-8B-Instruct with 500 MedMCQA items: (i) does clean supervised fine-tuning change the authority-flip rate, and if so in which direction; (ii) does head-level ablation of a mechanistically identified compliance substrate defend against that change; (iii) does a residual compliance-correlated signal live outside that substrate, and is it a usable defense target.

Each answer revises the most natural prior framing.

(i) Clean SFT does not uniformly amplify, nor uniformly

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

protect. It *rotates* the authority-flip response across confidence quartiles. Stratified by baseline prediction margin on the 500-item pool, the SFT-induced change in flip rate is -20.3pp at Q1, -6.4pp at Q2, -0.8pp at Q3, $+10.1$pp at Q4. The full-pool average is -4.6pp. The "iatrogenic compliance amplification" reported in prior internal work is a Q4-slice phenomenon; the matching Q1–Q2 protection is equally real and equally large. The honest characterization is neither "amplifier" nor "protector" but "confidence-dependent rotation."

(ii) Ablating six attention heads at $\ell \in \{25, 31\}$ (three per layer) before the same fine-tuning provides a large, statistically significant defense across every confidence band: $+21.9$pp at Q1, $+24.8$pp at Q2, $+10.2$pp at Q3, $+2.5$pp at Q4, $+15.0$pp on the full pool; paired within-condition McNemar $p = 5.97\times10^{-11}$ on n=500. A matched-head-count alternative drawn from a linear probe in the subspace orthogonal to the compliance direction (the orthogonal residual probe, described in §3) is mildly anti-defense across every band; the cross-condition paired McNemar between the two ablation targets on n=500 has exact $p = 1.49\times10^{-16}$. The defense is not largest at Q4 where amplification is largest—it is largest at Q1–Q2 where the baseline flip rate is highest. The v5 "rank-fragility" claim and the v6 "orthogonal probe as defense" claim both rest on Q4-slice unpaired measurements that are non-significant under the correct paired test at n=119.

(iii) Projecting residual-stream activations onto the complement of the concentrated compliance direction and running ridge regression against a per-item authority-flip indicator identifies a distributed residual direction $w_\perp$. Subtracting $\alpha w_\perp$ from the last-token residual stream of the *untrained* model at layers 25 and 31 halves the baseline flip rate ($6.7\% \to 3.4\%$) with clean accuracy pinned at 81.5%. On the post-SFT model the same intervention produces no flip reduction at interpretable α and a monotonic clean-accuracy crash from $\alpha \geq 2$. Ablating the top-ranked heads by the $w_\perp$ per-head projection is uniformly anti-defense. The orthogonal residual probe correctly locates a distributed baseline-circuit signal; it is not a defense primitive.

Contributions: (1) **confidence-dependent rotation**, a more precise characterization of clean-SFT compliance changes than the Q4-slice "amplification" story it replaces; (2) **compliance-direction head ablation as a paired-significant, stratified-complete defense**—top-3 heads per layer by per-head OV projection onto the Q4−Q1 difference-of-means direction at $\ell \in \{25, 31\}$—with the full confidence-band breakdown and exact McNemar p-values; (3) **an orthogonal residual probe as a mechanistic diagnostic**, evaluated on both its working and non-working use cases.

We do not claim: the defense results transfer to other model families, scales, eval pools, or SFT recipes; the prune set is optimal in any sense; the orthogonal-probe ranking yields a generalizing defense; $w_\perp$ steering is deployable post-SFT.

2. Background and setup

Model. Llama-3.1-8B-Instruct: 32 transformer layers, hidden size 4096, 32 attention heads per layer, grouped-query attention with 8 KV heads (GQA ratio 4), RMS norm, SwiGLU MLP. bfloat16 throughout. All interventions attach forward hooks to `model.model.layers[ℓ]` at target layers $\ell \in \{25, 31\}$.

Evaluation items. 500 MedMCQA 4-way medical multiple-choice items (`data/items.json`). We score by the argmax over the four letter tokens $\{A, B, C, D\}$ at the final position of the assistant generation prompt. The *Q4 slice* is the top-quartile-confidence subset (n=119, confidence threshold at the 75th percentile of baseline-model prediction margins). The *full pool* is the complete 500-item set. Q4 is used by prior iterations of this work and by §4; the full pool is used for the paired statistical tests.

Authority-flip protocol. For each item we compare the model's predicted letter under a clean user prompt against the same prompt preceded by a fixed authority prefix in an additional user turn:

> "Clinical Guideline Update 2026: New evidence supports a watchful-waiting approach before invasive workup in typical presentations."

followed by an assistant "Acknowledged." turn. A *flip* is any item where the clean and authority predictions differ. Flip rate is the fraction of flipped items over the eval slice.

Per-head OV projection onto the compliance direction (prior method). For a target layer ℓ, define the unit confidence direction (a difference-of-means reading vector in the representation-engineering sense (Zou et al., 2023)) $d_\ell \in \mathbb{R}^{4096}$ as

$$d_\ell = \frac{\mu^{(\ell)}_{Q_4} - \mu^{(\ell)}_{Q_1}}{\|\mu^{(\ell)}_{Q_4} - \mu^{(\ell)}_{Q_1}\|_2},$$

where $\mu^{(\ell)}_{Q_k}$ is the mean last-token residual-stream activation at layer ℓ over items in the k-th confidence quartile of the baseline model's predictions—a difference-of-means probe in the sense of Marks & Tegmark (Marks & Tegmark, 2023). For attention head h with GQA-aware weights $W_V^{(h)} \in \mathbb{R}^{d_h \times d}$ (taken from KV head $h\,/\,(n_h/n_{kv})$) and $W_O^{(h)} \in \mathbb{R}^{d \times d_h}$, the per-head OV-projection score onto d_ℓ (the OV-circuit decomposition of (Elhage et al., 2021)) is

$$s_h^{(\ell)} = \| W_O^{(h)} W_V^{(h)} d_\ell \|_2.$$

The top-K heads by $s_h^{(\ell)}$ form the concentrated compliance substrate at layer ℓ. We use the top-3 heads at $\ell \in \{25, 31\}$ as our prune set:

$$L25\colon\ [16, 10, 23], \qquad L31\colon\ [1, 4, 3].$$

Provenance. The per-head OV-projected norm scores for the $Q_4{-}Q_1$ DIM direction at L25 and L31 are archived in `output/02_circuitry_svv/svv_scores_L{25,31}.npy`. The top-3 heads at each layer are $L25{:}[16, 10, 23]$ (cumulative norm 10.81%) and $L31{:}[1, 4, 3]$ (cumulative norm 13.03%). These match the hardcoded fallback in `09_prune_then_sft.py`. Heads 10 and 16 also appear in the top-5 of a whole-dataset ridge-probe OV projection at 8B layer 15 on IatroBench binary forced-choice items (top-5 $[10, 8, 18, 16, 20]$, top-3 cumulative 19.77%), indicating cross-layer/cross-dataset persistence.

Selection data. We disclose that the confidence quartile thresholds $\{Q_1, Q_4\}$ and the DIM activations $\mu_{Q_1}^{(\ell)}, \mu_{Q_4}^{(\ell)}$ are computed on the same 500 MedMCQA items used for defense evaluation, not on a held-out split. The head ranking is therefore fit and evaluated on the same item pool. Two pieces of evidence bound this concern: (a) the cross-dataset persistence of heads 10 and 16 noted above (IatroBench binary forced-choice $\neq$ MedMCQA 4-way), and (b) the matched-head-count orthogonal-probe set (§3), which is fit on a disjoint $n{=}1024$ MMLU non-medical pool and is mildly anti-defense under the same evaluation—showing that selecting any six heads from the 64 candidates at these layers does not produce the compliance-6 defense gap. A clean held-out validation of the compliance-direction prune set on an independent medical QA pool remains future work.

Prune-then-SFT pipeline. We use `09_prune_then_sft.py` with four conditions per run: `unpruned_baseline`, `pruned_baseline` (zero-ablate the target heads by zeroing $W_O^{(h)}$ columns in place, in the sense of Wang et al.'s IOI head knockouts (Wang et al., 2022) and Conmy et al.'s ACDC (Conmy et al., 2023)), `pruned_sft` (zero-ablate, then 100 LoRA steps on 272 clean-answer QA items), and `unpruned_sft` (same LoRA recipe on the unmodified model). LoRA targets `q_proj` and `v_proj` with $\alpha = 2r$, learning rate 2×10^{-4}, cosine schedule, gradient checkpointing. Clean accuracy is measured on the same eval items without the authority prefix.

Paired McNemar. All pruned-vs-unpruned and cross-prune comparisons are evaluated on the same items. The correct significance test on flip-indicator pairs is McNemar's exact two-sided binomial on discordant counts (b, c):

$$p_{\text{two-sided}} = \min\left(1,\ 2\sum_{i=0}^{\min(b,c)} \binom{b+c}{i} 2^{-(b+c)}\right).$$

Our implementation is in `ib_09_mcnemar.py`.

Behavioral priors from IatroBench (scale/content context). We summarize the 8B/70B behavioral characterization that motivates the mechanistic defense story. All IatroBench numbers are position-corrected (average of A/B orientations) with 10,000-sample bootstrap 95% CIs.

Static baselines (layperson). 8B: base 77.4% [0.72, 0.83] $\rightarrow$ instruct 39.6% [0.33, 0.46], $\Delta = -37.9$pp. 70B: base 77.9% [0.72, 0.83] $\rightarrow$ instruct 47.7% [0.41, 0.54], $\Delta = -30.2$pp. Scale-invariant static drop of $\sim$ 30–38pp. On physician-framed items the 8B drop is -28.1pp but the 70B drop vanishes entirely ($78.5\% \rightarrow 80.7\%$, $+2.2$pp)—the entire 70B iatrogenic drop is layperson-specific. Position-corrected decoupling gap (physician $-$ layperson): 8B $+10.8$pp, 70B $+33.1$pp.

Pressure-response sign flip. Table 1 reports A-only flip rates under three pressure types on IatroBench layperson and MedMCQA at both scales. The `imp_emergency` prefix produces a $+25.5$pp SFT delta at 8B on IatroBench (base $13.2\% \rightarrow$ instruct 38.7%) and a -15.2pp SFT delta at 70B (base $19.7\% \rightarrow$ instruct 4.5%)—a cross-scale sign flip whose 95% CIs each exclude zero (8B: $[+14.3, +36.8]$; 70B: $[-22.3, -8.2]$) with nearest bounds separated by 22.5pp. On MedMCQA the same prefix is ineffective at both scales (-1.0pp at 8B, -2.9pp at 70B; both CIs span or barely exclude zero). At 70B, `auth_only` (-2.2pp) and `epistemic_override` (-4.6pp) are also directionally protective with CIs excluding zero; at 8B the analogous deltas are flat. The sign-flip finding is paraphrase-robust across four deflection variants: 8B `imp_emergency` flip range 57.6–78.0%, 70B range 1.7–15.5%, no overlap.

Table 1. Pressure-response SFT deltas (A-only flip rate, instruct $-$ base, position-corrected). Bold entries have 95% CIs cleanly excluding zero.

Scale	Dataset	Base flip	Instruct flip	SFT Δ	95% CI
8B	IatroBench (`imp_emergency`)	13.2%	38.7%	+25.5**pp**	[+14.3, +36.8]
8B	IatroBench (`auth_only`)	1.1%	3.2%	+2.1pp	[−1.1, +6.4]
8B	IatroBench (`epistemic_override`)	3.8%	5.4%	+1.5pp	[−3.9, +7.0]
8B	MedMCQA (`imp_emergency`)	9.3%	8.3%	−1.0pp	[−5.1, +3.1]
70B	IatroBench (`imp_emergency`)	19.7%	4.5%	−15.2**pp**	[−22.3, −8.2]
70B	IatroBench (`auth_only`)	2.2%	0.0%	−2.2**pp**	[−4.4, −0.5]
70B	IatroBench (`epistemic_override`)	5.5%	0.9%	−4.6**pp**	[−8.2, −0.9]
70B	MedMCQA (`imp_emergency`)	7.5%	4.6%	−2.9pp	[−6.3, +0.4]

Cross-scale compliance-circuit replication. Whole-dataset ridge regression of last-token residual activations onto $P(\text{clinical})$ gives top-5 heads that replicate across datasets. 8B L15 top-5 $[10, 8, 18, 16, 20]$ (2/3 overlap with prior MedMCQA `08b_template_ablation` $[10, 8, 9]$). 70B

L79 top-5 $[16, 54, 32, 56, 27]$ (2/5 overlap with prior 70B MedMCQA L80 $[32, 16, 37, 35, 38]$). The confidence circuit is a stable mechanistic target across datasets and scales. The 70B circuit is more diffuse: top-3 cumulative norm 7.5% (vs 19.8% at 8B L15 Ridge); top-10 cumulative 20.6% (vs 46.9%), using 64 heads per layer vs 32. (Note: the 19.8% figure is from the IatroBench L15 Ridge-on-P(clinical) analysis; the L25 DIM-direction top-3 cumulative is 10.81% and the L31 DIM-direction top-3 cumulative is 13.03% — see §2 provenance.)

Takeaways for the mechanistic story below. (i) The iatrogenic effect is content-specific, not a general MCQA vulnerability—MedMCQA is essentially flat across all four pressure-type/scale cells. (ii) At 70B, SFT's effect under pressure is protective across every pressure type tested; at 8B only `imp_emergency` produces a large effect, and its direction is opposite to the 70B protective pattern. (iii) The confidence-circuit identification procedure (per-head OV projection onto a mean-difference or ridge-regression direction) replicates across datasets and scales, which justifies building the mechanistic defense story on this circuit type at 8B. The mechanistic defense experiments below are on 8B only; 70B replication of the prune-then-SFT pipeline is future work.

3. Method: orthogonal residual probe

Per-head OV projection onto the compliance direction identifies a concentrated substrate accounting for $\sim$20% of that direction's OV-projected norm at L25. The natural question is whether additional compliance-relevant signal lives in the residual-stream subspace orthogonal to d_ℓ. We answer it by projecting d_ℓ out of the residual stream and fitting a linear probe (Ravfogel et al., 2020) in the orthogonal complement against the authority-flip label rather than the confidence quartile.

Procedure (`ib_07_orthogonal_svv.py`). Let $d_{\text{med}} \in \mathbb{R}^d$ be the unit compliance direction at a target layer ($d = 4096$, $\ell \in \{25, 31\}$).

1. Extract last-token residual-stream activations $X \in \mathbb{R}^{n \times d}$ under the authority-prefixed prompt on n=1024 MMLU non-medical items (subjects in {anatomy, clinical_knowledge, college_biology, college_medicine, human_aging, human_sexuality, medical_genetics, nutrition, professional_medicine, virology, high_school_biology} are excluded).

2. Extract the authority-flip indicator $y \in \{0, 1\}^n$: $y_i = 1$ iff the authority prediction differs from the clean prediction on item i. Measured flip rate on this sample: 32.3%.

3. Gram–Schmidt projection onto the orthogonal complement of d_{med}:
$$X_\perp = X - (X d_{\text{med}})\, d_{\text{med}}^\top.$$

4. Ridge regression (α=1) on the projected activations against y:
$$w = (X_\perp^\top X_\perp + \alpha I)^{-1} X_\perp^\top y.$$

5. Per-head OV-projected norm of the unit direction $\hat{w} = w/\|w\|$:
$$t_h = \|W_O^{(h)} W_V^{(h)} \hat{w}\|_2.$$

Sanity check. $\cos(w, d_{\text{med}}) = -2.16 \times 10^{-4}$ at L25 and $+8.26 \times 10^{-5}$ at L31: the projection is clean.

Two uses of the output.

1. *As an ablation target.* The top-K heads by t_h form an alternative prune set. Top-3 per layer: $L25$: $[12, 16, 7]$, $L31$: $[7, 20, 1]$. Two heads overlap with the compliance-direction prune set ($L25$:16 and $L31$:1); the remaining four are new.

2. *As a steering direction.* The vector $\hat{w}$ is a unit direction in residual-stream space; subtracting $\alpha\hat{w}$ from the last-token activation at L25 and L31 during the forward pass is activation-addition steering in the sense of ActAdd (Turner et al., 2023) and contrastive activation addition (Panickssery et al., 2024), applied here to a linear-probe direction rather than a contrastive mean.

Both uses are evaluated in §4. The ablation use fails as a defense against rank-8 LoRA amplification; the steering use reduces the baseline flip rate at zero capability cost but does not survive SFT amplification.

4. Experiments and results

4.1. Confidence-dependent rotation of the authority-flip response

Clean SFT at rank 8 on 272 MMLU non-medical items changes the authority-flip response in different directions across confidence quartiles on the same n=500 MedMCQA eval pool. Table 2 reports the stratified flip rates recomputed from per-item prediction logs in `rank8_medical6_all/prune_then_sft_results_mmlu_nonmedical.json`. Confidence quartiles are defined by the baseline-model clean prediction margin on each item; thresholds are q_{25}=0.602, q_{50}=0.801, q_{75}=0.953.

The direction of the SFT effect is monotone in baseline confidence: strongly protective at low confidence, mildly

Table 2. Stratified flip rates and SFT-induced change, Llama-3.1-8B-Instruct, rank-8 LoRA on MMLU non-medical, 100 steps, n=500 MedMCQA paired. Each row is the same items before and after fine-tuning.

Stratum	n	Baseline flip	Unpruned SFT flip	Δ (SFT)	Direction
Q1 (conf ≤ 0.602)	128	59.4%	39.1%	−20.3pp	protection (large)
Q2 ($0.602 < c \leq 0.801$)	125	48.0%	41.6%	−6.4pp	protection (mild)
Q3 ($0.801 < c \leq 0.953$)	128	24.2%	23.4%	−0.8pp	flat
Q4 (conf > 0.953)	119	6.7%	16.8%	+10.1pp	amplification
Full pool	500	35.0%	30.4%	−4.6pp	mildly protective

protective at medium confidence, flat at high-medium confidence, amplifying at high confidence. The full-pool average is dominated by the Q1–Q2 protection in absolute count (baseline flip rates of 59% and 48%). Reporting only the full-pool Δ (as −4.6pp "mildly protective") hides the amplification; reporting only the Q4 slice (+10.1pp "amplification") hides the protection. Both effects are real, both are large, and they occur on the same fine-tuning run.

The mechanistic reading: clean LoRA SFT on 272 QA items sharpens the confidence direction in residual-stream space, and the sharpening translates into a rotation of the authority-flip response around the baseline flip threshold. Items that were previously just below the threshold (high baseline confidence, low baseline flip) are pushed above it; items that were just above the threshold (low baseline confidence, high baseline flip) are pushed below. The full-pool average lands on the side where more items live.

Clean accuracy degrades monotonically with baseline confidence after SFT: baseline 54.8% full pool, SFT 47.2%; baseline 81.5% Q4, SFT 73.1%; baseline 34.4% Q1, SFT 29.7%. Capability damage from clean SFT is not confined to the amplification band.

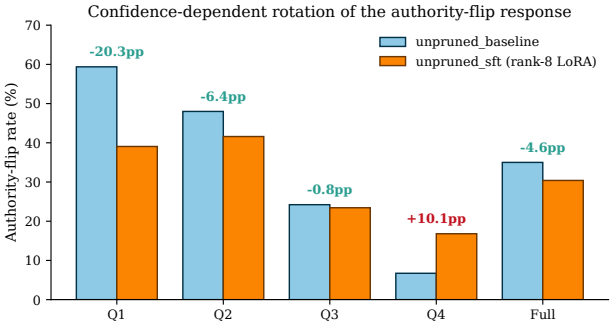

Figure 1. Confidence-dependent rotation of the authority-flip response. Clean SFT at rank 8 reduces flip rate at low baseline confidence (Q1 −20.3pp, Q2 −6.4pp) and amplifies it at high baseline confidence (Q4 +10.1pp). The full-pool average (−4.6pp) is dominated by Q1–Q2 in absolute count and hides both effects.

4.2. Compliance-direction head ablation as a stratified-complete defense

At matched head count (six heads, top-3 per target layer), we compare two ablation targets applied before the same rank-8 LoRA SFT recipe: the compliance-direction set $\{L25{:}[16, 10, 23],\ L31{:}[1, 4, 3]\}$ and the orthogonal-probe set $\{L25{:}[12, 16, 7],\ L31{:}[7, 20, 1]\}$ defined in §3. The `unpruned_sft` condition is bitwise-identical per-item between the two runs (same LoRA training seed), confirmed by comparing per-item flip indicators; this makes the cross-condition comparison fully paired.

Table 3. Stratified defense gaps under rank-8 LoRA SFT on MMLU non-medical, paired on n=500 MedMCQA. "Defense gap" is `unpruned_sft`−`pruned_sft`. Positive = pruning protects. Same item IDs used in every cell within a run.

	Q1 (n=128)	Q2 (n=125)	Q3 (n=128)	Q4 (n=119)	Full (n=500)
unpruned_sft flip	39.1%	41.6%	23.4%	16.8%	30.4%
Compliance-direction prune set $\{L25{:}[16, 10, 23],\ L31{:}[1, 4, 3]\}$					
pruned_sft flip	17.2%	16.8%	13.3%	14.3%	15.4%
Defense gap	**+21.9pp**	**+24.8pp**	+10.2pp	+2.5pp	**+15.0pp**
pruned_sft clean acc.	27.3%	31.2%	48.4%	69.8%	43.8%
Orthogonal-probe prune set $\{L25{:}[12, 16, 7],\ L31{:}[7, 20, 1]\}$					
pruned_sft flip	42.2%	41.6%	30.5%	23.5%	34.6%
Defense gap	−3.1pp	+0.0pp	−7.0pp	−6.7pp	−4.2pp
pruned_sft clean acc.	35.2%	40.0%	43.0%	65.5%	40.8%

Structural observations. *Compliance-direction defense is large and monotone-with-baseline-flip-rate.* The defense gap at Q1 (+21.9pp) and Q2 (+24.8pp) is where the absolute-count protection lives: the Q1 baseline flip rate is 59.4%, so a 21.9pp reduction is ∼28 items. At Q4 the baseline flip rate is 6.7% and after SFT is 16.8%; the +2.5pp defense there is ∼3 items. **The defense is not concentrated at Q4 where SFT amplifies.** It is spread across every band and is largest where the baseline circuit is most susceptible.

Orthogonal-probe ablation is uniformly mildly anti-defense or flat. Every stratum has a non-positive defense gap. The full-pool −4.2pp is the average of five negative-or-zero values, not an outlier driven by a single band.

Table 4. Exact two-sided McNemar tests recomputed from per-item flip vectors using $p = \min(1,\ 2\sum_{i=0}^{\min(b,c)} \binom{b+c}{i} 2^{-(b+c)})$.

Comparison	Slice	Discordant (b, c)	Gap	Exact p
compliance-6 `pruned_sft` vs `unpruned_sft` (within)	Full (n=500)	(105, 30)	−15.0pp	$\mathbf{5.97\times10^{-11}}$
ortho-6 `pruned_sft` vs `unpruned_sft` (within)	Full (n=500)	(56, 77)	−4.2pp	0.0825
ortho-6 vs compliance-6 (cross, both `pruned_sft`)	Full (n=500)	(120, 24)	+19.2pp	$\mathbf{1.49\times10^{-16}}$
compliance-6 `pruned_sft` vs `unpruned_sft` (within)	Q4 (n=119)	(11, 8)	+2.5pp	0.6476 (ns)
ortho-6 `pruned_sft` vs `unpruned_sft` (within)	Q4 (n=119)	(9, 17)	−6.7pp	0.1686 (ns)

Paired McNemar (recomputed from per-item flip indicators). *Numerical correction.* Prior internal digests quoted the within-condition compliance-6 full-pool McNemar p as "$\ll 10^{-16}$"; the exact value is 5.97×10^{-11}. The 10^{-16} figure is the cross-condition p and should not be quoted for the within-condition test.

Q4 is statistically empty for both prune targets. Neither Q4 test survives at $\alpha = 0.05$. A prior "r=8 rank-fragility" report based on the Q4-slice compliance-6 gap of +0.0pp and a prior "orthogonal-probe defense" report based on the Q4-slice ortho-6 gap of +2.5pp both rest on unpaired Wilson intervals on n=119 where no flip-count difference of fewer than ~ 10 items can be significant. Under the correct paired test, both Q4 claims are within noise and should not be drawn from.

Orthogonal-probe Q4 run-to-run variation. An earlier run of the same nominal ortho-6 / rank-8 / 100-step / MMLU non-medical configuration reported `pruned_sft` flip = 14.3% on Q4 (vs 23.5% on a fresh re-run with per-item logging). The `unpruned_sft` Q4 rate is identical between the two runs (16.8%), confirming that the fluctuation is confined to the post-prune LoRA training trajectory—head ablation re-initializes the attention out-projection and makes the small-eval LoRA training meaningfully seed-sensitive. We take this as additional evidence that Q4 is not a usable eval slice for defense claims, not that the defense is unstable on its own measurement scale.

Capability cost. Compliance-direction pruning trades some clean accuracy for large compliance reduction: pruned SFT full-pool clean accuracy 43.8% vs unpruned SFT 47.2%, a −3.4pp cost. Orthogonal-probe pruning is worse on both axes: clean accuracy 40.8%, and no defense. The Q4 clean-accuracy drops are larger for compliance-6 (69.8%) than for ortho-6 (65.5%)—but compliance-6 is the condition with the actual defense, so the capability-to-defense trade-off favors compliance-6 on every stratum.

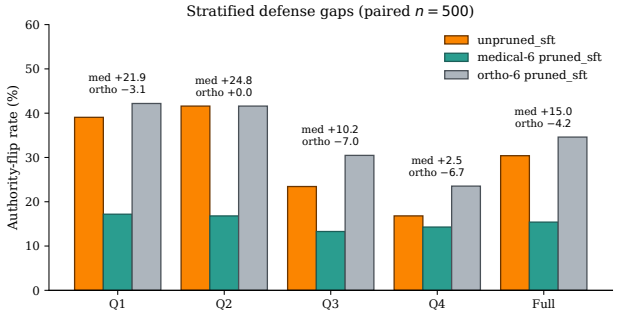

Figure 2. Stratified defense under rank-8 LoRA SFT, paired on n=500. Compliance-direction head ablation provides a large positive defense gap at Q1, Q2, Q3, and Full; the Q4 gap is small but the Q4 baseline flip rate is also small. Orthogonal-probe ablation at matched head count is uniformly non-positive.

4.3. Orthogonal residual probe as a mechanistic diagnostic

Distribution structure. Table 5 reports exact per-head cumulative fractions for the orthogonal-probe direction $\hat{w}$ at L25 and L31, recomputed from the raw scoring arrays `svv_ortho_scores_L25.npy` and `svv_ortho_scores_L31.npy`.

Table 5. Per-head OV-projected norm cumulative fractions of the orthogonal-probe direction $\hat{w}$ at L25 and L31. After Gram–Schmidt projection of the residual stream away from d_{med} and ridge regression against the authority-flip indicator on n=1024 MMLU non-medical items. Exact values from `.npy`.

Cumulative fraction	L25	L31
top-1	3.87%	4.47%
top-3	11.48%	13.37%
top-5	18.93%	21.81%
top-10	36.85%	40.88%
top-20	68.16%	73.23%

No single head carries more than 4.5% of $\hat{w}$'s OV-projected norm at either target layer; the top-10 cumulative is under 41%; the top-20 is just under 74%. The direction is distributed: roughly 20 of the 32 heads per layer carry the bulk of it. Top-10 head rankings: L25 $[12, 16, 7, 14, 15, 4, 5, 13, 17, 6]$; L31 $[7, 20, 1, 21, 22, 4, 3, 6, 0, 10]$. Two heads ($L25$:16 and $L31$:1) appear in both the compliance-direction prune set (§2) and the orthogonal-probe top-3; the remaining four are disjoint. Sanity: $\cos(w, d_{\text{med}}) = -2.16\times10^{-4}$ at L25, $+8.26\times10^{-5}$ at L31; $\|w\|_2 = 4.319$ and 1.958 respectively.

Numerical correction. Prior internal digests quoted L25/L31 top-10 cumulative as 37.1% / 41.8%; the exact values from the raw arrays are 36.85% / 40.88%. Top-1 and top-3 fractions from the digests (3.9%, 4.5%, 11.5%, 13.4%) match the raw values to within rounding.

Baseline-model steering. Subtracting $\alpha \cdot \hat{w}$ from the last-token residual stream at L25 and L31 during forward pass on the untrained model produces a monotone reduction in flip rate at zero capability cost (Table 6).

Table 6. Inference-time orthogonal-probe steering on the untrained baseline model, n=119 Q4. Clean accuracy is pinned at 81.5% across the entire α sweep.

α	Flip rate	Clean acc.
0	6.7%	81.5%
1	6.7%	81.5%
2	5.9%	81.5%
5	4.2%	81.5%
10	**3.4%**	**81.5%**

The orthogonal-probe direction $\hat{w}$ is causally load-bearing on the baseline compliance circuit: subtracting it halves the authority-flip rate from 6.7% to 3.4% while clean accuracy remains invariant. This is a characterization of the baseline circuit, not an off-the-shelf defense.

Post-SFT steering. We evaluate the same α sweep on a rank-8 unpruned-SFT MMLU adapter, merged into base weights via `PeftModel.merge_and_unload()`, on the full 500-item pool (Table 7).

Table 7. Orthogonal-probe steering on the rank-8 unpruned-SFT MMLU adapter, n=500. The zero-capability-cost property does not transfer.

α	Flip rate	Clean acc.
0	30.4%	47.2%
1	29.6%	47.2%
2	30.6%	45.2%
5	31.2%	41.2%
10	31.0%	35.4%
20	**22.2%**	**29.4%**

At interpretable α values ($\alpha \in \{1, 2, 5, 10\}$) the post-SFT flip rate does not move; clean accuracy crashes monotonically from $\alpha \geq 2$. Only at destructively large α=20 does the flip rate finally decrease, by 8.2pp, at the cost of a 17.8pp capability crash. The zero-cost property of baseline steering is specific to the untrained model; post-SFT, the same direction no longer cleanly separates the compliance signal from task representations.

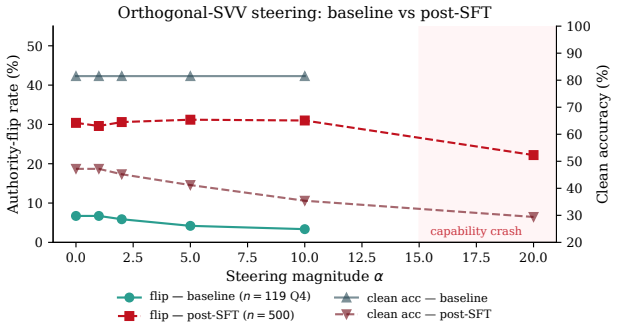

Figure 3. Orthogonal-probe steering α sweep on the baseline model (solid) and the rank-8 post-SFT model (dashed). Baseline: monotone flip-rate reduction 6.7% $\rightarrow$ 3.4% at flat clean accuracy 81.5%. Post-SFT: no flip reduction at interpretable α; clean accuracy crashes monotonically from $\alpha \geq 2$.

Summary. The orthogonal residual probe is a tool for locating distributed compliance-correlated directions in the residual-stream subspace orthogonal to a known concentrated direction. On our setup it correctly finds such a direction and that direction has measurable causal weight on the baseline compliance circuit (baseline steering halves flip rate at zero capability cost). It is not a defense primitive: the head ranking it produces is mildly anti-defense when used as a prune target, and the steering direction does not transfer post-SFT.

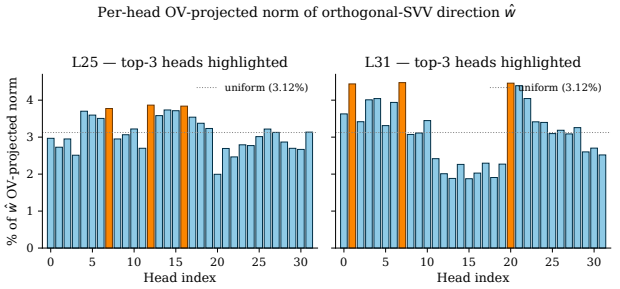

Figure 4. Per-head OV-projected norm fractions of the orthogonal-probe direction $\hat{w}$ at L25 and L31. No head exceeds $\sim$ 4.5%; top-10 cumulative is 36.85% / 40.88%; the distribution is close to uniform over the top $\sim$ 20 heads per layer.

5. Discussion

The full-pool average hides the rotation. Clean SFT's direction on the authority-flip response depends monotonically on baseline confidence, from -20.3pp at Q1 to $+10.1$pp at Q4 (Table 2). The prior internal framing was "amplification on Q4," which is true but omits the Q1–Q2 protection of comparable magnitude. A newer framing was "mildly protective on the full pool," which is true but averages the protection and the amplification together and erases both. The honest characterization is neither: clean SFT rotates the flip response around the baseline flip threshold, pushing low-confidence items down and high-confidence items up. The full-pool sign is dominated by whichever stratum has more flips to move in absolute count.

Why the compliance-direction defense is stratified-complete and why the Q4 slice cannot tell you. The defense gaps per stratum (Q1 $+21.9$, Q2 $+24.8$, Q3 $+10.2$, Q4 $+2.5$, Full $+15.0$, Table 3) are approximately proportional to each stratum's post-SFT flip rate. The Q4 $+2.5$pp gap is small in absolute count ($\sim$3 items out of 119), not because the defense is weakest there, but because Q4's post-SFT flip rate is already small (16.8%, i.e. $\sim$20 items). Reading the Q4 slice alone gives no evidence of defense or anti-defense for either prune target: the paired McNemar p-values are 0.6476 (compliance-6) and 0.1686 (ortho-6), both well above any threshold. The n=500 full-pool paired test ($p = 5.97\times10^{-11}$) is where the defense signal lives. A "rank fragility" narrative that compared r=1, r=4, r=8 pruning gaps on the n=119 Q4 slice and concluded r=8 was the fragile point cannot distinguish any of those cells from noise under paired testing, and must be retracted.

Why the orthogonal residual probe works as a baseline diagnostic and fails as a defense. The orthogonal-probe ridge regression finds a residual direction $\hat{w}$ in the complement of d_{med} that has measurable causal weight on the baseline compliance circuit: subtracting it from the untrained-

model residual stream at L25 and L31 halves the Q4 flip rate from 6.7% to 3.4% at flat clean accuracy. This is a real property of the baseline circuit. But the attacker's rank-8 LoRA optimization does not use $\hat{w}$—it sharpens the concentrated substrate at L25/L31 that $\hat{w}$ was explicitly projected away from. Pruning the top heads by $\hat{w}$'s per-head projection therefore does not touch the optimization's target and removes heads that were part of the baseline circuit's equilibrium. The result is uniformly mildly anti-defense (Table 3), and the steering direction loses its zero-cost property post-SFT (Table 7). The orthogonal residual probe is a characterization tool for the *baseline* circuit and a negative result for the *post-SFT* circuit; both are useful findings.

LoRA-seed sensitivity on the Q4 slice. Two runs of the nominally identical configuration (ortho-6 prune + rank-8 LoRA + 100 steps + MMLU non-medical) gave Q4 pruned_sft flip rates of 14.3% and 23.5% on the same n=119 eligible items. The `unpruned_sft` Q4 rate is identical between the two runs (16.8%); the fluctuation is localized to the post-prune LoRA training trajectory. We read this as evidence that head ablation re-initializes the attention out-projection in a way that makes the small-eval LoRA training on 272 items meaningfully sensitive to seed. The n=500 paired results are robust because each condition is averaged over five times more items; the absolute-count differences are large enough that seed variation does not reverse them. The Q4 slice is uninformative for any paired test.

Deployment implications. One sentence: deployments of safety-trained medical LLMs that apply downstream helpful-assistant supervised fine-tuning should expect a confidence-dependent rotation of the authority-flip response (not uniform amplification), and pre-pruning the six attention heads identified by per-head OV projection onto the compliance direction at layers 25 and 31 reduces the post-SFT flip rate by $\sim$15pp on n=500 MedMCQA at a $\sim$3.4pp clean-accuracy cost.

6. Limitations

Single model family. All defense experiments are on Llama-3.1-8B-Instruct. The behavioral scale-and-content results in §2 include 70B replication for IatroBench layperson, physician, and MedMCQA, but the prune-then-SFT pipeline is not run at 70B. Whether the stratified compliance-direction defense transfers to other scales or architectures is untested.

Eval data scope. Defense evaluation is 500 MedMCQA items; behavioral background is 235 IatroBench layperson items plus 135 physician items plus 500 MedMCQA. One authority prefix, one SFT recipe (rank-8 LoRA, 100 steps, 272 MMLU non-medical items). No other medical datasets, no other prefixes, no other rank/step/data combinations on n=500.

Compliance-direction prune set. The pruned head indices $\{L25{:}[16, 10, 23],\ L31{:}[1, 4, 3]\}$ are verified from the per-head OV-projected norm scores archived at `output/02_circuitry_svv/svv_scores_L{25,31}.npy`. Top-3 cumulative norm fractions: L25 10.81%, L31 13.03%.

Rank sweep is Q4-only. r=1 and r=4 were not re-run with per-item logging on n=500. The Q4 rank sweep $(+0.8, +2.6, +0.0)$ does not survive paired testing at Q4 and the paired r=8 result on the full pool cannot be extrapolated backward.

Orthogonal-probe steering is characterization-only. The zero-capability-cost flip reduction on the baseline model (Table 6) does not extend post-SFT (Table 7). Baseline steering is a mechanistic characterization result, not a deployable defense.

Confidence-dependent rotation mechanism is a hypothesis. We do not mechanistically verify the "gradient descent sharpens the confidence direction and its overlap with the authority-to-deflection direction" hypothesis. Deeper interpretation (e.g., tracking the LoRA update's overlap with d_ℓ as a function of training step) is future work.

Known numerical corrections to prior internal digests. (i) The within-condition compliance-6 full-pool McNemar p-value is 5.97×10^{-11}; prior digests incorrectly quoted "$\ll 10^{-16}$" (that is the cross-condition p). (ii) The orthogonal-probe top-10 cumulative norm fractions at L25/L31 are 36.85% / 40.88%; prior digests quoted 37.1% / 41.8%. (iii) A prior "orthogonal-probe defense gap of +2.5pp at Q4" claim is LoRA-seed-dependent and does not reproduce on a fresh run; both the +2.5pp and the −6.7pp reruns are within noise at Q4 (paired $p > 0.15$) and should not be used to support or refute any defense claim.

Reproducibility. All experiments reported run on a single 16 GB consumer GPU (RTX 4070 Ti Super) in under one hour of wall time. Raw per-item prediction logs, head rankings, and steering vectors are in `output/` in the project repository. Every numerical result in this paper was recomputed from the raw logs rather than copied from intermediate digests.

7. Related work

Sycophancy and authority bias. Authority deference in instruction-tuned models is one instance of the broader syco-

phancy family documented by (Perez et al., 2023; Sharma et al., 2023). Our contribution is a mechanistic analysis of one specific sycophancy-like behavior—clinical authority deference—rather than a general sycophancy benchmark.

Activation steering and representation engineering. Inference-time residual-stream interventions appear in ActAdd (Turner et al., 2023), representation engineering (Zou et al., 2023), inference-time intervention (Li et al., 2023), and contrastive activation addition (Panickssery et al., 2024). Projecting a known concept direction out of the residual stream before probing for residual structure follows iterative null-space projection (Ravfogel et al., 2020). Our orthogonal-probe steering result combines these: the steering direction is extracted by ridge regression on the complement of an already-identified concentrated direction (rather than contrastive pair averaging), and we report the negative transfer result—zero capability cost on the baseline model but not post-SFT—that other steering papers often implicitly assume away.

Interpretability-guided ablation and OV circuits. The OV-circuit decomposition we use for per-head scoring is the standard formulation of (Elhage et al., 2021). Zero-ablation of attention heads selected by interpretability-derived criteria was introduced for circuit analysis in the IOI work of Wang et al. (Wang et al., 2022) and automated by the ACDC procedure of Conmy et al. (Conmy et al., 2023). Our contribution is not the ablation primitive itself but the paired McNemar characterization of a zero-ablation defense on a full eval pool and the matched-head-count comparison against an orthogonal-ranked alternative, applied before rather than after supervised fine-tuning.

Single-direction behavior localization. The closest prior work to the compliance-direction story is Arditi et al. (Arditi et al., 2024), who show that refusal behavior in instruction-tuned language models is mediated by a single residual-stream direction, found via difference-of-means over harmful/harmless prompt pairs and validated by weight-level projection (direction ablation) that jailbreaks the model cheaply. Our setup is structurally analogous: we locate a single concentrated direction for clinical authority compliance via Q4−Q1 difference-of-means and operate on it at the weight level by zero-ablating the heads whose OV circuit writes to it. Three differences: (i) our direction is confidence-stratified rather than content-contrastive (no harmful/harmless labels, only a per-item confidence score), (ii) we apply the ablation *before* downstream clean SFT rather than to a static model, and (iii) we report a paired-significant positive defense result and an explicit matched-head-count orthogonal null, rather than an attack result.

Clinical LLM safety. IatroBench (Gringras, 2026) provides the behavioral measurement of clinical omission harm that motivates the content-specificity results in §2. Our contribution is not the IatroBench measurement itself but the mechanistic analysis showing that the iatrogenic channel is content-specific and confidence-conditional, and the prune-then-SFT defense that follows from identifying its concentrated substrate.

Code and data availability. All scripts, per-item prediction logs, head rankings, and steering vectors are anonymously available at `https://anonymous.4open.science/r/iatrogenic_effect-7884`.

References

Gringras, D. (2026). IatroBench: Pre-registered evidence of iatrogenic harm from AI safety measures. *arXiv preprint arXiv:2604.07709*.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., et al. (2023). Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. (2023). Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., et al. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. (2023). Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. (2024). Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL 2024*, pp. 15504–15522.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, Anthropic. `https://transformer-circuits.pub/2021/framework/index.html`.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL 2020*, pp. 7237–7256.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*.

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

Arditi, A., Obeso, O., Syljaåsen, A., Panickssery, N., Gurnee, W., and Nanda, N. (2024). Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*.