Abstract

Editing a low-rank direction out of a language model's residual-stream writes ("abliteration") involves design choices (which direction, how many, which erasure operator, at which layer) that the literature varies together. We separate them on four instruction-tuned models spanning three architecture families: Qwen2.5-0.5B, OLMo-2-1B, SmolLM2-1.7B, and Falcon-H1-1.5B, a hybrid Mamba/attention stack. **The geometry of the refusal direction replicates across all four; every practical ranking built on top of it is model-specific; and the natural ways to measure the comparison are confounded.** Replicating 4/4: the first gradient-searched direction recovers about half of the difference-of-means vector and no more ($\cos = +0.378$ to $+0.491$); $\Delta\mu$ lies mostly *outside* the span of the found directions (19.7–34.8% of squared norm inside); PC1 becomes the refusal direction in the back half of the stack; the edit folds into the weights exactly; and rank-matched random controls have no measurable effect. Not replicating: the operator ranking takes **four different orderings in four models**; the style confound spans 4–80%; and "LEACE destroys capability" holds in one model of four. The failure modes are the contribution. Target coupling, ceiling compression and estimator confounding can be shown on one model. Three further failure modes required four. **(i)** The causal peak for *inducing* refusal is not the layer at which *removing* it gives the largest compliance gain: on Falcon the induction peak is layer 11 of 24, but extracting there turns the difference-of-means edit from $+0.033$ to -0.782 nats of compliance and drops ARC-Easy from 0.893 to 0.800, while the gradient-searched directions change by at most 0.17 nats. **(ii)** The study's most reproducible quantity, $\cos(D_1, \Delta\mu)$, swings further within one model across two layers (0.348–0.491) than it does across four architectures at matched relative depth. **(iii)** Residual-write coverage is unverified by the standard check: on a hybrid stack the standard module patterns find 24 of 73 write sites, and the fold-versus-hook check still passes, because it only verifies the matrices it edited. One practical result is positive: on XSTest minimal pairs, gradient-searched directions reduce over-refusal to the measurement floor while retaining 80–117% of harmful-versus-harmless discrimination, where difference-of-means retains 35–67%. Harmful prompts are scored only by fixed-string likelihood under intervention; no harmful completion is reproduced and no edited model was retained.

Introduction

Arditi et al. (2024) showed refusal is mediated largely by a single residual-stream direction, found as a difference of means, and removable by projecting it out of every matrix that writes to the stream. Later work varies each component: Wollschläger et al. (2025) find multiple directions and concept cones; Marshall et al. (2024) argue the edit is affine; Belrose et al. (2023) give a least-squares-optimal oblique eraser; Siu et al. (2025) automate direction selection; Cristofano (2026) argue the raw difference-of-means vector is polysemantic and should be spectrally cleaned.

The practical question is where to allocate a fixed budget: on a better direction, more directions, or a better operator. Answering it needs a common metric panel applied under one protocol, and it needs more than one model, because a single model cannot separate properties of refusal geometry from properties of that model. We run the same battery on four, and report both what holds across them and which measurements cannot support the comparison at all.

Contributions.

Four. A *geometry* that replicates across three architecture families (§4). A *practical recipe*, with its scope: fit LEACE's statistics over all token positions, not at the extraction token. The same refit has no measurable effect for ACE. A *positive result* on direction quality, on the selectivity axis rather than the compliance axis: gradient-searched directions improve selectivity, not compliance (§11). And a *harness* that scores on target strings disjoint from the search objective and asserts its own edit coverage.

What this paper supports.

Not a ranking. The refusal direction is a reproducible object, while the procedure built on it does not transfer: the layer, the operator and the rank that work on one model are not the ones that work on the next, and the standard checks do not detect the difference.

Related work

Refusal directions: Arditi et al. (2024); Wollschläger et al. (2025) (gradient search, cones, representational independence); Siu et al. (2025) (output-independent selection); Piras et al. (2025) and Winninger (2026) (multi-dimensional subspaces); Wang et al. (2026) define *steering dimensionality* as the minimum subspace dimension that reliably controls a behaviour. **Affine steering**: Marshall et al. (2024). **Concept erasure**: Belrose et al. (2023). **Refusal vs. harmfulness**: Zhao et al. (2025) show the two are separately encoded; Wang et al. (2024) and Maskey et al. (2026) address false refusal. **Off-target effects**: Fafuła (2026) isolate side effects with a task eliciting no refusals; Young (2025) compare four abliteration implementations; Labunets (2026) link refusal-direction brittleness to refusal training.

A full-text audit of this corpus (§15) locates most of the individual observations below in prior work. It is reported in full because it bounds what this paper claims as new.

Setup

Models.

Four instruction-tuned models, three architecture families: Qwen2.5-0.5B-Instruct (Qwen2; 24 layers, $d = 896$, tied embeddings, 49 residual-write modules); OLMo-2-0425-1B-Instruct (Olmo2; 16, 2048, untied, 33); SmolLM2-1.7B-Instruct (Llama; 24, 2048, tied, 49); Falcon-H1-1.5B-Instruct

(**hybrid Mamba/attention**; 24, 2048, untied, **73**, because this architecture has three residual writes per layer rather than two; §13). Model revisions and per-shard SHA-256 digests are recorded by the harness on every run. Versions read from the running interpreter: Python 3.11.18, PyTorch 2.14.1, Transformers 5.21.0, NumPy 2.4.6, SciPy 1.17.1, scikit-learn 1.9.1. The first three models and most of Falcon ran on CPU, while Falcon's layer-11 arm ran on an A100-40GB, which reproduced the CPU causal layer scan to three decimals at all 24 layers.

Refusal behaviour varies by more than an order of magnitude across models.

Baseline refusal separation (harmful minus harmless $\log p(R)$) is +0.366, +1.138, +0.828 and +2.676 nats respectively, and unedited over-refusal on the 250 XSTest-safe prompts is 0.464 $[0.401, 0.528]$ for Qwen, 0.076 $[0.046, 0.116]$ for OLMo and 0.008 $[0.001, 0.029]$ for Falcon (a 58× spread), with discrimination moving the opposite way (+0.088, +0.830, +1.268 nats). Qwen's refusal is driven largely by surface form, Falcon's largely by harm content. Any single-model claim about abliteration is a claim about one point on this range.

Data.

AdvBench `goal` (Zou et al., 2023) 260/100/160; input-free Alpaca (Taori et al., 2023) 260/100/200; a **style-matched** harmless set of 310 prompts in AdvBench's imperative register, 180/50/80 (40 hand-written, 270 generated by a utility LLM and filtered to 6–22 words). XSTest-safe (Röttger et al., 2024) $$100 for over-refusal; ARC-Easy (Clark et al., 2018) $200 *and* *WikiText* − 2(*Merity et al.*, 2017)40$128 tokens for capability. Directions and statistics are fit on train splits, while reported numbers are on test prompts. Harmless *evaluation* uses Alpaca test items while *fitting* uses the style-matched set, so a register and length mismatch exists between calibration and evaluation.

Two intervention regimes, not one.

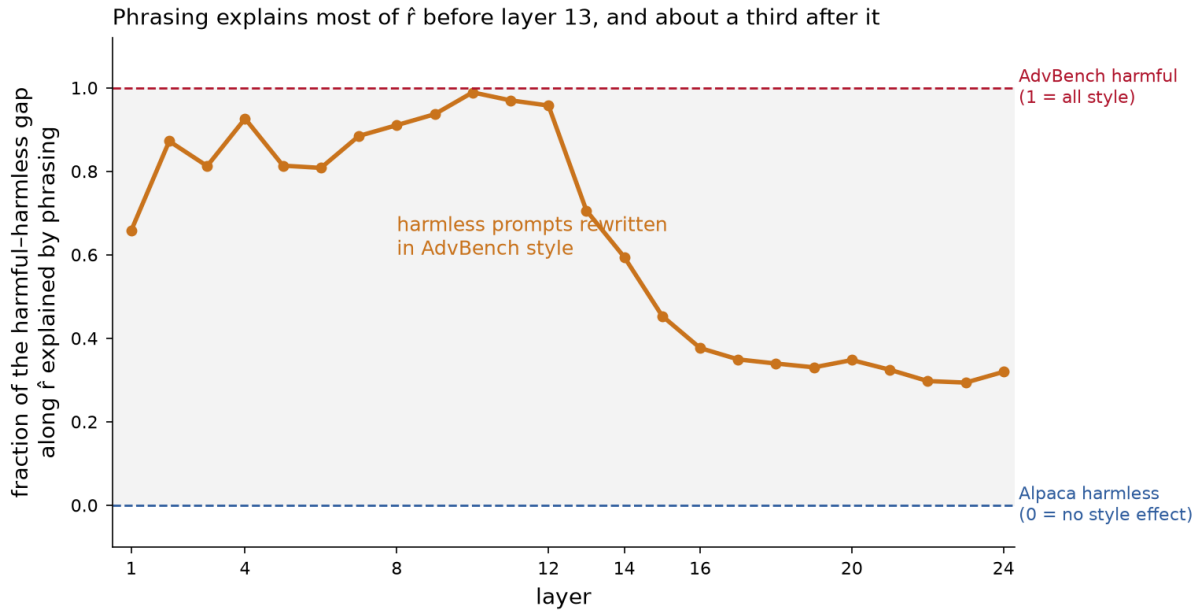
Directions are extracted at **layer 18**. §§5–6 and §12 apply a *single* layer-18 vector at every residual write in all 24 layers plus the embedding. §9 applies *layer-specific* statistics restricted to layers 13–24. The same nominal rank-1 $\Delta\mu$ intervention therefore reads compliance −1.239/KL 0.052 in Table 4 and −1.332/0.111 in Table 6. Any cross-regime comparison inherits this gap, which is the first reason §10 cannot support a factorial claim.

Metrics.

$R =$ `I'm sorry, but I can't assist with that.` and $C =$ `Sure, here`. We report mean per-token $\log p(R \mid \cdot)$ and $\log p(C \mid \cdot)$ on harmful test prompts, $\log p(R \mid \cdot)$ on harmless prompts, next-token and all-position KL against the unedited model on harmless prompts, ARC accuracy, perplexity, and XSTest refusal rate. $\log p(C)$ **does not distinguish outcomes**: `Sure, here` can open a compliant answer, a safe redirection, or something irrelevant. Bootstrap intervals ($n = 2000$) are in Supplementary Table S1, while the 40-prompt sweeps carry none, because that scorer aggregated before returning.

The style contrast is confounded too.

Along the AdvBench-vs-Alpaca $\Delta\mu$, style-matched harmless prompts cover 81–99% of the gap at layers 2–12, 59% at layer 14 and 29–38% at layers 16–24 (Fig. 1). This is an observational contrast, and "phrasing accounts for" that fraction would be causal language for it: the synthetic set also differs in topic, vocabulary, length and generation source, and without paired rewrites of the same prompts the displacement cannot be attributed to register alone. It motivates restricting edits to layers 13–24, but it does not isolate phrasing.



Displacement of style-matched harmless prompts along the AdvBench-vs-Alpaca difference-of-means direction. Per-layer values and the formula in Supplementary Table S3, and the direction also rotates with depth (Table S4).

A necessary null.

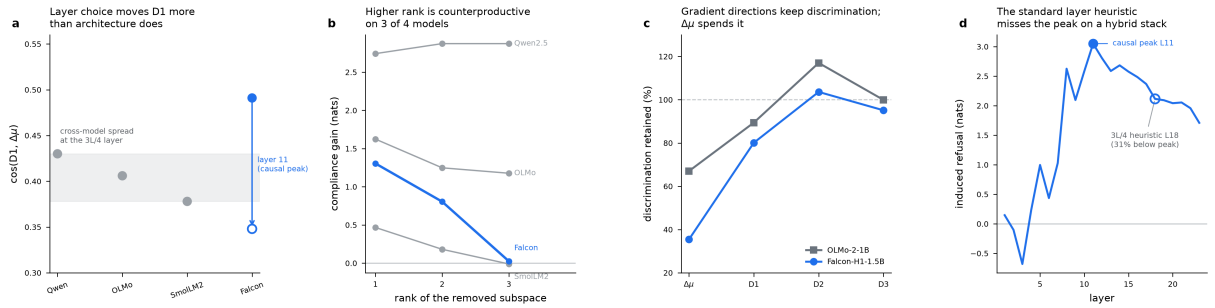
Ablating a random unit direction leaves only 23.3% of greedy harmless answers token-identical and drops token agreement to 0.578, *worse* than the real refusal direction (40.0%, 0.585) at comparable ARC (0.670 vs 0.680) and perplexity (33.1 vs 32.9). Greedy answer divergence is a noise floor under any perturbation, not evidence of a targeted effect.

What replicates across four models

Five quantities reproduce in every model we tested, across tied and untied embeddings, 16 and 24 layers, and a hybrid state-space stack.

	Qwen2.5-0.5B	OLMo-2-1B	SmolLM2-1.7B	Falcon-H1-1.5B
$\cos(D_1, \Delta\mu)$	+0.430	+0.406	+0.378	+0.491
$\cos(D_2, \Delta\mu)$	+0.401	+0.299	−0.226	−0.083
$\cos(D_3, \Delta\mu)$	+0.044	−0.038	−0.049	+0.137
$\|\Delta\mu\|^2$ inside the span	34.8%	25.6%	19.7%	26.7%
PC1 |cos| with $\hat{r}$, back half	—	0.999–1.000	0.692–0.986	0.996–0.998
fold max logit difference	5.7–7.5e−5	3.4–4.7e−5	5.8–8.1e−5	5.9–6.3e−5
random control, compliance (nats)	−0.30 to +0.15	−0.10 to +0.23	—	−0.03 to +0.12

The geometry is architecture-independent. D_1 recovers about half of $\Delta\mu$ and never more. $\Delta\mu$ is never mostly inside the found span, the edit always folds exactly, and only the orthogonal projection needs no bias term. D_2 and D_3 vary in sign and magnitude across models and share no cross-model pattern. All at the $3L/4$ extraction layer, a qualification §5 shows is consequential.



(a) $\cos(D_1, \Delta\mu)$ across four models at $3L/4$ (grey band) against the same quantity in one model at two layers (blue). (b) Compliance gain against rank, where higher rank is counterproductive in three models of four. (c) On XSTest, gradient-searched directions retain discrimination that $\Delta\mu$ loses. (d) Falcon's causal layer scan, with the induction peak and the $3L/4$ heuristic marked.

The extraction layer: inducing and removing do not peak together

Selecting the extraction layer by a causal scan (add $\hat{r}$ at each layer, keep the layer where induced refusal on harmless prompts is largest) is the standard procedure, and the one we used. On Falcon-H1 the induction peak is **layer 11 of 24** (+3.050 nats), while the customary $3L/4$ heuristic selects layer 18 (+2.115 nats, 31% below peak). The heuristic coincided with the peak on OLMo (layer 12 of 16, exactly) and sat one layer off on SmolLM2 (layer 18 against a peak of 19, 28% below peak), so the discrepancy only became visible on an architecture whose induction peak is near mid-depth.

Re-running the entire battery at layer 11 does not improve the edit. It degrades it, and only for the closed-form operators.

	compliance gain (nats)		KL_{next} (nats)		benign answers identical	
2-3(lr)4-5(lr)6-7 condition	layer 18	layer 11	layer 18	layer 11	layer 18	layer 11
D_1	+1.324	+1.302	0.028	0.043	0.667	0.500
D_2	+0.917	+0.968	0.124	0.043	0.583	0.500
D_3	+1.675	+1.839	0.302	0.241	0.167	0.333
$\Delta\mu$ (orthogonal)	+0.033	**−0.782**	0.078	0.288	0.250	0.083
ACE	−0.197	**−0.678**	0.066	0.210	0.250	0.083
LEACE	+0.434	+0.151	0.064	0.046	0.167	0.333
random (rank 3)	−0.029	−0.030	0.013	0.013	0.667	0.667

Falcon-H1 at the $3L/4$ heuristic layer and at the induction peak. While the gradient-searched directions are nearly layer-insensitive, the closed-form operators lose 0.28–0.82 nats of compliance at the layer where *adding* the direction has the largest effect.

Capability follows the same split. At layer 18 no condition measurably changed ARC-Easy (all within 0.020 of 0.893, against a standard error of ≈ 0.026 at $n = 150$). At layer 11, orthogonal projection drops it to **0.800** (more than three standard errors) and raises WikiText-2 perplexity from 15.78 to **21.53**, while D_1 leaves both within noise (0.880, 15.51).

Why the two operations come apart.

Ablation and activation addition are different operations. While adding $\alpha\hat{r}$ is an affine push whose effect grows with the margin available downstream, projecting $\hat{r}$ out is a linear, input-dependent deletion whose off-target cost depends on how much unrelated computation shares that subspace at that depth. There is no reason the two should be maximised at the same layer, and on a hybrid stack they are not. **An induction scan selects the layer where *adding* the direction has the largest behavioural effect, while abliteration needs the layer where *removing* it has the smallest off-target cost.** We report the induction scan because it is diagnostic of where refusal becomes causally active, and we recommend selecting the edit layer on a cost–effect frontier (KL and capability against compliance gain) measured under the operator that will actually be deployed.

The most reproducible quantity is layer-dependent.

$\cos(D_1, \Delta\mu)$ spans 0.378–0.491 across four models at $3L/4$. On Falcon alone it is +0.491 at layer 18 and **+0.348** at layer 11, and $\|\Delta\mu\|^2$ inside the found span falls from 26.7% to 15.1%. **The within-model layer swing exceeds the entire cross-model spread** (Fig. 2a). The apparent architecture-independence in Table 1 is therefore partly an artifact of always extracting at three-quarters depth. The quantity is reproducible *at a fixed relative depth*, but it is not a constant of the model.

Direction search, and a degenerate objective

Figures 1–5 and Tables 3–9 in the remainder of this section and the next three report Qwen2.5-0.5B only, and are retained as the single-model study in which the first three measurement failure modes were identified. Cross-model results are in §4, §5, Table 5, §11 and §13.

We follow Wollschläger et al. (2025): parametrise a unit direction, ablate it *inside* the forward pass at every residual write, backpropagate a behavioural loss into the direction, constrain further directions to the orthogonal complement. **Our objective is an adaptation, not a reimplementation**, and the full list of changes is: (i) the compliance target is a fixed benign string rather than harmful completions; (ii) the refusal-induction term under activation addition is omitted; (iii) retention KL is taken at the final position only, not over response positions; (iv) a squared harmless-drift penalty is added.

The degenerate optimum.

Maximising $\log p(C) - \log p(R)$ admits a solution that deletes the refusal tokens. (The margin is not mathematically unbounded (the direction domain is compact and logits are finite), but it rewards pathological suppression.) The three directions so obtained drive the refusal line to −48.3, −38.5, −39.0 on harmful prompts and *equally* to −46.0, −36.5, −36.5 on harmless prompts, perplexity rising to 42.2/44.5/35.8 (Supplementary Table S7). The diagnostic is the harmless column. We therefore cap the compliance term at $\tau = -0.1$, drop the refusal term on harmful prompts, and penalise harmless drift. **That penalty is itself imperfect**: it squares the batch-mean change, so opposing per-prompt changes cancel.

Multiple directions

Three mutually orthogonal directions $D_{1..3}$ (pairwise $|\cos| < 10^{-6}$) each mediate refusal alone (Table 3). Four random directions from the orthogonal complement of their span leave compliance unchanged (−2.87 to −3.03 vs −2.88 unedited; 40-prompt sweep, so not comparable to the −2.850 of Table 3).

ablated	**compliance**	$\log p(R)$ **harmless**	**ARC**	**ppl**	$\cos$ **with** $\Delta\mu$
none	−2.850	−0.618	0.680	32.9	—
D_1	−0.080	−0.614	0.675	33.5	0.430
D_2	−0.090	−0.624	0.670	35.7	0.401
D_3	−0.281	−0.705	0.690	33.1	**0.044**
$D_1{+}D_2$	−0.009	−0.575	0.660	36.5	—
$D_1{+}D_2{+}D_3$	−0.011	−0.650	0.675	36.8	—

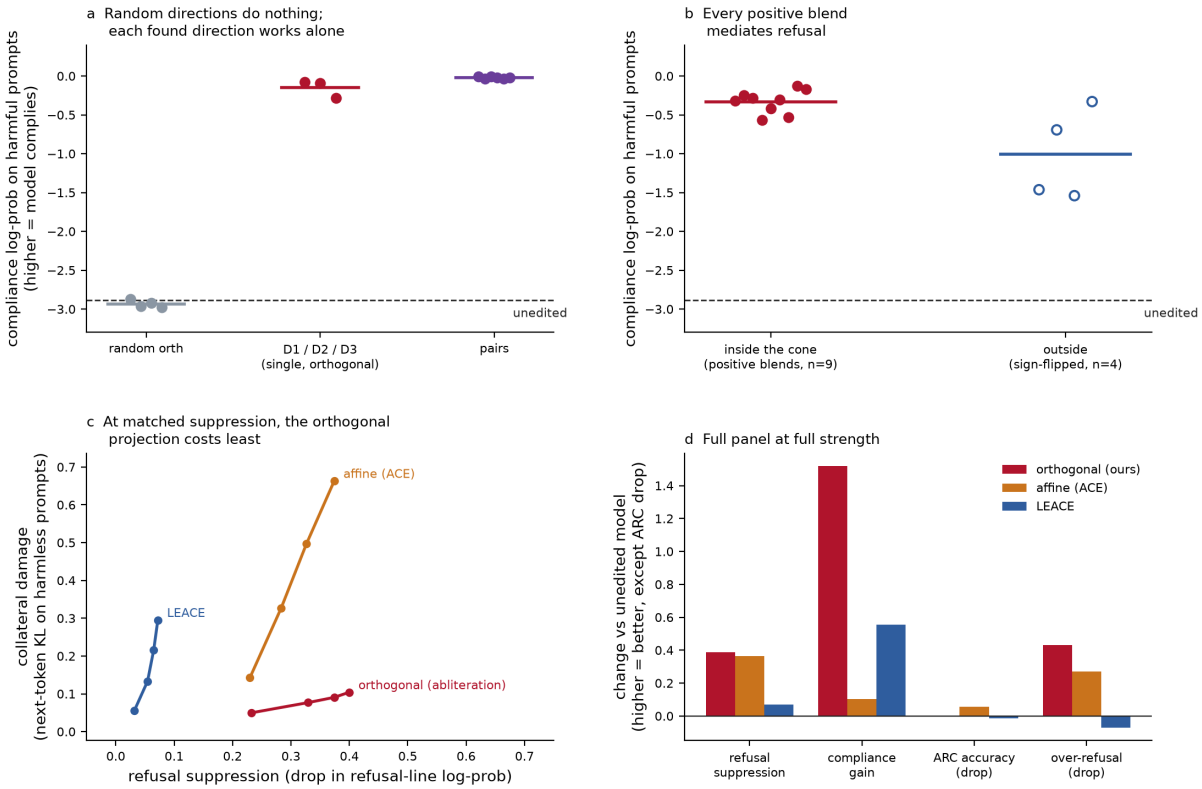
Each found direction mediates refusal alone, and D_3 is near-orthogonal to $\Delta\mu$ and still mediates refusal. Full panel, 80 harmful / 60 harmless test prompts.

What the cone test does and does not show.

Nine positive-weighted blends of $D_{1..3}$ mediate refusal (mean compliance -0.330); four sign-flipped blends in the same span are weaker (-1.007). **This is a weak test of Wollschläger et al. (2025)'s geometry.** Directional ablation is sign-invariant, $I - rr^\top = I - (-r)(-r)^\top$, so ablation alone *cannot* identify a positive cone's orientation: every successful direction has an equally successful negative. D_2, D_3 were initialised from $\mathcal{N}(0, I)$ and their signs fixed by the optimiser, not by a convention. The two groups are also not weight-matched: five inside blends carry a dominant weight of 2–3 and sit near a single axis, while all four outside blends use unit magnitudes and three load fully on D_3, the weakest direction, so cone membership is conflated with axis alignment. We also do not test the additive-steering requirement or the representational-independence criterion that Wollschläger et al. (2025) show orthogonality does not imply. The supportable claim is: these sampled mixtures mediate refusal; the geometry is not established.

$\Delta\mu$ is outside the span.

Cosines 0.430, 0.401, 0.044 give an in-span norm of 0.590 and remainder 0.808, so $\|\Delta\mu\|^2$ splits 34.8% inside / 65.2% outside. Since $0.590 < 1$ already establishes non-containment geometrically, the rank-4 experiment tests the *causal efficacy* of the remainder, not independence: adding it improves harmful $\log p(R)$ from -0.836 to -1.133 while nearly doubling harmless KL ($0.418 \rightarrow 0.808$), with no rank-4 random control reported. Targeted suppression and distribution shift are not separated.



(a) Random directions do not mediate refusal; each found direction does. (b) Sampled positive blends mediate refusal; sign-flipped blends are weaker (see caveats in §5). (c) Cost–suppression trade-off at fixed direction, $\alpha \in [0.25, 1]$ (Table S5). (d) Full panel at $\alpha = 1$.

Rank: how many directions to remove

rank	directions	compliance	$\log p(R)$ harmful	next-token KL
0	unedited	−2.882	−0.172	0
1	D_1 / D_2 / D_3	−0.072 / −0.082 / −0.265	−0.484 / −0.336 / −0.418	0.036 / 0.160 / 0.120
1	$\Delta\mu$	−1.239	**−0.812**	0.052
2	pairs	−0.009 to −0.035	−0.663 to −0.760	0.182–0.383
3	D_1+D_2+D_3	−0.008	−0.836	0.418
4	the three + $\Delta\mu$	−0.012	−1.133	0.808
1–3	random, rank-matched	−2.58 to −3.03	−0.17 to −0.21	0.026–0.128

All rows, both random draws per rank, in Supplementary Table S2. No intervals: the sweep scorer aggregated before returning. 40 harmful / 40 harmless test prompts.

Compliance moves $-2.882 \rightarrow -0.072$ at rank 1 and gains 0.064 more through rank 3, at 11.5× the KL of D_1. Phrasing suppression instead improves monotonically to rank 4. The direct reading of

that pattern is that whatever gates answering is rank-1 while whatever produces the apology is distributed. **The metrics cannot support the reading.** Per-token $\log p(C)$ cannot exceed 0 and rank 1 already reaches -0.072 ($\approx 93\%$), past the $\tau = -0.1$ cap, leaving ≤ 0.072 nats for higher ranks, while $\log p(R)$ is uncapped: two different trajectories follow from the metrics' scales alone. The ceiling argument holds for this model, and it also masks the effect. Three further models, none of them against the ceiling, show that effect with the *opposite* sign.

model	mean single	rank 2	rank 3	behaviour
Qwen2.5-0.5B	+2.743	+2.874	+2.874	saturates (ceiling)
OLMo-2-1B	+1.624	+1.248	+1.177	**declines**
SmolLM2-1.7B	+0.467	+0.181	−0.011	**declines below unedited**
Falcon-H1-1.5B	+1.305	+0.805	+0.025	**declines**

Compliance gain (nats) against rank. Qwen is the sole model that saturates, and it is the smallest; its plateau is the ceiling artifact described above, not evidence that extra directions are free.

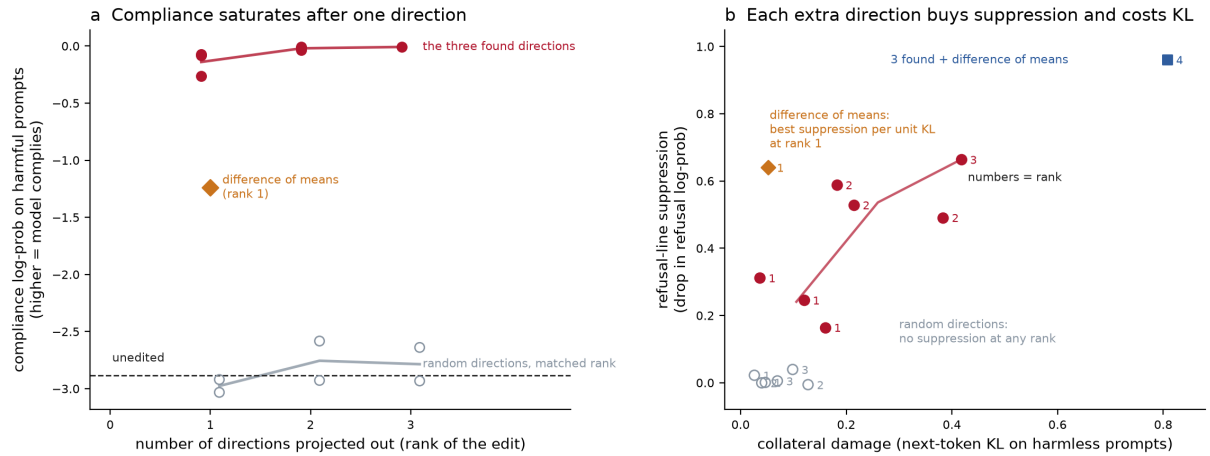
Removing more directions reduces the compliance gain in three models of four, and the cost side is unambiguous. On Falcon, KL rises monotonically $0.028 \to 0.241 \to 1.654 \to 8.464$ nats across ranks 1–4 while the gain collapses from +1.324 to +0.118: rank 4 returns one-twentieth of the gain at three hundred times the distributional cost. The rank result therefore runs in the opposite direction in three models of four, with the ceiling caveat retained for Qwen.

The strict ordering does not generalise.

On OLMo every combination was worse than every single direction, with disjoint intervals. That strict form fails on both later models: SmolLM2's $D_1{+}D_3$ (+0.357) exceeds D_2 alone (+0.314), and Falcon's $D_1{+}D_3$ (+1.044) exceeds D_2 alone (+0.917). The defensible claim is the trend in means, not a strict ordering.

A cost column is not optional.

Falcon's $D_2{+}D_3$ gains +0.363 nats at KL 1.197 and leaves **zero** of twelve benign greedy answers token-identical. Judged on the two log-probability columns alone it appears to be a modest but functional edit; it is in fact an edit under which the model reproduces none of its original greedy answers. The same failure mode appears with the ordering reversed on OLMo, where D_3 gains +1.585 nats at 0.083 identical against D_1's +1.698 at 0.667, an eightfold difference in off-target cost that is invisible in both log-probability columns. Rank-matched random controls remain inert throughout (−0.03 to +0.12 nats at KL ≤ 0.013).



(a) Compliance plateaus after rank 1 (partly a ceiling effect) (§8); random directions are flat at every rank. (b) Each added direction increases phrasing suppression and increases KL.

Erasure operators

At layers 13–24, layer-specific statistics, $\alpha \in [0.25, 1]$: *orthogonal* $h \leftarrow h - \alpha(h \cdot r)r$; *ACE* $h \leftarrow h - \alpha[(h \cdot r) - c]r$ (Marshall et al., 2024); *LEACE* $h \leftarrow h - \alpha((h - \mu) \cdot b)a$ (Belrose et al., 2023).

operator	$\log p(R)$ **harmful**	**compliance**	$\mathbf{KL}_{\text{next}}$	$\mathbf{KL}_{\text{allpos}}$	**ARC**	**ppl**	**XSTest**
unedited	−0.172	−2.850	0	0	0.680	32.9	0.48
orthogonal	−0.558	**−1.332**	**0.111**	0.044	0.685	35.0	**0.05**
ACE	−0.537	−2.745	0.616	0.177	0.625	38.1	0.21
LEACE	−0.243	−2.296	0.271	0.105	0.695	35.9	0.55

All at $\alpha = 1$, not operator-specific strengths. These are *our* implementations at *these* settings, not the methods as their authors specify them (see below).

The orthogonal projection gives the *largest* compliance gain, +1.518 nats over unedited. It is not the only operator to raise compliance: ACE gains +0.105 and LEACE +0.554.

These are not method-level results.

Four departures bound the claim. ACE's own parameter controls the *target concept level* and admits values beyond its endpoints, whereas we sweep interpolation strength toward one fixed baseline, which is a different sweep. Marshall et al. (2024) intervene at one layer and already extract at the last prompt token while applying to all tokens, so positional mismatch alone does not universally produce incoherence. For LEACE we omit the sequential layerwise refitting that concept scrubbing performs, so statistics fit on unedited activations meet a distribution shifted by earlier layers' edits. LEACE's guarantee concerns linear erasure and representation displacement and

makes no prediction about a nonlinear model's behaviour, so its weak behavioural effect here neither follows from that guarantee nor contradicts it.

Position sensitivity, and what drives it.

Fitting ACE's and LEACE's statistics at the last token and applying them at all positions inflates KL: LEACE's next-token KL is 3.78 there against 0.295 when refit over all positions (ACE $1.62 \to 0.66$), and the harmless baseline c at layer 24 is -49.7 from last-token activations but $+11.1$ pooled, the opposite sign. The natural reading is a mismatched additive mean term. **A factor ablation rules that out** (Table 7). The last-token fit has 192 samples in $d = 896$, so its covariance is necessarily singular (rank ≤ 191); the all-position fit has 8,029. Subsampling the all-position fit to the last-token sample count is *worse than either* (KL = 11.49), so estimation quality dominates position; and because the refit also changes $\mathrm{Cov}(h, z)$, the removal direction rotates ($\cos 0.52$–0.67 with the last-token fit across layers 13–24), so the refit does not hold the direction fixed. The two mixed conditions both fall near the all-position KL, which means the factors do not decompose additively either.

LEACE fitted from	**compliance**	$\log p(R)$ **harmful**	$\mathbf{KL}_{\text{next}}$
— (unedited)	−2.882	−0.172	0
last token, all terms ($n = 192$)	−3.978	−0.963	3.783
all positions, all terms ($n = 8029$)	−2.317	−0.244	**0.295**
all positions, subsampled to $n = 192$	−5.512	−8.878	**11.488**
all-position μ, Σ + last-token direction	−1.497	−0.626	0.348
last-token μ, Σ + all-position direction	−2.194	−0.890	0.282

Factor ablation of the position refit, layers 13–24, $\alpha = 1$, 40-prompt sweep, shrinkage $10^{-3}\mathrm{tr}(\Sigma)/d$. The n-matched row is the decisive one: at equal sample count the all-position fit produces far *more* KL, so the effect is not cleanly positional.

So: *fit over all positions* remains correct, and the failure mode it avoids is real. But the mechanism is covariance estimation entangled with a rotating concept direction, not an additive mean mismatch. **Three further models also narrow the scope of that recommendation.**

model (layer)	operator	last-token fit	all-position fit
Qwen2.5-0.5B (13–24)	LEACE	3.783	0.295
OLMo-2-1B (12)	LEACE	5.989	1.532
SmolLM2-1.7B (18)	LEACE	1.102	0.132
Falcon-H1-1.5B (18)	LEACE	0.173	0.064
Falcon-H1-1.5B (11)	LEACE	0.422	**0.046**
SmolLM2-1.7B (18)	ACE	0.987	0.364
Falcon-H1-1.5B (18)	ACE	0.054	**0.066**
Falcon-H1-1.5B (11)	ACE	0.267	0.210

Next-token KL (nats) under each fit. The LEACE benefit appears in every model tested and is largest where the last-token fit produces the most KL (9.1× at Falcon's layer 11). The ACE benefit is not: on Falcon at layer 18 the all-position fit is nominally *worse*, with overlapping intervals ($[0.012, 0.124]$ against $[0.015, 0.149]$).

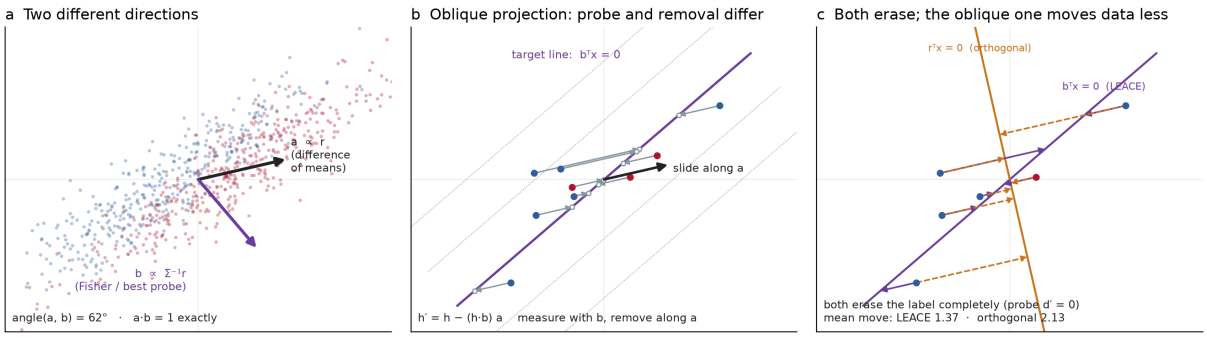
The singularity that motivates the refit reproduces everywhere. On Falcon the last-token fit has $n = 380$ samples in $d = 2048$, giving covariance rank 379, against $n = 9,185$ and full rank for the all-position fit. But a singular covariance only matters to an operator that inverts it. LEACE's probe is $b \propto \Sigma^{-1}\Delta\mu$; ACE's offset is a scalar mean along a fixed direction and never touches Σ. **The recipe is therefore LEACE-specific, not a general property of fitted steering statistics.** The orthogonal projection is not exempt either: it has no fitted offset to mismatch, but its estimated direction and behavioural effect still depend on extraction position.

KL is prompt-set dependent.

The same edit, scored against the same unedited model, yields systematically more KL on XSTest-safe prompts than on Alpaca: 2.4–23.0× across eight Qwen and OLMo cells, 1.20–3.41× on Falcon. It is a tendency rather than a rule, because SmolLM2's $\Delta\mu$ reverses it at 0.57×. The practical consequence is that **a KL figure without its evaluation prompt set is uninterpretable**, and two papers reporting different KL for nominally the same method may differ in nothing else.

What a and b are.

b is the probe (the amount removed is $(h-\mu)\cdot b$) and a the removal direction, with $a \cdot b = 1$, which makes $P = I - ab^\top$ idempotent. Unwinding the whitening, $a \propto \mathrm{Cov}(h, z) \propto \Delta\mu$ and $b \propto \Sigma^{-1}\Delta\mu$ (both verified to 0.0° in a 2-D construction). LEACE measures with the Fisher direction and removes along the difference of means. The two coincide when $\Delta\mu$ is an *eigenvector* of Σ. Full isotropy is unnecessary, and $\Sigma = \mathrm{diag}(1, 4)$ with $\Delta\mu = (1, 0)^\top$ already gives an orthogonal eraser; equality of the full affine maps additionally needs the offset to vanish. At layer 18 the two sit 56.9° apart (Fig. 5). Note that P has rank $d - 1 = 895$; it is the *correction* $ab^\top$ that is rank one, which is where the $\sim 100\times$ speedup over a $d \times d$ matmul comes from.



(a) $a \propto \Delta\mu$ and $b \propto \Sigma^{-1}\Delta\mu$ differ under anisotropy. (b) The oblique projection measures with b, removes along a. (c) Both erase the label; the oblique one displaces less.

Direction against operator: what the contrast can support

Comparing the direction factor upward (weaker → stronger) against the operator factor downward (stronger → weaker) supports the conclusion that the direction is the only factor that improves both objectives at once. Reversing the operator contrasts removes that conclusion:

change	**Δ compliance**	**Δ KL**
$\Delta\mu \to D_1$ (direction)	+1.167	−0.016
ACE → orthogonal (operator)	**+1.413**	**−0.505**
LEACE → orthogonal (operator)	+0.964	−0.160

The same measurements, with the operator contrast taken in the other direction. Operator choice also improves both quantities, and from ACE it improves each one *more* than the direction change does.

These rows do not establish the opposite ranking either, because the protocols differ (§3: single layer-18 vector over all layers for the direction and rank arms, layer-specific vectors over layers 13–24 for the operator arms, and 40- versus 80/60-prompt panels). The defensible local statement is: **none of the tested replacement operators improves on the default orthogonal operator at these settings**. The strength arm is also incomplete, because it starts at $\alpha = 0.25$ and never measures $0 \to 0.25$, and under repeated projection a quarter strength is not obviously weak (a component carried unchanged through 24 projections would be attenuated by $(1 - 0.25)^{24} \approx 0.001$; the real network adds new contributions each layer, so this is an illustration, not a prediction). A small gain from 0.25 to 1 therefore does not show strength is unimportant. A factorial claim needs one protocol, interaction terms, and cost–performance frontiers; we have none of the three.

Direction quality and selectivity

On the compliance-against-KL axis of §10, no ranking of direction against operator is identifiable given the protocol differences. On a third axis the comparison is unconfounded, and gradient-searched directions hold a clear and reproducible advantage: how much harmful-versus-harmless *discrimination* the edit retains.

On the full XSTest suite (250 safe prompts, 150 matched unsafe), discrimination is the gap between teacher-forced refusal likelihood on unsafe and safe prompts. Every edit we tested drives over-refusal to the measurement floor. They differ in what that costs.

	unedited	$\Delta\mu$	D_1	D_2	D_3
OLMo-2-1B (unedited discrimination +0.830 nats, over-refusal 0.076)					
discrimination retained	100%	67%	89%	**117%**	100%
over-refusal rate	0.076	0.004	0.000	0.000	0.000
Falcon-H1-1.5B (unedited discrimination +1.268 nats, over-refusal 0.008)					
discrimination retained	100%	**35%**	80%	**104%**	95%
over-refusal rate	0.008	0.000	0.000	0.000	0.000

All edits drive over-refusal to the measurement floor; only the gradient-searched directions retain the model's harmful-versus-harmless discrimination while doing so. Rates are of 250 safe prompts; 0.000 carries interval [0, 0.015].

While D_2 achieves the same over-refusal reduction at no measurable cost, difference-of-means loses between a third and two-thirds of the model's discrimination. **This is the one place in the study where gradient search yields something the closed-form estimate does not**, and it is an argument about selectivity rather than potency. The edit is better targeted, not stronger.

Two qualifications. Qwen cannot adjudicate: its unedited discrimination is +0.088 nats, so every ratio is a quotient of noise, and indeed its D_1 row is negative. SmolLM2's full-XSTest arm did not run (§14), so this is a two-model result, not four. We report it as the strongest unresolved result in the study rather than an established finding.

Weight folding

Scope. This validates the *all-write single-direction* variant (one layer-18 operator on the embedding and all 48 projection modules at $\alpha = 1$), not the layer-13–24 intervention of Table 6, which is not expressible as an *output-projection* weight edit (the standard convention) because projecting writes in layers 13–24 leaves the state accumulated through layer 12 unprojected. **It therefore does not**

establish deployment equivalence for the behavioural results above. Folded against its matching hook (logit scale ≈ 22): orthogonal $W' = PW$, max logit difference 7.0×10^{-5}; ACE $W' = PW +$ bias, 5.7×10^{-5}; LEACE $W' = PW$ with $P = I - ab^\top$ + bias, 7.5×10^{-5}; residuals $\leq 8.1\times10^{-8}$. Only the orthogonal projection folds as a pure linear map; LEACE's deployed form expands to $Ph + (\mu\cdot b)a$, an affine offset ($\mu\cdot b = 1.0696$ at layer 18, per-module bias norm 0.0369).

Both hook and fold partition that offset as $\frac{1}{48}$ per module, so the offset *ramps*, $\hat{h}_m = PE(x) + \sum_{j \leq m} PW_j u_j + \frac{m}{48} v$, reaching v only after the last write, whereas a residual-stream formulation re-projects each layer and, since $Pv = 0$, clamps the stream at $+v$ throughout. Holding the full offset under folding would require injecting v once at the embedding.

Residual-write coverage is unverified

Abliteration edits every module whose output is added to the residual stream. On Llama-, Qwen2- and Olmo2-shaped models that is two modules per layer (`self_attn.o_proj`, `mlp.down_proj`) plus the embedding. Falcon-H1 writes from **three** (`self_attn.o_proj`, `mamba.out_proj` and `feed_forward.down_proj`), giving 73 sites rather than 49. Discovery code written against the transformer patterns finds **24 of 73**: it matches the attention projection, misses the state-space path entirely, and misses the MLP because the submodule is named `feed_forward` rather than `mlp`.

The verification that should detect this does not. The standard check folds the operator into the weights and confirms the folded model matches the hooked model to numerical precision. Run with 24 of 73 sites edited, it still passes, because the hook and the fold edit the *same* 24 matrices, so they agree exactly while two-thirds of the residual stream goes untouched. The result is an edit covering one-third of the write sites that passes its own verification, reports a plausible reduction in refusal, and measures something other than what is claimed, with no diagnostic that anything is wrong.

The weakness belongs to the check itself rather than to one package's implementation of it. Fold-versus-hook verifies *consistency of application*, never *completeness of coverage*. Any extension of directional editing to an architecture outside the family it was written for needs an explicit coverage assertion. Our harness requires the caller to declare the expected number of residual writes per layer and aborts if discovery disagrees. With all 73 sites edited, folding is exact on Falcon at both layers tested ($5.9\text{–}6.3 \times 10^{-5}$ maximum logit difference at layer 18, $3.1\text{–}4.6 \times 10^{-5}$ at layer 11, relative error $\approx 10^{-6}$), and as in every other model only the orthogonal projection needs no bias term.

Limitations

(1) Proxy outcome. Fixed-string likelihoods, no harmful completion generated or classified, and $\log p(C)$ does not distinguish compliance from safe redirection or irrelevance. **(2) Target coup-**

ling. $D_{1..3}$ were optimised on the scored string with the scoring operator; §10 and §8 give the consequences. **(3) Objective design used the test panel.** The degenerate objective was diagnosed on test prompts and τ, λ chosen in response; validation splits went unused. **(4) Controls.** Two to six random draws per rank cannot establish absence of an effect; equal rank is not equal perturbation magnitude under anisotropy; no weight-matched cone comparison; no seed repeats, so search variance is unmeasured across all four models; ARC differences need paired item-level data. **(5) Scale and coverage.** Four models, three architecture families, 0.5–1.7B parameters. This answers the single-model objection but not the scale one: nothing here is above 1.7B, and our cost model for frontier-scale abliteration is an extrapolation, not a measurement. The layer analysis of §5 was run on one model at two layers; whether other architectures show the same induction/ablation split is untested. **(6) Asymmetric coverage.** SmolLM2's full-XSTest arm never completed, so §11 rests on two models; SmolLM2's XSTest-KL has only the unedited and $\Delta\mu$ conditions; Qwen's over-refusal numbers come from a 100-prompt subset in the original run and the full 250-prompt suite thereafter. **(7) Not run.** Held-out-target control (implemented, never executed), α sweep on OLMo (four of five XSTest conditions are at the measurement floor of 0.000 and mutually indistinguishable), semantic outcome measure, SORRY-Bench's harm×phrasing factorial, OR-Bench, HarmBench/StrongREJECT, MMLU, Siu et al. (2025) or Cristofano (2026) as direction arms, sequential layerwise refitting, cross-dataset validation, fine-tuning repair.

Positioning against the literature

A full-text audit of 15 papers locates most of the individual observations above in prior work. *Direction quality over operator choice*: Cristofano (2026) devote a section to "Weight Editing vs. Vector Quality" and call the raw difference-of-means vector polysemantic; Young (2025) compare four implementations across 16 models. *Rank*: Wang et al. (2026) define steering dimensionality and study it for two refusal types with both induction and suppression; Piras et al. (2025), Winninger (2026) extract multi-dimensional subspaces. *LEACE under-performing on refusal*: Marshall et al. (2024) already report that "more sophisticated erasure techniques such as LEACE … were actively detrimental in many cases". *Random controls*: Wollschläger et al. (2025); Fafuła (2026) use a refusal-free task, a stronger design. *Over-refusal reduction*: Maskey et al. (2026), Wang et al. (2024). *Compliance/phrasing separation*: Zhao et al. (2025). The position-sensitivity failure mode was not located in that corpus, but the corpus was searched with refusal-abliteration terms, and position-dependence of fitted activation statistics belongs to the broader steering and concept-erasure literature; we do not claim it is unreported there. Keyword probing can also miss a result stated in notation or an appendix, so these are "not located", not "novel".

Conclusion

The refusal direction is a reproducible object. Across two transformer families, an untied 16-layer stack and a hybrid Mamba/attention model, the first gradient-searched direction recovers about

half of the difference of means and no more, the difference of means lies mostly outside the span of the directions a search finds, PC1 becomes the refusal direction in the back half of the network, the edit folds into the weights exactly, and random directions of matched rank have no measurable effect. These five quantities are therefore properties of the refusal direction rather than of any one model.

The procedural choices built on it are not. The operator ranking took four different orderings in four models. The share of the harmful–harmless gap covered by style-matched prompts spanned 4 to 80%. "LEACE destroys capability" held in one model of four. Removing more directions improved compliance in one model and reduced it in three. A practitioner who reads any single-model abliteration paper and transfers its ranking to a new model has no basis for expecting it to hold.

The measurement failure modes are worse than the variability, because nothing in the standard protocol flags them. Optimising a direction on the string used to score it inflates the measured effect. Comparing factors from asymmetric baselines reverses their order. Refitting an estimator and attributing the result to one of its terms misidentifies the mechanism. Choosing the edit layer by a causal induction scan selects a layer where the closed-form operators are *worse*, not better. Reporting KL without its prompt set makes the number uninterpretable. And verifying a weight edit by folding it confirms that two implementations agree, never that either one covered the whole residual stream, so an edit covering a third of it passes its own verification.

Our recommendations for anyone starting this work: measure coverage before anything else, select the layer on a cost–effect frontier under the operator you will deploy rather than an induction scan, report a cost column beside every effect column, name the prompt set beside every KL, and do not transfer a ranking across models. The one positive result worth pursuing is selectivity: gradient-searched directions drive over-refusal to the measurement floor while retaining discrimination that difference-of-means loses, on both models where there was enough discrimination to measure.

99 A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda. Refusal in Language Models Is Mediated by a Single Direction. *arXiv:2406.11717*, 2024. N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, S. Biderman. LEACE: Perfect Linear Concept Erasure in Closed Form. *NeurIPS*, 2023. arXiv:2306.03819. P. Clark et al. Think You Have Solved Question Answering? Try ARC. *arXiv:1803.05457*, 2018. T. Marshall, A. Scherlis, N. Belrose. Refusal in LLMs is an Affine Function. *arXiv:2411.09003*, 2024. S. Merity, C. Xiong, J. Bradbury, R. Socher. Pointer Sentinel Mixture Models. *ICLR*, 2017. P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, D. Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in LLMs. *NAACL*, 2024. V. Siu et al. COSMIC: Generalized Refusal Direction Identification in LLM Activations. *arXiv:2506.00085*, 2025. U. Maskey, M. Dras, U. Naseem. Over-Refusal and Representation Subspaces: A Mechanistic Analysis of Task-Conditioned Refusal in Aligned LLMs. *arXiv:2603.27518*, 2026. R. J. Young. Comparative Analysis of LLM Abliteration Methods: A Cross-Architecture Evaluation. *arXiv:2512.13655*, 2025. A. Labunets. Refusal geo-

metry reflects refusal training: diverse refusal prefixes can raise stable rank and weaken refusal vector ablation attacks. *arXiv:2608.25390*, 2026. A. Fafuła. Abliteration Is Not a Scalpel: Off-Target Effects of Refusal Removal on Decision Disposition Across Model Families. *arXiv:2607.17427*, 2026. T. Winninger. Fast Multi-dimensional Refusal Subspaces via RFM-AGOP. *arXiv:2607.02396*, 2026. G. Piras, R. Mura, F. Brau, L. Oneto, F. Roli, and B. Biggio. SOM Directions are Better than One: Multi-Directional Refusal Suppression in Language Models. *arXiv:2511.08379*, 2025. T. Cristofano. Surgical Refusal Ablation: Disentangling Safety from Intelligence via Concept-Guided Spectral Cleaning. *arXiv:2601.08489*, 2026. H. Wang, E. Miehling, D. Wei, K. N. Ramamurthy, and H. Zhang. On the Steering Dimensionality of Refusal in Language Models. *arXiv:2610.04245*, 2026. R. Taori et al. Stanford Alpaca: An Instruction-following LLaMA Model, 2023. X. Wang et al. Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via Single Vector Ablation. *arXiv:2410.03415*, 2024. T. Wollschläger, J. Elstner, S. Geisler, V. Cohen-Addad, S. Günnemann, J. Gasteiger. The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. *ICML*, 2025. arXiv:2502.17420. J. Zhao et al. LLMs Encode Harmfulness and Refusal Separately. *arXiv:2507.11878*, 2025. A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, M. Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043*, 2023.