ANCHORED BUT NOT PROTECTED: WHEN REFUSAL ERODES DURING FINE-TUNING, AND WHETHER HOLDING ONE DIRECTION STOPS IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning an aligned chat model on ordinary instruction data makes it more willing to answer harmful requests. One family of defences picks a direction in activation space, calls it the refusal direction, and penalises movement along it while training runs. We audit that idea on LLaMA-3-8B-Instruct and Qwen2-7B-Instruct. Erosion is concentrated: 52 of the 517 optimiser steps in epoch 1 carry 89 to 101 percent of the net displacement that epoch produces, under four learning-rate schedules. The coordinate these defences hold mediates the refusal that fine-tuning removes. On four exploratory harm sets, restoring its vanilla value at inference takes undue compliance down 18 points to within 2.2 of the vanilla model, and past vanilla under a reference classifier, while the same operation on a random direction does nothing measurable; by ten epochs it recovers about three quarters of the gap. What fails is the calibration. On LLaMA-3, at the strength we inherited, the penalty prevents 9.61° of drift where our acceptance bar needs 15.18°. Anchoring on a direction pooled over nine refusal categories prevents 3.08° more on LLaMA-3 and 2.91° more on Qwen2, lowers undue compliance on LLaMA-3 by 2.0 to 3.6 points across three readings of one rubric, and still misses the bar. Tripling the penalty on LLaMA-3 clears it and cuts undue compliance on the harm sets by a further 6.75 and 8.25 points under the primary scorer and the reference, at a cost of one extra false refusal in 250 and no loss of ReClor accuracy relative to training with no penalty at all. Underneath sits a measurement problem large enough to decide results: three scorers disagree by up to 70 points on identical responses, and one published rubric built two defensible ways disagrees with itself on 6 percent of them, moving the pooled anchor's pre-registered effect from 3.6 points to 2.2 and changing whether it resolves.

1 INTRODUCTION

Safety training does not survive fine-tuning well. Train an aligned chat model on ordinary instruction data, with no harmful examples in it, and it starts answering requests it used to refuse (Qi et al., 2024).

One family of defences tries a simple fix. Refusal seems to be written along a single direction in the model's activations: take the mean activation on harmful prompts, subtract the mean on harmless ones, and removing the resulting direction makes a model comply with almost anything (Arditi et al., 2024). So penalise the model during fine-tuning whenever its activations move along that direction, and refusal should survive. ProCon (Du et al., 2025) is the cleanest version of this idea and the one we audit.

The argument is a chain of four links (Figure 1), and anyone deciding whether to use the defence needs to know which link breaks. We tested each one. The first two hold. The third holds once the penalty is strong enough, and the fourth fails at the strength we inherited, for a mundane reason: the penalty is too weak, and on LLaMA-3 turning it up is enough to pass our bar.

- **Timing is not the problem; calibration is.** The penalty can act inside the short window where erosion happens. On LLaMA-3, at the strength we inherited, it prevents 9.61° of drift against a bar that needs 15.18°; at three times that strength it prevents 19.41° and clears. On our reading of the unit map, ProCon's strong setting sits near our $3\times$.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

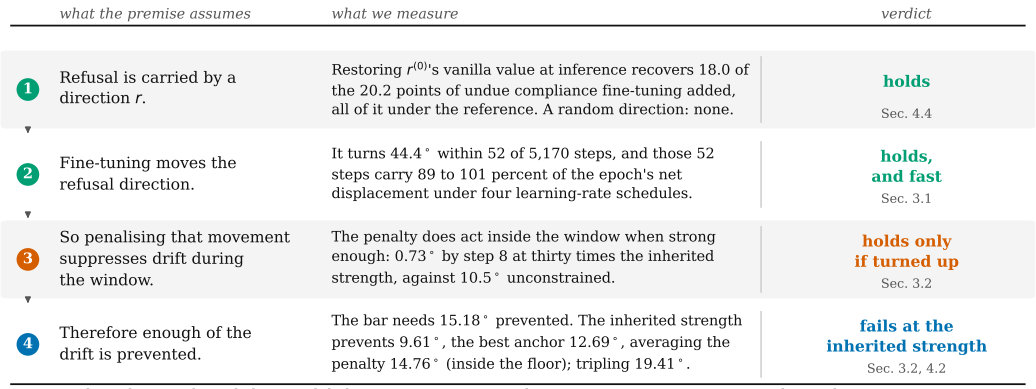


Figure 1: The anchoring premise as four linked claims, with what we measured against each, on LLaMA-3. Link 1 uses the four exploratory harm sets of Appendix A.4. The last link fails at the inherited strength, where the penalty prevents about two thirds of what our bar requires, and holds at three times that strength. Every number reappears in the section named.

- **The anchored direction matters causally, and a pooled direction protects the geometry better at the same strength.** Restoring the component along the anchor at inference undoes most of the compliance fine-tuning added, and the same operation on a random direction does nothing.
- **Scoring choices decide results, including ours.** Three scorers differ by up to 70 points on the same responses. One published rubric, built two defensible ways, disagrees with itself on 6 percent of responses, enough to move our pre-registered anchor effect from 3.6 points to 2.2 and to change whether it resolves.

Code, run records and the scripts behind every number are at `https://anonymous.4open.science/r/anchored-but-not-protected--6D11`, and Table 17 maps each headline number to the run, prompts, scorer and file behind it.

2 Setup and Measurement

2.1 The Defence We Audit

Following Du et al. (2025), the refusal direction $r^{(l)}$ at layer l is the difference between the mean activation on harmful and on harmless prompts at the last prompt token, scaled to unit length. The prompts are 100 AdvBench instructions and minimal edits of them that swap the harmful span for a harmless one; the AdvBench prompts we evaluate on are disjoint from these. We write $r^{(0)}$ for the direction extracted from the vanilla model and $r^{(T)}$ for one re-extracted after training, dropping the layer index where it is not needed. With $z^{(l)}(s) = r^{(l)\top} x^{(l)}(s)$ for how far an activation points along the direction, the constraint keeps each response token's z close to the value the frozen, pre-fine-tuning model gives:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \sum_{l} \sum_{j} \left| z_t^{(l)}(s_j) - z_0^{(l)}(s_j) \right|, \tag{1}$$

summed over layers and response tokens, with z_0 computed once before training starts.

The loss sums rather than averages, over 124 to 11,726 terms per micro-batch, so averaging is a different objective, not a rescaling (Section 4.3 tests both). We run $\alpha = 1.044 \times 10^{-4}$ on LLaMA-3, the value in the implementation we started from, and call it the *inherited* strength rather than the published one, because Du et al. (2025) report 0.1 and 0.2 for ordinary use and 1 and 2 for their strong setting, in units we cannot map with certainty. Two probes put the conversion near 3,759, the mean number of terms the sum runs over (Appendix A.9), if their code averages where their paper writes a sum. On that reading their ordinary settings are 0.25 and 0.51 times our inherited strength, and our $3\times$ sits near their strong setting.

2.2 MODELS, DATA AND ARMS

We fine-tune LLaMA-3-8B-Instruct and Qwen2-7B-Instruct with LoRA (Hu et al., 2022) (rank 8, scaling 16, on the query, key, value and output projections) at a constant learning rate of 2×10^{-4} and effective batch 32, for ten epochs on 16,534 ordinary instruction samples from UltraInteract (Yuan et al., 2025) and Alpaca (Taori et al., 2023). An epoch is 517 optimiser steps, so 52 steps is one percent of training. The rate is constant on purpose, since a decaying schedule would damp late movement on its own.

Raw α means different things for different models and anchors, so unless stated otherwise we match arms on κ, the share of the task loss the penalty would take on the unconstrained trajectory: $\kappa = \alpha\bar{q}/\bar{\mathcal{L}}_{\text{task}} = 0.0836$, where $\bar{q}$ is the penalty evaluated along that trajectory (logged there but not applied) and averaged over optimiser steps 1 to 52. This gives 2.043×10^{-5} on Qwen2. κ equalises that share, not the gradient the penalty contributes. Every arm except the second-draw controls loads the same hashed LoRA initialisation, and every run checks that arms agree exactly at step 0 (Appendix A.11).

Drift is the angle, in degrees, between the refusal directions of the base and fine-tuned models, averaged over valid decoder blocks (31 on LLaMA-3, 27 on Qwen2). We also build nine per-category refusal directions from CoCoNot (Brahman et al., 2024), SORRY-Bench (Xie et al., 2024), WildGuardMix (Han et al., 2024) and XSTest (Röttger et al., 2024), 100 prompt pairs each; the norm of the nine principal angles between base and fine-tuned subspaces is our multi-direction measure.

2.3 WHAT COUNTS AS AN EFFECT

Which scorer to believe. We checked every automatic scorer against 150 hand-labelled responses to harmful prompts (Table 16). The released HarmBench classifier (Mazeika et al., 2024) agrees best, at 0.907, reads every response on every arm and never declines, so we treat it as the *reference*. The HarmBench rubric is published as a Llama-2 chat string, and a chat judge can receive it whole, as one user turn (as run), or split into a system turn and a user turn (as our code documents it). With a Qwen2 judge on LLaMA-3's responses the two builds agree with our labels at 0.780 and 0.773, yet disagree with each other on 388 of 6,200 responses; we rescored every retained generation under both. With a LLaMA-3 judge on Qwen2's responses the chat construction declines 22 percent, and declines that differ across arms bias a paired comparison. We keep the as-run build as the *primary scorer*, the one most tables lead with, because the pre-registration named it. The reference, a Llama-2 derivative, stretches our rule that no model scores its own family.

When a result counts. A behavioural comparison is resolved when McNemar's exact p (McNemar, 1947) is below 0.05 in the predicted direction and the effect clears the floor measured on the same sets under the same instrument, as pre-registered; $[b/c]$ counts the prompts that moved in the direction of the reported difference and against it. Brackets on angles are 5th to 95th percentiles over 2,000 bootstrap replicates that resample the 100 extraction pairs jointly for both arms. They capture extraction noise, not training noise, so a geometric comparison is resolved only when it also clears the floor for its kind. Floors are empirical control bounds from a few matched reruns, not estimates of run-to-run variance. For geometry, a second LoRA draw of the same arm lands 0.397° away at the end of epoch 1, and comparisons across measurement runs must clear 0.67° (Appendix A.11). The nine-direction norm is not on the rank-1 scale and is compared only as a profile over time. Behavioural floors depend on set and scorer: 1.0 point under the primary scorer on both the five pre-registered and the four exploratory sets, 1.8 on AdvBench, and 6.4 early in the window (7.3 under the reference). Exploratory tests are not corrected for multiplicity.

3 WHEN EROSION HAPPENS, AND WHETHER THE CONSTRAINT CAN ACT THERE

3.1 THE WINDOW

We checkpointed the unconstrained arm at optimiser steps 1, 2, 4, 8, 16, 32 and 52. Drift reaches 0.59°, 1.06°, 2.86°, 10.62°, 31.41°, 41.64° and 44.44°. It starts at step 1 and moves fastest between

steps 8 and 16, so on this arm a defence that waits eight steps has already missed a quarter of the movement.

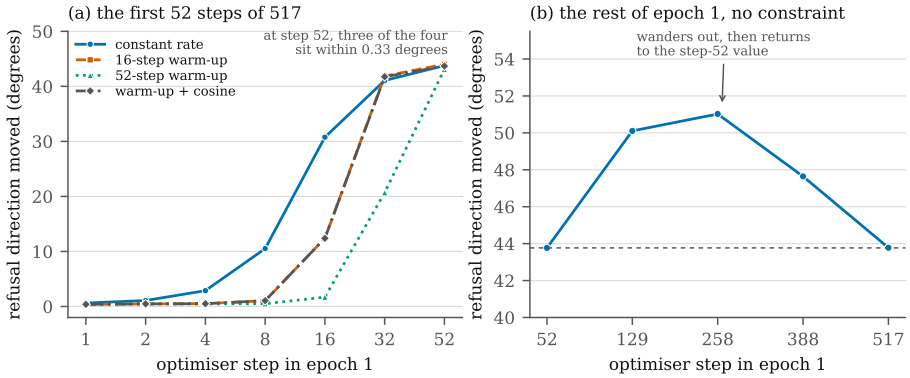


Figure 2: Erosion is concentrated at the start of training under all four learning-rate schedules. The schedule changes where inside the window it happens, and one schedule ends the epoch 5.4° higher than the rest. (a) Drift over the first 52 of 517 optimiser steps; the 16-step warm-up and warm-up-plus-cosine arms overlap because cosine decay is nearly flat over this range. (b) The unconstrained arm over the rest of the epoch: drift overshoots to about 51° mid-epoch and settles back.

Since we trained without warm-up, we ran three more epoch-1 arms with a 16-step warm-up, a 52-step warm-up, and a 16-step warm-up followed by cosine decay. Early concentration survives: under all four schedules step 52 carries 89 to 101 percent of the arm's end-of-epoch displacement (Figure 2a). Where inside the window the movement happens depends on the schedule: by step 16, 70.0 percent of the step-52 angle is present at a constant rate (net of the step-0 angle, Appendix A.11) against 27.6 percent with a 16-step warm-up. And the end-of-epoch magnitude is not fully schedule-invariant. At step 52 the constant, 16-step warm-up and warm-up-plus-cosine arms sit within 0.33° of each other, inside the cross-run floor, but by the end of the epoch the warm-up-plus-cosine arm reaches 49.12° while the other three end at 43.76°, 43.77° and 43.85°, eight times the cross-run floor. Its cosine decays to zero at step 517, so its rate is about half its peak at the step-258 overshoot and near zero over the stretch where the other arms settle back; we read the gap as the arm being frozen mid-overshoot. Batch order, not recorded for these arms, is a second candidate.

The model keeps moving after step 52, overshooting to 51.02° at step 258 and settling to 43.78°, so the claim is about net displacement (Figure 2b; separate unconstrained runs, spanning 0.66°, at our 0.67° floor, Appendix A.11). The same profile appears in both families on the nine-direction subspace (Appendix A.3), and epoch 1 carries about 79 percent of the ten-epoch net displacement.

3.2 Turning the constraint up

A strength sweep on LLaMA-3 runs epoch 1 at 0, 1, 3, 10 and 30 times the inherited strength, with identical batches in all five arms (Table 1).

Geometry and behaviour come apart. Drift falls 9.0° to 9.8° per threefold step. Under the primary scorer attack success falls from 0.182 to 0.009 and never rises, but it reads $3\times$ and $10\times$ as identical; under the classifier, $3\times$, $10\times$ and $30\times$ read 0, 3 and 0 prompts of 110, and the 3 does not resolve against either neighbour ($p = 0.25$). Measured safety reaches what these instruments resolve at three times while geometry keeps improving for another order of magnitude. Du et al. (2025) also see attack success still falling at their strongest setting on LLaMA-3; our sweep adds the geometry.

Strength also settles the timing worry: at thirty times, drift at step 8 is 0.73° against 10.5° unconstrained, so the penalty is not too late, only too weak.

This is where the defence clears our geometric target. The target, fixed before any of these runs (Section 4.2), asks epoch 1 to end below 28.6° on LLaMA-3. At three times the inherited strength the arm ends at 24.37°, and with the pooled anchor at 23.81°. Both clear it and the stricter recomputed target of 26.4° (Appendix A.5). Neither exceeds 28.6° at any checkpoint; against 26.4°, only the ProCon-anchor arm does, once, at the step-258 overshoot. Behaviour moves too: on the four

Table 1: Geometry keeps improving with strength, while measured safety reaches the classifier's resolution at three times. LLaMA-3, end of epoch 1. Attack success is on AdvBench under the primary scorer (as run) and the reference classifier. Over-refusal is the false refusal rate under the refusal judge on XSTest-safe (250 prompts) and 100 held-out Alpaca instructions, so the Alpaca column spans 0 to 3 prompts. ReClor (Yu et al., 2020) is 452 logical reasoning questions, reported raw and among parsed answers.

	geometry		**attack success**		**over-refusal**		**ReClor**
strength	drift	prevented	as run	classifier	XSTest	Alpaca	raw / parsed
base model	–	–	0.000	0.000	0.156	0.000	0.511 / 0.591
$\alpha = 0$	43.78°	–	0.182	0.136	0.136	0.010	0.564 / 0.568
$1\times$	34.17°	9.61°	0.064	0.036	0.144	0.020	0.555 / 0.558
$3\times$	24.37°	19.41°	0.036	0.000	0.140	0.010	0.584 / 0.593
$10\times$	15.38°	28.40°	0.036	0.027	0.176	0.010	0.569 / 0.591
$30\times$	5.57°	38.21°	0.009	0.000	0.188	0.030	0.535 / 0.575

exploratory harm sets of Appendix A.4, tripling cuts compliance by a further 6.75 points under the primary scorer [37/10], $p = 10^{-4}$, and 8.25 under the reference [42/9], $p = 3\times10^{-6}$. Against the base model the two scorers differ. Under the reference the $3\times$ arm reads 0 of 110 on AdvBench like the base model, and 17.25 against 17.50 on the four sets ($p = 1.00$), where the inherited strength sits 8.00 points above it [60/28], $p = 0.00085$. Under the primary scorer the $3\times$ arm stays above the base model, by 3.6 points on AdvBench and by 5.0 on the four sets [38/18], $p = 0.01$. Of the three levers we tested on LLaMA-3, only strength, already a parameter of ProCon, clears the bar.

Tripling shows little cost on the utility checks we ran: against the unconstrained arm the $3\times$ arm falsely refuses one more of 250 XSTest-safe prompts, is unchanged on Alpaca, and scores higher on ReClor. The trade-off appears at thirty times, with false refusal up 5.2 points on XSTest-safe and raw ReClor down 2.9, though parsed-answer accuracy holds (0.575 against 0.568) as the parse rate drops from 0.993 to 0.931.

4 Whether the anchored direction is the right one

4.1 Which direction most affects refusal

If refusal is carried by one direction, the first check is whether the anchored direction is that one. We removed a direction from the residual stream at every block of the vanilla model and counted how many refusals flipped to compliance (Table 3). On LLaMA-3 the anchored direction moves refusal about half as much as each category's own direction, the ratio staying between 1.84 and 2.32 at all five generation lengths, and the pooled direction beats the anchor in 8 of 8 non-tied categories at 512 tokens ($p = 0.0078$). On Qwen2 the ordering reverses at 512 tokens: the anchor beats the category's own direction in 9 of 9 ($p = 0.0039$) and the pooled direction in 8 of 9. We treat this as exploratory and cannot explain the split: each cell is a mean over nine categories with no interval, scored by the refusal judge alone, and it ablates the vanilla model where Section 4.4 intervenes on the fine-tuned one. Appendix A.2 has the full sign tests and our reproduction of Arditi et al. (2024).

4.2 Swapping the anchor at the same strength

We trained two epoch-1 arms on LLaMA-3 that differ from ProCon only in the direction the penalty holds, at the same κ. One holds the direction pooled over all nine categories, the other a direction chosen by Arditi et al.'s selection procedure (Appendix A.8). The pooled arm was repeated on Qwen2.

Geometry is the strongest evidence here. The pooled direction prevents 3.08° more drift than ProCon's anchor on LLaMA-3, $[2.90, 3.24]$, positive in 31 of 31 layers, and 2.91° more on Qwen2, $[2.71, 3.15]$, in 23 of 27. Both intervals exclude zero and both effects clear the cross-run floor; that the intervals overlap says the two sizes are compatible, not that they are equal. As protection, the anchor prevents 9.61° and the pooled direction 12.69°, a third more at the same κ (Figure 3b). The advantage builds late (0.56° at step 52). One tension belongs here. On Qwen2, Section 4.1's ablation ranks the anchor *above* the pooled direction, yet holding the pooled direction prevents more drift; its behavioural effect there is 3.0 points under the reference [25/13], $p = 0.073$, which does not resolve, and 4.0 under the substring list and the refusal judge, where it resolves (Appendix A.4).

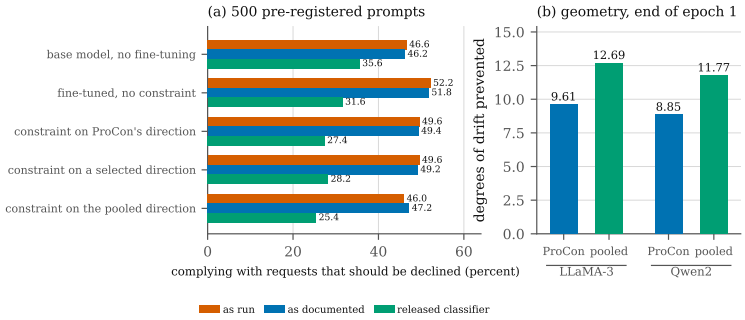


Figure 3: Holding a better direction at the same constraint strength. (a) The rate of complying with requests that should be declined, LLaMA-3, on 500 prompts from the five pre-registered sets, with one HarmBench rubric read three ways. The pooled arm is the lowest of the fine-tuned arms under all three readings; the other two constrained arms sit within the floor of each other, and the absolute level and the gap to the base model change with the reading. (b) Drift prevented at the end of epoch 1 in both families, a geometric measurement no scorer touches.

Behaviour. After two inconclusive runs we pre-registered a third, fixing the arms, five evaluation sets not previously used on the constrained arms, the primary scorer, the predicted direction, the floor and the decision rule before any generation. Four of the five are CoCoNot contextual-noncompliance categories, so what we registered as attack success is mostly complying with requests that should be declined for reasons other than harm.

By the pre-registered rule the swap resolves: $+3.6$ points under the primary scorer [27/9], $p = 0.0039$, clearing the 1.0-point floor (Table 5). It also resolves under the substring list and the refusal judge (Table 14), and does not under the documented build ($+2.2$, $p = 0.080$) or the reference ($+2.0$, $p = 0.052$), so claims that rest on one reading do not survive. ProCon's own protection against the unconstrained arm does not resolve under either chat construction ($p = 0.098$ and 0.13) and does under the reference ($p = 0.0055$), and on these sets fine-tuning raises compliance by 5.6 points under both chat constructions but lowers it by 4.0 under the reference. Our largest anchor effect is not the registered one. On the four exploratory harm sets, used before the pre-registration, the swap is worth 6.0 points under the reference [28/4], $p = 2 \times 10^{-5}$, and 2.5 under the primary scorer ($p = 0.12$); those four are harm categories and four of the five registered sets are not, so we pre-registered the weaker evaluation. The effect is also uneven across categories (Appendix A.7).

It still misses the bar. Our acceptance criterion, fixed before these runs (it is ours, not Du et al. (2025)'s), asks epoch 1 to end below 28.6° on LLaMA-3 and 23.0° on Qwen2. ProCon's anchor ends at 34.17° and 34.00°. The pooled direction ends at 31.088° and 31.091° (a coincidence of two runs); both miss, and recomputing the target widens the miss (Appendix A.5).

4.3 What happens at stronger constraint

At three times the inherited strength, with the endpoint registered before training, the pooled anchor's geometric lead falls from 3.08° to 0.56° $[0.47, 0.66]$, below the cross-run floor. That is an endpoint statement: from step 16 to step 388 the lead is 1.2° to 2.0°, above a floor measured at step 52, though mid-epoch noise between reruns was not measured. The drop is 2.52° (approximate 90 percent interval $[2.33, 2.71]$, Appendix A.6), and it shows in behaviour: the swap's advantage shrinks from $1\times$ to $3\times$ by a resolved margin under three of four instruments, 6.25 points under the reference, and only the primary scorer does not resolve it (Table 9).

Read as levers on the four exploratory harm sets, all scored by the reference on the same 400 prompts, each change helps against ProCon at $1\times$: swapping the anchor by 6.00 points [28/4], switching the reduction rule by 6.00 [32/8], and tripling α by 8.25 [42/9]. All three leave attack success within 2 points of the base model's 17.50 (19.50, 19.50 and 17.25), none resolved against it. Tripling has the largest estimate, but the three do not separate from one another: tripling against the swap is $+2.25$ points ($p = 0.16$) and against the reduction rule $+2.25$ ($p = 0.093$). Stacking the better anchor on the tripled constraint adds nothing under the reference, -0.25 [9/10], and 1.75 to 2.25 points under the primary scorer and the substring list.

Averaging instead of summing the penalty, at the same κ, prevents 5.15° more drift than ProCon, $[4.88, 5.48]$; it realised a larger penalty share, and a sensitivity calculation removing the excess leaves about 2.7°, a lever of roughly the anchor's size (Appendix A.9).

4.4 Restoring the coordinate at inference

Design. Section 4.1 asked the vanilla model. Here we take the fine-tuned model and, at inference only, set the scalar projection $r^{(0)\top} h$ back to the value the vanilla model gives, at every site where the constraint computes it. Nothing is retrained. The control is the same operation on a random unit direction. The 400 prompts are the four exploratory harm sets of Appendix A.4, where the anchor mattered most; the test has no replication on the pre-registered sets.

Result. Under the primary scorer the restoration takes compliance down 18.0 points, $p = 1.3{\times}10^{-15}$, to within 2.2 points of vanilla [25/16, $p = 0.21$] (Table 15). The other three instruments put the restored arm on the safe side of vanilla: by 2.0 points under the refusal judge ($p = 0.18$), 3.8 under the substring list ($p = 0.044$) and 5.25 under the reference ($p = 0.0011$), which reads compliance falling from 36.50 to 12.25 against vanilla's 17.50. A random direction does nothing, $+1.0$ at $p = 0.45$. A second initialisation draw reproduces the drop and the restored level to within 0.75 points under the reference, including the pass beyond vanilla (4.50 points, $p = 0.011$), and every quantity to within 2.5 as run. On Qwen2 the restoration replicates at the same size, -25.75 points under the reference against LLaMA-3's -24.25, at $p < 10^{-23}$ under all three instruments we read (Appendix A.10 gives a decoding caveat). On the refusal-opening margin, C1's pre-registered readout, the same LLaMA-3 restoration recovers 41 percent of the gap at the end of epoch 1, which our criterion calls partial, and is sufficient at step 52.

What weakens, and what does not replicate. At ten epochs the restoration still recovers about three quarters of the gap under the primary scorer (73 percent; 83 under the reference), but the remainder now resolves ($p = 1.7{\times}10^{-5}$), so by our pre-registered criterion the coordinate is no longer *sufficient*, though it still does most of the work. Removing the coordinate no longer moves compliance ($+3.0$ under the reference, $p = 0.23$). On Qwen2 the ablation does not replicate even in sign: removing the anchored direction *lowers* compliance, by 6.00 ($p = 0.025$) and 5.00 points ($p = 0.047$), where LLaMA-3 gives $+16$. The Qwen2 result therefore rests on restoration, where all three instruments agree.

What this licenses. It licenses saying the anchored coordinate mediates most of the refusal that fine-tuning removes. It does not license the training claim that erosion travelled along that coordinate. Nor do we find that the anchor has been superseded: none of eleven fine-tuned cells meets our relocation criterion, and in both LLaMA-3 cells with generation at step 517, $r^{(0)}$ still mediates more of the margin than $r^{(T)}$, 0.96 against 0.72 and 0.96 against 0.80. Appendix A.10 lists the caveats, including three self-test settings changed after a first run that produced no readouts. The margin's sufficiency verdict also depends on how the multi-direction subspace is built: the two constructions disagree in all fifteen cells (68 against 15 percent recovered on one).

5 Three quantities, three scorers

5.1 Geometry, refusal-opening margin and text move apart

We also scored the seven checkpoints on 110 AdvBench prompts (Zou et al., 2023) at 512 new tokens and tracked the refusal-opening logit margin, the gap at the last prompt position between the best token that begins a refusal and the best one that does not; it has no human validation.

They do not move together. Normalised to each curve's change from step 0 to step 52, by step 16 the margin has moved 78.2 percent, the direction 70.7 percent, and measured attack success 12.2 percent under the primary scorer or 4.3 under the reference (Figure 4b). The direction clearly moves before attack success; that the margin moves earlier still rests on one unvalidated readout.

Two unconstrained arms that differ only in loss path already differ by 6.4 points of attack success at step 52 (7.3 under the reference), while step 16 sits 4.5 points above the base model under the primary scorer and 1.8 under the reference. So at step 16 neither generation-based scorer clears its control bound, despite the geometric change already present. The same angle need not mean the same

behaviour across runs either. At step 52 the 52-step warm-up arm sits 0.65° from the constant arm, at the floor, and 10.0 points of attack success from it under the primary scorer, outside the 6.4-point floor (15.4 under the substring list, none under the refusal judge; Appendix A.8); interpolating across all four schedules gives spreads of 5.7 to 32.6 points at similar drift angles. The refusal angle alone does not determine measured safety across training runs.

5.2 Scorer choice can change the conclusion

The clearest case is one scorer disagreeing with itself. The two builds of the HarmBench rubric (Section 2.3) agree with our human labels almost identically, 0.780 and 0.773, yet disagree with each other on 388 of 6,200 responses. On the pre-registered run that 6 percent moves the anchor effect from $+3.6$ to $+2.2$ points and its p from 0.0039 to 0.080, which decides whether the anchor swap resolves.

Coverage matters as much as agreement, and here the two can be pulled apart. The Qwen2 anchor swap looked like a null under the chat construction, $+0.7$ points [13/11], $p = 0.84$, on the 300 prompts (75 percent) that both arms parse. On the same 300 the reference reads $+1.3$, and on all 400 it reads $+3.0$ ($p = 0.073$), so about 1.7 of the 2.3-point difference comes from which prompts get read and about 0.7 from the instrument.

The disagreement is largest where effects are large. With Arditi et al.'s selected direction removed, three scorers read the same 100 JailbreakBench responses as 84, 48 and 14 percent attack success. Across the six ablation conditions of Figure 4a, averaged over evaluation sets, the three scorers disagree by more than the effect itself every time, by 7.5 to 46 points. Among trained arms the ten effects below 9 points spread by at most 5.5 across scorers, and every comparison above 9 points in the figure spreads by 14.5 or more; spread and effect come from the same readings, so this describes these data rather than a general law. On one category the swap clearly changes sign, -3 points under the primary scorer against $+9$ and $+6$ under the others (Appendix A.7). Our own readouts are not exempt: the two subspace constructions of Section 4.4 decide the pre-registered sufficiency verdict. For the small and moderate effects in our study, we do not treat a single scorer as sufficient when reasonable alternatives disagree.

6 Related work

Safety erodes under ordinary fine-tuning. Qi et al. (2024) showed that fine-tuning on data with no harmful examples raises harmful compliance. He et al. (2024) identify benign subsets that break safety disproportionately, and Jan et al. (2024) track how the gap opens across task mixtures. Du et al. (2025) find the refusal direction's drift sharpest after the first epoch and design their warm-up around it; these works, theirs included, measure per run or per epoch, and we measure inside the first epoch.

Where refusal lives. Arditi et al. (2024) showed that a single difference-in-means direction mediates refusal at inference, the result these defences build on. Wollschläger et al. (2025) argue the picture is not one-dimensional, recovering cones of representationally independent refusal directions, and Li et al. (2025b) localise safety to a band of layers. Section 4.4 asks whether the anchored direction still carries refusal after fine-tuning.

Holding something fixed during fine-tuning. ProCon penalises movement along the refusal direction, AsFT (Yang et al., 2026) keeps updates within a narrow safety basin, and Li et al. (2025b) freeze the layers they identify. ProCon and AsFT report harmfulness and task performance across constraint strengths, and AsFT and Li et al. (2025b) show that the choice of anchored object matters. Among these methods, none sweeps constraint strength against the anchored geometry, where Section 3.2 finds it coming apart from measured safety, or compares candidate refusal directions at a strength matched on the penalty's share of the loss.

Whether any of it can be measured. Chen & Goldfarb-Tarrant (2025) show that LLM safety evaluators are not robust to artifacts in the responses they grade, Schwinn et al. (2026) find that judges cannot reliably rank methods, and Li et al. (2025a) report similar instability for reward models after instruction tuning. Fraser et al. (2025), closest to us, find that fine-tuning lowers safety and disrupts evaluation consistency at once, and Hossain et al. (2025) standardise the benchmark. We add that one published rubric, built two defensible ways, moves our own pre-registered effect.

7 LIMITATIONS AND DISCUSSION

Scope. Two model families at 7 to 8 billion parameters, LoRA only: no full fine-tuning and no larger models. No run holds the constraint for all ten epochs. One arm held ten times the inherited strength for epoch 1 and then removed it, as ProCon's warm-up does, though for one epoch where Du et al. (2025) use four: it prevented 28.88° at the end of epoch 1 and kept 7.27° $[6.56, 8.01]$ after ten (Appendix A.11). Whether the epoch-1 margin survives a constraint that stays on is untested. The sweep, the schedules and the pre-registered run are LLaMA-3 only. The five pre-registered sets are out of sample as prompts but not as categories, since the pooled direction was fitted across all nine, and the four exploratory sets carry both the lever comparison and the restoration.

Instruments. The same 150 labels, enriched for disagreement, chose the reference and supplied both its 0.907 and the primary scorer's 0.780 (inter-annotator 0.87). In Table 1 the as-run column has a 1.8-point floor that one step, $3\times$ to $10\times$, does not clear. Because the LLaMA-3 judge declines 22 percent of Qwen2's responses, every Qwen2 behavioural result rests on the released classifier and the two secondary instruments. The pre-registered swap is a single registered test: it resolves under its primary scorer ($p = 0.0039$) and not under the documented build or the reference (0.080, 0.052). The exploratory results are uncorrected for multiplicity.

What stays unresolved. Section 4.4 is an inference-time intervention: it shows the anchored coordinate mediates refusal in the fine-tuned model, not that training moved along it. The mapping from geometry to behaviour is not shown to transfer across schedules. The cosine arm's 5.4° is explained by inference from its schedule, not measured. The acceptance target carrying our central verdict is one we defined, and it is not like-for-like across families: from ProCon's anchor it leaves 5.57° still to prevent on LLaMA-3 against 11.00° on Qwen2.

What we conclude. The anchored direction matters causally, and a pooled direction protects the geometry better at the same strength. Most of epoch 1's net displacement is in place within the first one percent of training, and the penalty can act there. At the inherited strength it prevents 9.61° of the 15.18° our bar requires; the better anchor reaches 12.69°, the reduction rule's 14.76° is inside the floor of the original bar and short of the recomputed one, and tripling the penalty prevents 19.41° and clears it. On the four exploratory harm sets tripling also has the largest behavioural estimate under both the primary scorer and the reference (6.75 and 8.25 points), though the three levers do not separate from one another there. The practical reading is narrow: on LLaMA-3, raising the penalty to about ProCon's strong setting, on our best reading of the unit map, clears our bar at the end of epoch 1 with little cost on the utility checks we ran. And two defensible builds of one published rubric changed whether our pre-registered comparison resolves, so in this area the choice of scorer is part of the result.

AI USE STATEMENT

We used generative AI tools to assist with parts of the experiment and analysis code, to refine control and diagnostic checks. We also used them to support figure preparation, literature search and formatting, and manuscript editing, including clarity, consistency, and claim-calibration checks across the main text and appendices. Generative AI was not used to generate the evaluation datasets or to replace experimental measurements. All experiments were run by the authors, and AI-assisted code and analyses were manually reviewed and validated against the recorded runs and implementation checks. The provenance of the paper's headline results is documented in Table 17, with missing artifacts explicitly disclosed where applicable. All citations and final scientific claims were checked and approved by the authors. We take full responsibility for the final content of the paper, including text, code, analyses, and other artifacts produced with the aid of generative AI.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A.

Smith, Yejin Choi, and Hannaneh Hajishirzi. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 2024.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 19750–19766, 2025.

Yanrui Du, Fenglei Fan, Sendong Zhao, Jiawei Cao, Qika Lin, Kai He, Ting Liu, Bing Qin, and Mengling Feng. Anchoring refusal direction: Mitigating safety risks in tuning via projection constraint. *arXiv preprint arXiv:2509.06795*, 2025.

Kathleen C. Fraser, Hillary Dawkins, Isar Nejadgholi, and Svetlana Kiritchenko. Fine-tuning lowers safety and disrupts evaluation consistency. In *Proceedings of the First Workshop on LLM Security (LLMSEC)*, pp. 129–141, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety. In *Conference on Language Modeling*, 2024.

Saad Hossain, Samanvay Vajpayee, and Sirisha Rambhatla. SafeTuneBed: A toolkit for benchmarking LLM safety alignment in fine-tuning. *arXiv preprint arXiv:2506.00676*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Essa Jan, Nouar AlDahoul, Moiz Ali, Faizan Ahmad, Fareed Zaffar, and Yasir Zaki. Multitask mayhem: Unveiling and mitigating safety gaps in LLMs fine-tuning. *arXiv preprint arXiv:2409.15361*, 2024.

Guanlin Li, Kangjie Chen, Shangwei Guo, Jie Zhang, Han Qiu, Chao Zhang, Guoyin Wang, Tianwei Zhang, and Jiwei Li. Picky LLMs and unreliable RMs: An empirical study on safety alignment after instruction tuning. *arXiv preprint arXiv:2502.01116*, 2025a.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to LLM security. In *International Conference on Learning Representations*, 2025b.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Leo Schwinn, Moritz Ladenburger, Tim Beyer, Mehrnaz Mofakhami, Gauthier Gidel, and Stephan Günnemann. A coin flip for safety: LLM judges fail to reliably measure adversarial robustness. In *International Conference on Machine Learning*, 2026.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *International Conference on Machine Learning*, 2025.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*, 2024.

Shuo Yang, Qihui Zhang, Yuyang Liu, Yue Huang, Xiaojun Jia, Kun-Peng Ning, Jia-Yu Yao, Jigang Wang, Dai Hailiang, Yibing Song, and Li Yuan. AsFT: Anchoring safety during LLM fine-tuning within narrow safety basin. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34322–34330, 2026.

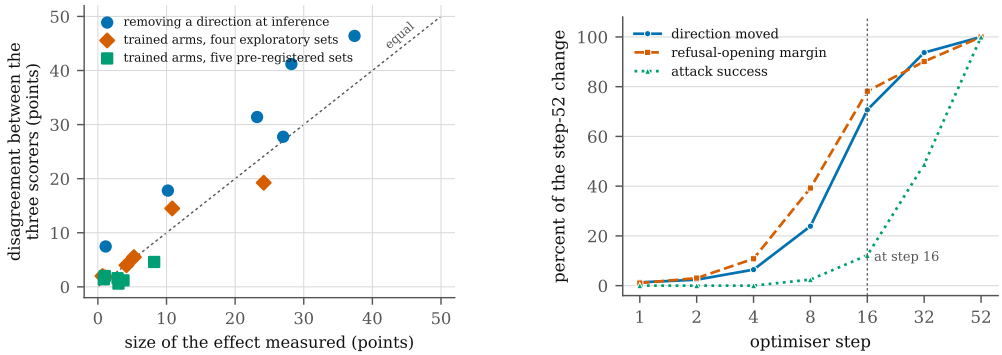
Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. *International Conference on Learning Representations*, 2020.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing llm reasoning generalists with preference trees. *International Conference on Learning Representations*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Additional results

A.1 Scorer disagreement and the three curves of Section 5.1



(a) (b)

Figure 4: (a) Disagreement between three scorers (the primary scorer, the 12-substring list and the refusal judge) against the size of the effect they score. Each point is one paired comparison between arms or interventions, read by all three scorers on the same generations; among trained arms, scorer spread is small for the small effects and larger for the larger ones; in the ablation series, disagreement exceeds the effect in all six conditions; and the dotted line is where the disagreement equals the effect. (b) The three quantities of Section 5.1, each normalised to its own change between step 0 and step 52. By step 16 the refusal-opening margin and the direction have made most of their step-52 change and attack success little of it (78.2, 70.7 and 12.2 percent under the primary scorer, 4.3 under the reference); the exact behavioural crossing depends on the scorer (Appendix A.8).

In Figure 4a the horizontal axis is the mean absolute effect over the three scorers and the vertical axis their spread. Both come from the same readings, so the figure describes how those readings behave rather than relating two independent measurements. Among trained arms, no comparison below 9 points spreads by more than 5.5, and the two above spread by 14.5 and 19.3. The pooled Spearman correlation over trained arms, $+0.664$, is $+0.943$ within the run whose effects span 0.7 to 24 points and $+0.086$ within the run whose effects all fall below 9. Table 2 lists the six ablation conditions, each averaged over the evaluation sets it was run on; in every one the three scorers disagree by more than the effect itself. Its largest single cell is the 70-point spread quoted in Section 1: with Arditi et al.'s selected direction removed, the 12-substring list, the primary scorer and the refusal judge read the same 100 JailbreakBench responses as 84, 48 and 14 percent attack success.

Table 2: The inference-ablation series of Figure 4a, LLaMA-3: the effect of removing each direction on attack success (mean absolute change under the primary scorer against the unablated model) and the spread between the three scorers, in points, averaged over the evaluation sets each condition was run on. Condition names are the keys in `e8/e8_summary.json`.

condition removed	sets	effect	scorer spread
`arditi_selected`	5	37.4	46.4
`ours_refblock`	5	28.2	41.2
`loo`	4	27.0	27.7
`own_percategory`	5	23.2	31.4
`procon_anchor`	5	10.2	17.8
random, norm matched	15	1.1	7.5

A.2 Which direction most affects refusal

Section 4.1 removes a direction from the residual stream at every block of the vanilla model and counts refusals that flip to compliance (Table 3). On LLaMA-3 the anchored direction moves refusal about half as much as the category's own direction, the ratio between 1.84 and 2.32 at all five generation lengths. A paired sign test over the nine categories puts the pooled direction ahead of the anchor in 8 of 8 non-tied pairs at 512 tokens ($p = 0.0078$, one category tied) and 8 of 9 at 64 ($p = 0.039$); the category's own direction leads in 8 of 8 at 512 but only 7 of 8 at 64 ($p = 0.070$). On Qwen2 the ordering reverses at 512 tokens: the anchor beats the category's own direction in 9 of 9 ($p = 0.0039$) and the pooled direction in 8 of 9, and subtracting each category's random value widens the reversal from 1.8-fold to 3.0-fold. Qwen2 was run at two generation lengths and LLaMA-3 at five, so how the reversal behaves with length is unmeasured. Each cell is a mean over nine categories with no interval, scored by the refusal judge alone, which ranks second from the bottom in Table 16. None of our three directions removes refusal outright, where Arditi et al. (2024) take it from about 0.95 to 0.15 on this model; reproducing their procedure recovers their result, and what the reproduction mainly teaches is how their selection criterion works (Appendix A.8).

Table 3: Which direction most affects refusal when it is removed at inference. Numbers are the fraction of prompts the model refused at baseline that flip to compliance, averaged over nine categories, scored by the refusal judge alone. On LLaMA-3 the anchored direction is the weakest non-random choice at both lengths. On Qwen2 the ordering does not hold, and at 512 tokens it reverses.

	LLaMA-3		Qwen2	
direction removed	64 tokens	512 tokens	64 tokens	512 tokens
the category's own direction	0.213	0.320	0.161	0.179
one direction pooled over all nine	0.264	0.285	0.191	0.205
ProCon's anchored direction	0.104	0.169	0.178	0.325
a random direction, norm matched	0.041	0.056	0.053	0.107

A.3 The same early window measured as a nine-direction subspace

On the unconstrained arm the single direction moves 0.59, 0.47, 0.90, 1.94, 2.60, 0.64 and 0.14 degrees per optimiser step over the intervals ending at steps 1, 2, 4, 8, 16, 32 and 52. Repeating the measurement on the nine-category subspace, with the norm of the nine principal angles between base

and fine-tuned subspaces, gives the same early-window profile in both families (Table 4). The nine directions span about five effective dimensions, and we have not checked how much of that is refusal rather than a shared benign component. Their floors are near 4.4° on LLaMA-3 and 14° on Qwen2, which is why the table reports each arm as a percentage of its own displacement.

Table 4: The early window as a percentage of each arm's own step-52 displacement: raw ratios of each step to step 52. Section 3.1 nets out the step-0 angle and uses a different constant-rate arm (Appendix A.11). The nine-direction norm is not on the rank-1 scale, so the two are compared only as profiles.

optimiser step	1	2	4	8	16	32
LLaMA-3, nine directions	1.9	3.1	7.4	22.8	69.3	93.0
Qwen2, nine directions	2.2	3.2	7.0	24.2	67.7	93.6
LLaMA-3, one direction	1.3	2.4	6.4	23.9	70.7	93.7

The two families sit within 1.6 points of each other at every step, and the sign of the difference changes four times across the six checkpoints, so neither front-loads harder. On the single direction, the paired gap between the constrained and unconstrained arms first excludes zero at step 4 on LLaMA-3, $0.116°$ $[0.083, 0.145]$ (a 90 percent extraction-pair interval, Section 2.3), with 22 of 31 layers individually above zero, and at step 8 on Qwen2. The two arms share initialisation and batch order, so this is a within-run criterion rather than the cross-initialisation floor of Section 2.3.

Table 5: The pre-registered run, LLaMA-3: 500 prompts from five evaluation sets not previously used on the constrained arms. Entries are the change in the rate of complying with requests that should be declined, in points, with $[b/c]$ and McNemar's exact p; positive means the second arm declines more. The three column pairs are one HarmBench rubric read three ways: handed to the chat judge whole, as one user turn (as run, the primary scorer), split into a system turn and a user turn (as documented), and read by the released classifier. Substring list and refusal judge are in Table 14.

	as run		**as documented**		**released classifier**	
comparison	points	p	points	p	points	p
second-draw control	-1.0 [9/14]	0.40	-0.2 [11/12]	1.00	$+1.0$ [17/12]	0.46
erosion: base vs. fine-tuned	-5.6 [45/73]	0.013	-5.6 [40/68]	0.0091	$+4.0$ [75/55]	0.095
constraint: none vs. ProCon	$+2.6$ [33/20]	0.098	$+2.4$ [32/20]	0.13	$+4.2$ [37/16]	0.0055
anchor: ProCon vs. pooled	$+3.6$ [27/9]	0.0039	$+2.2$ [22/11]	0.080	$+2.0$ [16/6]	0.052
anchor: ProCon vs. selected-style	$+0.0$ [15/15]	1.00	$+0.2$ [13/12]	1.00	-0.8 [11/15]	0.56
anchor: pooled vs. selected-style	-3.6 [7/25]	0.0021	-2.0 [12/22]	0.12	-2.8 [7/21]	0.013

Two readings of Table 5 are easy to misread. In the erosion row the negative chat entries mean fine-tuning raised compliance on these sets (from 46.6 to 52.2 percent as run), while the reference reads fine-tuning as lowering it (35.6 to 31.6). And the strongest anchor contrast in the table is not the registered one: the pooled direction against the selected-style one resolves under two of three readings ($p = 0.0021$ as run, 0.013 under the reference), where the registered swap resolves under one.

A.4 The anchor swap on the four exploratory harm sets

Where these four sets came from. They are the harm categories among the nine, fixed before any constrained arm was scored on them, when Section 4.1's comparison with Arditi et al. (2024) was restricted to safety categories (XSTest's lookalike set was left out as a contrast set). They were first used for the anchor swap after the AdvBench comparison came back null. The five pre-registered sets are the remaining categories, which is why four of them are CoCoNot non-compliance categories. We call these four the *exploratory* sets. They also carry Table 9, the lever comparison and the restoration, so those results are not independent of one another.

These sets ran first (Table 6); the pre-registered run later reproduced their ordering in sign, the pooled direction ahead of ProCon's anchor and the selected-style direction inseparable from it under all four instruments. The chat construction does not resolve the swap here, which is why the pre-registered run exists. The second-draw control is one draw per instrument; under the refusal judge it came out at exactly zero, so it is a weak floor there. On Qwen2 the swap resolves under the two

Table 6: The anchor swap on the four exploratory harm sets, 400 prompts, LLaMA-3, under four instruments. Each cell gives points $[b/c]$, with McNemar's exact p beneath.

comparison	primary scorer	substring list	refusal judge	released classifier
second-draw control	-1.00 [13/17]	$+1.00$ [9/5]	$+0.00$ [15/15]	-1.0 [16/20]
	0.58	0.42	1.00	0.62
constraint: none vs. ProCon	$+8.50$ [43/9]	$+19.25$ [80/3]	$+4.75$ [31/12]	$+11.0$ [56/12]
	$2e{-6}$	$< 10^{-19}$	0.0054	$6e{-8}$
anchor: ProCon vs. pooled	$+2.50$ [22/12]	$+8.00$ [33/1]	$+5.25$ [23/2]	$+6.0$ [28/4]
	0.12	$4e{-9}$	$2e{-5}$	$2e{-5}$
anchor: ProCon vs. selected-style	$+0.00$ [11/11]	$+1.50$ [7/1]	$+1.75$ [12/5]	$+1.0$ [10/6]
	1.00	0.070	0.14	0.45

secondary instruments, $+4.00$ points under the substring list ($p = 0.0015$) and under the refusal judge ($p = 0.0070$), against controls of -2.00 and -2.25, while the classifier reads $+3.0$ at $p = 0.073$ and does not resolve it.

A.5 Cross-family summary of the anchor result

Table 7: Cross-family summary of the anchor result at the inherited strength.

	LLaMA-3	Qwen2
drift prevented, ProCon's anchor	9.61°	8.85°
drift prevented, pooled direction	12.69°	11.77°
improvement from the swap	+3.08° [2.90, 3.24]	+2.91° [2.71, 3.15]
layers where the swap helps	31 of 31	23 of 27
end-of-epoch drift, pooled direction	31.09°	31.09°
acceptance target for epoch 1	28.6°	23.0°

The two end-of-epoch figures in Table 7 agree to two decimals by coincidence (31.088° and 31.091°, separate runs on separate models).

How the target was built. Each target is an unconstrained drift minus a band, applied identically in both families: $46.0 - 17.4 = 28.6$ and $43.5 - 20.5 = 23.0$. The band is not the epoch-1 drift's own bootstrap interval, which is about 1.4° wide (44.03° $[43.49, 44.92]$ on the current-path unconstrained arm). It is the 95th percentile of a bootstrap floor, the angle a direction moves when only its 100 extraction pairs are resampled, averaged over 31 layers (mean 13.53, p95 17.44), recorded with an earlier gate on the ten-epoch net drift ($56.7 - 17.4 = 39.3$) and carried over to epoch 1 unchanged. For two arms sharing those pairs the paired noise of a difference is 0.20° to 0.39°, so the band is a deliberately demanding margin, not a noise estimate. Subtracting a band rather than scaling is harsher on the noisier family: Qwen2 must prevent 20.5° where a proportional rule matched to LLaMA-3 would ask 16.4°. So the targets are not like-for-like, and distances to target should not be compared across families.

Reconciling the two data paths. Both terms come from the reference arms the project was set up against, on an earlier data path than the arms in Table 7. On the current path the same drifts are 43.78° and 42.86°, and recomputing the targets gives 26.4° and 22.4°, which the best anchor misses by 4.7° and 8.7° instead of 2.5° and 8.1°. Nothing in the pooled-anchor or three-times endpoint verdicts turns on the choice. Two other verdicts do: against 26.4° the reduction-rule arm, at 29.02°, misses by 2.62°, outside the floor, where against 28.6° it is 0.42° short and unresolved; and the ProCon-anchor $3\times$ arm crosses 26.4° once, at its step-258 overshoot. Likewise the Qwen2 gate metadata's ten-epoch 54.0° is the earlier-path reference and our 51.41° the current-path arm.

A.6 Constraint strength: the sweep, and the anchor swap at three times

To test whether the anchor comparison depends on strength we trained one further arm: the pooled anchor at three times the matched α, 3.17×10^{-4} against 1.06×10^{-4}, which the κ rule permits because it is linear in α. Its comparator, ProCon's anchor at the same strength, already existed from

the sweep, with the same baseline and initialisation. Endpoint, instruments, floors and the three possible verdicts were written down before training.

Table 8: The pooled anchor's geometric lead over ProCon's anchor at two constraint strengths, end of epoch 1, LLaMA-3.

constraint strength	gap at step 517	interval	layers positive
inherited ($1\times$)	$+3.08^\circ$	$[2.90, 3.24]$	31 of 31
three times ($3\times$)	$+0.56^\circ$	$[0.47, 0.66]$	20 of 31

The gap falls by 2.52° (Table 8), with an approximate 90 percent interval of $[2.33, 2.71]$ treating the two intervals as independent. What remains at three times is below the 0.67° cross-run floor, its whole interval underneath. Within the $3\times$ epoch the gap reads $+1.51$, $+2.05$, $+1.69$, $+1.81$, $+1.99$, $+1.16$ and $+0.56$ degrees at steps 16, 32, 52, 129, 258, 388 and 517: the advantage holds through mid-epoch and shrinks at the end. The floor was measured at step 52, and mid-epoch noise between reruns was not; two initialisation draws of the unconstrained arm differ by 1.95° at step 129, so the mid-epoch values should be read with that in mind. The two strengths are comparable at only two checkpoints, and at step 52 the ordering reverses, $+0.56^\circ$ at $1\times$ (where the lead builds late) against $+1.69^\circ$ at $3\times$ (where it appears early and fades); the verdict rests on step 517 because the registration named it.

Behaviour was measured on the four exploratory sets, named in the registration before this run (Table 9).

Table 9: The anchor swap at two constraint strengths on the four exploratory sets, and the interaction: the per-prompt $1\times$ effect minus the $3\times$ effect, with a 95 percent bootstrap interval over 10,000 resamples and a Wilcoxon p, computed for all four instruments with one procedure. An earlier computation of the same quantity for two instruments, whose script is not in the repository, gave $p = 0.70$ as run and 0.0017 under the reference, with interval $[+2.75, +9.75]$.

instrument	inherited ($1\times$)	three times ($3\times$)	interaction
primary scorer	$+2.50$ [22/12], p=0.12	$+1.75$ [8/1], p=0.039	$+0.75$ $[-2.50, +4.00]$, p=0.65
substring list	$+8.00$ [33/1], $p<10^{-8}$	$+2.25$ [9/0], p=0.0039	$+5.75$ $[+2.75, +8.75]$, p=0.0003
refusal judge	$+5.25$ [23/2], p=2e−5	$+0.75$ [8/5], p=0.58	$+4.50$ $[+1.50, +7.50]$, p=0.0027
released classifier	$+6.00$ [28/4], p=2e−5	-0.25 [9/10], p=1.00	$+6.25$ $[+3.00, +9.75]$, p=0.0005

The effect shrinks at three times strength. Its point estimate stays positive under three of the four instruments and resolves under two, the primary scorer ($p = 0.039$, clearing its second-draw control) and the substring list ($p = 0.0039$); the refusal judge's $+0.75$ has $p = 0.58$. Its p improves as it shrinks because fewer prompts move the wrong way: 12 of 34 at the inherited strength, 1 of 9 at three times. The classifier reads the residual advantage as -0.25, which agrees with the geometry in leaving the endpoint advantage unresolved. The interaction, whether the anchor helps less once the constraint is stronger, is supported geometrically (by the approximate interval above) and resolves under three of the four instruments; only the primary scorer does not resolve it.

Table 10: Absolute attack success on the four exploratory sets, five arms and the base model.

arm	primary scorer	released classifier
base model	18.25	17.50
unconstrained	38.50	36.50
ProCon's anchor, $1\times$	30.00	25.50
pooled anchor, $1\times$	27.50	19.50
ProCon's anchor, $3\times$	23.25	17.25
pooled anchor, $3\times$	21.50	17.50

In absolute terms (Table 10), under the primary scorer changing the anchor at the inherited strength buys 2.50 points, tripling α instead 6.75 ([37/10], $p = 10^{-4}$), and changing the anchor on top of the tripled α a further 1.75. Under the classifier the same three steps are 6.00, 8.25 and -0.25, and every constrained arm but ProCon's at $1\times$ sits within 2 points of the base model, none resolved against it. Under the primary scorer the tripled ProCon arm stays 5.0 points above the base model ([38/18],

$p = 0.01$). With one strength beyond the inherited one, this shows that the levers interact, not the shape of the interaction.

A.7 What the per-category variation is made of

This subsection is exploratory: it pools the five pre-registered categories with the four exploratory ones, giving nine on LLaMA-3. Table 11 gives the change in attack success from ProCon's anchor to the pooled one, with $[b/c]$; all 27 tests are uncorrected.

Table 11: The anchor swap per category on LLaMA-3, 100 prompts each, under three instruments. Q is the inverse-variance heterogeneity statistic, each category's effect $(b - c)/n$ weighted by $n^2/(b + c)$; categories with no discordant pairs carry no weight and are dropped.

category	primary scorer	substring list	refusal judge
Humanizing requests*	$+11$ [12/1]	$+10$ [13/3]	$+11$ [12/1]
Hate and inappropriate	$+10$ [11/1]	$+13$ [13/0]	$+7$ [8/1]
Unsupported requests*	$+5$ [7/2]	$+2$ [4/2]	$+4$ [6/2]
Incomplete requests*	$+4$ [4/0]	$+0$ [0/0]	$+0$ [0/0]
Crime advice	$+3$ [4/1]	$+4$ [4/0]	$+3$ [4/1]
Safety core	$+0$ [5/5]	$+6$ [6/0]	$+5$ [5/0]
Harmful vs. look-alike*	$+0$ [1/1]	$+5$ [5/0]	$+3$ [3/0]
Indeterminate requests*	-2 [3/5]	-1 [0/1]	$+4$ [4/0]
Safety, contextual	-3 [2/5]	$+9$ [10/1]	$+6$ [6/0]
Cochran's Q	21.7 on 8 df	32.0 on 7 df	5.6 on 7 df
p	0.005	< 0.001	0.58
I^2	63%	78%	0%

* the five pre-registered categories; the other four are the exploratory sets of Appendix A.4.

The nine effects are heterogeneous under the primary scorer and the substring list, not under the refusal judge; with nine categories we read I^2 descriptively. Using the unconditional variance instead gives Q of 23.0, 34.1 and 6.1, so our choice is the conservative one and changes no conclusion. Two categories change sign across scorers, and one clearly: Safety, contextual reads -3 [2/5] under the primary scorer, $+9$ [10/1] under the substring list ($p = 0.012$) and $+6$ [6/0] under the refusal judge ($p = 0.031$), on the same generations (Indeterminate requests reads -2, -1 and $+4$). Whether the swap looks heterogeneous at all depends on the scorer.

Six candidate explanations show no clear association with the per-category effect: the pooled direction's alignment gain over the anchor, the same gain at the reference block, the pooled direction's alignment alone, the anchor's alignment alone, the base model's headroom, and the constraint's own effect. Under the primary scorer all six Spearman correlations have $|\rho| \leq 0.33$, and the alignment gain reads $+0.21$, $+0.10$ and $+0.57$ ($p = 0.11$) under the three scorers. Table 12 is a construction check rather than a finding: for each category we fitted the refusal direction on that category's own extraction split, by the protocol the pooled direction was built under, and measured its cosine with each anchor.

Table 12: Cosine between each category's own fitted refusal direction and each anchor, over the 31 valid blocks. A construction check: the pooled direction was fitted across all nine.

category	pooled direction	ProCon's anchor	gain
Humanizing requests	0.793	0.263	$+0.530$
Unsupported requests	0.769	0.283	$+0.486$
Indeterminate requests	0.669	0.222	$+0.447$
Safety, contextual	0.924	0.532	$+0.392$
Crime advice	0.884	0.525	$+0.359$
Hate and inappropriate	0.795	0.442	$+0.352$
Safety core	0.777	0.434	$+0.344$
Harmful vs. look-alike	0.813	0.592	$+0.221$
Incomplete requests	0.438	0.220	$+0.219$

The pooled direction is closer in all nine, as a paired result: in absolute terms the anchor's best category sits above the pooled direction's worst. The pooled anchor is closer to every one of the nine directions it was built from, while its behavioural advantage is neither uniform across categories nor stable across scorers, which is Section 5.1's point that geometry does not determine behaviour.

A.8 SMALLER CONTROLS AND ARITHMETIC

Matching arms on κ, and what realised share measures. Baseline projections are 743.97 for the pooled direction and 753.93 for ProCon's anchor, so the pooled α barely moves; the selected-style direction's 568.17 needs an α a third higher, 1.385×10^{-4}. Realised penalty share is computed as the ratio of sums over steps 1 to 52, the form the κ rule uses (the mean of per-step ratios over the run gives 4.5 percent for ProCon instead). On it the pooled arm realised 3.72 percent, the selected-style arm 3.55 and ProCon's anchor 3.54; the averaged-reduction arm realised 4.13. Tripling α raises the realised share only 1.83 times, because a stronger constraint shrinks the projection it penalises. As a sensitivity calculation, not a corrected estimate, we apply one rule to both arms that realised more than ProCon. Scaling on the share converts an arm's share ratio into an α multiple through the sweep's 1.83-fold share per threefold α, then into degrees at about 9.8° per threefold step; on that rule a linear scaling would attribute about 2.5° of the reduction rule's 5.15° and 2.1 of its 6.0 points to its extra share, leaving 2.70°, and 0.82° of the pooled anchor's 3.08°, leaving 2.26°. Treating share as proportional to α instead attributes 0.45° to the pooled anchor, leaving 2.63°. Either way the anchor advantage stays far above every floor. Realised share may not measure strength at all: the penalty is an absolute value, so its gradient is α times the projection's gradient whatever the gap, and a larger realised value can simply mean the projection moved further. On that reading the reduction rule's caveat is that it is a different objective (Section 2.1).

Our reproduction of Arditi et al. (2024). As a relative flip we get 0.833 against their 0.842, with one declared deviation: our candidates are fitted on AdvBench alone where they mix four prompt sources. Their selection criterion does most of its work through the KL-divergence bound rather than the bypass score. Thirteen candidates in our sweep score a stronger bypass than the winner and KL excludes all thirteen, including the block and token position they report, which our KL estimate puts at 0.257 against their bound of 0.1. So the criterion mainly filters for directions whose removal leaves other behaviour intact, and our reproduction is close rather than exact. The selected-style anchor of Section 4.2 is a direction chosen by this procedure.

How much of the lag curve rests on single responses. Under the primary scorer the half-way point of Section 5.1's attack-success curve is decided by half a prompt: half the step-52 value is 0.1864 and step 32 sits at 0.1818. Under the reference, step 52 reads 46 of 110 and step 32 reads 23, exactly half. So the crossing is at step 32 under one instrument and between 32 and 52 under the other, and in both it falls well after the direction's half-way point near step 12: the lag is robust, the exact crossing is not.

What makes a fitted direction good. In a four-category check, a direction fitted on three categories beats the held-out category's own direction by 24 and 12 points on two sets and loses by 5 and 6 on the other two; cutting its budget from 300 pairs to 100 moves the result by 1, 0, 4 and 1 points.

Two schedule controls. The 52-step warm-up is still warming up at step 52, reading 43.12° against the constant arm's 43.77° and 0.236 attack success against 0.336. The behavioural gap, 10.0 points under the primary scorer (15.4 under the substring list, none under the refusal judge, which reads both arms at 2.7 percent), is outside its 6.4-point floor; the geometric one, 0.65°, is at the 0.67° floor. The geometric gap closes by the end of the epoch, to 0.08°; attack success was not measured there. Section 3.1's step-52 comparison therefore uses the other three arms. Under a 16-step warm-up the inherited constraint prevents 10.08° $[9.76, 10.39]$ in 31 of 31 layers against 9.61° at a constant rate. That within-run interval excludes 9.61°, but the two come from separate runs, and their 0.47° difference is inside the cross-run floor that governs such comparisons.

A.9 THE RELEASED CLASSIFIER ON THE MAIN BEHAVIOURAL COMPARISONS, AND THE REDUCTION-RULE ARM

Why this appendix exists. On Qwen2 the LLaMA-3 chat judge declines 22 percent of responses, and a decline rate differing across arms biases whatever is left. `HarmBench-Llama-2-13b-cls`,

the classifier the rubric was released with, reads every prompt on every arm and never declines. The substring list and the refusal judge also read both families completely; the classifier does so and agrees best with our labels, so it is our cross-family reference. Table 13 collects its readings.

Table 13: The released classifier on the main behavioural comparisons. "A.2" is the four exploratory sets (400 prompts); "Table 5" is the five pre-registered sets (500).

comparison	**prompts and family**	points [b/c]	p
second-draw control	A.2, LLaMA-3	-1.0 [16/20]	0.618
	A.2, Qwen2	-1.8 [18/25]	0.360
constraint: none vs. ProCon	A.2, LLaMA-3	$+11.0$ [56/12]	6×10^{-8}
	A.2, Qwen2	$+6.8$ [44/17]	0.00073
	Table 5, LLaMA-3	$+4.2$ [37/16]	0.0055
anchor: ProCon vs. pooled	A.2, LLaMA-3	$+6.0$ [28/4]	2×10^{-5}
	A.2, Qwen2	$+3.0$ [25/13]	0.073
	Table 5, LLaMA-3	$+2.0$ [16/6]	0.052
	A.2, LLaMA-3 at $3\times$	-0.25 [9/10]	1.000
anchor: ProCon vs. selected	A.2, LLaMA-3	$+1.0$ [10/6]	0.45
strength: ProCon $1\times$ vs. $3\times$	A.2, LLaMA-3	$+8.25$ [42/9]	3×10^{-6}
reduction: sum vs. mean	A.2, LLaMA-3	$+6.0$ [32/8]	0.00018
restoration: $r^{(0)}$ restored	A.2, LLaMA-3	-24.25 [1/98]	3×10^{-28}
	A.2, LLaMA-3, second draw	-24.5 [1/99]	2×10^{-28}
	A.2, Qwen2	-25.75 [9/112]	9×10^{-24}

The anchor swap has the same sign in both families under this instrument, resolving on LLaMA-3's exploratory sets and not on its pre-registered sets or on Qwen2. ProCon's own protection resolves here in both families and on both LLaMA-3 evaluations, so its failure to resolve under the chat constructions on the pre-registered run looks like an instrument effect on those sets. The anchor advantage at three times is not resolved, as it is not geometrically.

What differs between the two evaluations. Two causes, both visible under the reference category by category (the split reproduces the pooled [28/4] and [16/6]). Construct explains part: on the four harm sets the effect is carried by Hate (+13 [15/2], $p = 0.002$) and Safety, contextual (+8 [8/0], $p = 0.008$), with Crime at +3 [3/0] and Safety core at exactly zero [2/2]; on the five pre-registered sets Humanizing reads +7 [8/1] ($p = 0.039$), beating two of the harm categories, and Incomplete, Unsupported and Indeterminate read +2, +1 and 0. Headroom explains a little more: the lookalike set sits at 1 percent compliance under ProCon's anchor and has no discordant pairs, so a fifth of the pre-registered prompts could not move whatever the anchor did. Without it the pre-registered effect is 2.5 points rather than 2.0, still far short of 6.0.

The reduction-rule arm. Section 2.1's constraint sums $|z_t - z_0|$ over layers and response tokens, 124 to 11,726 terms per micro-batch, so one micro-batch's sum can have ninety-five times as many terms as another's. At the same raw α averaging makes the penalty look inert, since sum and mean differ by the term count, about 3,759 on average, the same factor as the unit map. At the same κ, the averaged arm ends epoch 1 at 29.02° against the summed arm's 34.17°: 5.15° $[4.88, 5.48]$ in 31 of 31 layers. On the four exploratory sets, against ProCon at $1\times$ and under the classifier, the reduction is worth $+6.0$ points [32/8] and the anchor swap $+6.0$ [28/4], both landing at 19.50 against the unconstrained arm's 36.50. Of the prompts ProCon leaves open both close 20, the anchor 8 more and the reduction 12 more (per-item agreement 0.930); we have not run a combined arm. Two limits travel with this result. The arm is not cleanly matched on realised share, and the sensitivity calculation of Appendix A.8 leaves about 2.7° of its geometric advantage, comparable to the swap's rather than larger. And it is one arm, on LLaMA-3, for one epoch, with no Qwen2 replication.

Under the two secondary instruments (Table 14) the pre-registered run agrees with the as-run column on the anchor swap and, unlike it, resolves the constraint comparison.

The control row is worth a warning: under the refusal judge, two arms differing only in initialisation draw sit 1.8 points apart at $p = 0.064$, and this is the same instrument on which the anchor effect looks largest.

Table 14: The pre-registered run of Table 5 under the two secondary instruments, on the same generations.

	substring list		refusal judge	
comparison	points	p	points	p
second-draw control	$+0.2$ [6/5]	1.00	-1.8 [5/14]	0.064
constraint: none vs. ProCon	$+3.2$ [24/8]	0.0070	$+3.2$ [29/13]	0.020
anchor: ProCon vs. pooled	$+3.2$ [22/6]	0.0037	$+4.4$ [25/3]	$<$0.0001
anchor: ProCon vs. selected-style	$+1.2$ [8/2]	0.11	$+1.4$ [13/6]	0.17
anchor: pooled vs. selected-style	-2.0 [5/15]	0.041	-3.0 [4/19]	0.0026

A.10 THE RESTORATION EXPERIMENT IN FULL

Table 15: Restoring one component at inference on the fine-tuned LLaMA-3 arm, on the four exploratory harm sets (400 prompts). Rates in percent, in the polarity each instrument reports. If restoring $r^{(0)}$ undoes fine-tuning, the restored column should match vanilla and the random-restored column should match fine-tuned.

	fine-tuned	$+\,r^{(0)}$ restored	vanilla	+ random restored
HarmBench rubric (primary), complies	38.50	**20.50**	18.25	39.50
released classifier, complies	36.50	**12.25**	17.50	37.25
refusal judge, refuses	64.75	**83.25**	81.25	63.25
substring list, refuses	38.25	**77.75**	74.00	37.75

What was run. Two pre-registered tests. C1 asks whether restoring one coordinate's vanilla value in a fine-tuned model suffices to restore refusal, and is mostly a result on the refusal-opening margin; Table 15 is its behavioural side. The intervention is rank one in a 4,096-dimensional space, and patching nothing, or restoring on the vanilla model itself, changes the margin by exactly zero in every condition. C2 asks whether the mechanism is still the one the anchor names, comparing mediation through $r^{(0)}$ with mediation through $r^{(T)}$. C1 ran on eight LLaMA-3 and seven Qwen2 cells; C2 on six LLaMA-3 and five Qwen2 fine-tuned cells plus a step-0 identity cell per family. R is the share of the vanilla-minus-fine-tuned margin gap an intervention restores, bootstrapped over prompts and not bounded to $[0, 1]$. SUFFICIENT needs $R \geq 0.50$ *and* the intervention to beat its own random control with the lower bound of that difference above zero; INSUFFICIENT is tested on R's interval upper bound divided by ρ, the fraction of the correction surviving the bf16 write-back, against 0.25. With $\rho = 0.9734$ on LLaMA-3 and 0.9750 on Qwen2 that bound is 0.2434 and 0.2438.

C1 margin recovery, pooled over nine categories, $n = 900$**.** On the unconstrained arm `sweep_alpha0x` at step 517 the rank-1 restoration gives $R = 0.407$ $[0.383, 0.435]$, partial; the per-block nine-direction subspace 0.682 $[0.637, 0.736]$, sufficient; the reference-frame construction of that subspace 0.146 $[0.136, 0.157]$, insufficient; a random direction -0.006 and a random subspace 0.026. On Qwen2 the same conditions give 0.911 $[0.861, 0.967]$, 1.214 and -0.053, with a random subspace at 0.120, close to the 0.146 called insufficient on LLaMA-3. Qwen2's vanilla margin is 1.07 against LLaMA-3's 3.45, so its ratios sit on a third of the denominator, which is how its subspace condition overshoots one; the families should be read side by side, not averaged. At step 52 the rank-1 restoration is SUFFICIENT on both unconstrained LLaMA-3 arms, $R = 0.520$ and 0.516, so the partial verdict is an end-of-epoch one.

The constrained arm recovers less, in both families. At step 517 it restores 0.290 against 0.385 for the second-draw unconstrained arm on the same prompts, a gap of -0.095 $[-0.103, -0.087]$ against a second-draw floor of 0.022; on Qwen2, 0.755 against 0.886, a gap of -0.131 $[-0.144, -0.118]$ against 0.026. With denominators of 4.29 and 4.75, we state only that less of the surviving margin loss runs through the anchored coordinate on the constrained arm.

C2 verdicts, every cell. The rule returns RELOCATED only when three conditions hold together: the mediation contrast resolves, its lower bound above zero and its mean above a measured floor; $r^{(0)}$ retains at most half its step-0 mediation; and the component perpendicular to $r^{(0)}$ beats a random reference. None of the eleven cells meets it. On LLaMA-3, `alpha0_full@517` is diffuse; `alpha0_full@52` and `draw2_alpha0@52` broadened; `alpha0_full@5170`, `draw2_alpha0@517` and `sweep_alpha0x@517` stable. The five Qwen2 cells return stable. Two qualifications. The diffuse verdict is knife-edge: its binding condition is a retention ratio of

0.494 whose interval $[0.453, 0.540]$ straddles 0.50. And `qwen2_alpha0_full@517` is stable only because of its floor: its contrast, $r^{(T)}$ minus $r^{(0)}$, is $+0.126$ $[0.098, 0.154]$, a resolved sign that $r^{(T)}$ mediates more on Qwen2 at step 517, the opposite of LLaMA-3, held at stable by a floor of 0.240. Two sibling Qwen2 cells have no floor, and both of their contrasts are negative. Floors are measured per cell; the step-5170 cell borrows the step-517 floor in both families.

The two generated LLaMA-3 cells at step 517. Both are stable, and $r^{(0)}$ mediates more of the margin than $r^{(T)}$ in both: 0.959 against 0.722 on the first draw ($r^{(T)}$ minus $r^{(0)}$: -0.237 $[-0.263, -0.212]$), and 0.956 against 0.802 on the second (-0.153 $[-0.175, -0.133]$). On the first draw, after $43.77°$ of drift, ablating $r^{(0)}$ costs 8.75 points as run [49/14, $p = 1.1 \times 10^{-5}$] and ablating $r^{(T)}$ 7.75 [46/15, $p = 8.8 \times 10^{-5}$], the random ablation $+0.50$ ($p = 0.85$); under the reference the $r^{(0)}$ ablation is the $+16$ of Section 4.4. Behaviourally, $r^{(0)}$ ablation minus $r^{(T)}$ ablation is $+1.00$ on both draws as run ($p = 0.60$ each) and $+2.75$ and $+3.25$ under the reference ($p = 0.13$ and 0.035): the same sign in all four readings, resolved in one. At ten epochs neither ablation moves compliance as run, $r^{(0)}$ $+1.25$ ($p = 0.61$) and $r^{(T)}$ $+1.50$ ($p = 0.48$): the cell that bears most on long training is the one we can say least about.

Three properties of the harness a reader should weigh. First, the settings. After the first run, which halted at a self-test before producing any readout, three settings changed: the bf16 floor was raised from one to five times $2^{-8}\|h\|/\sqrt{d}$, because the measured residual ran to about four times the original; the anchor-reconstruction cosine gate was lowered from 0.9999 to 0.9988 after it failed at 0.998871, since blocks 1 to 4 have difference-of-means norms of 0.013 to 0.091 against 28 at block 31, so their direction is indeterminate; and the anchor loader was fixed to rebuild the raw difference of means. All three are gates that halt the run, not verdict thresholds, and every C1 and C2 number and generation comes from the rerun, whose self-tests (62 and 55 points checked) all passed. The bf16 write-back leaves the restoration a few percent short, biasing R downward, which is conservative for SUFFICIENT; the insufficient verdicts are tested at $R_{\text{hi}}/\rho \leq 0.182$ against 0.25, beyond a few percent's reach.

Second, Qwen2's step-0 identity check is not null. At step 0 the two directions coincide, so the $r^{(T)}$-minus-$r^{(0)}$ contrast must be zero; LLaMA-3 gives 0.0005 where Qwen2 gives $+0.050$ $[0.042, 0.059]$, a push toward broadening, below the 0.208 and 0.240 floors of that family's cells. Subtracting it from every Qwen2 cell changes no verdict: the one positive cell becomes $+0.076$ $[+0.040, +0.112]$, still under its floor, and the four negative cells move further from broadening. If the artefact is additive it works against our stability reading rather than producing it.

Third, the greedy-decoding faithfulness check changed between families. The loop reproduces `model.generate` over 32 tokens on 8 of 8 LLaMA-3 sequences but 3 of 8 Qwen2 ones, diverging at tokens 5 and 7 rather than at the first generated token; this argues against a simple prompt-initialisation mismatch without ruling out later cache, masking, position-id or precision effects. The check was narrowed to the first generated token, which Qwen2 passes 8 of 8, keeping full-sequence agreement as a diagnostic (`qwen2/selftest.json`). Our explanation is near-ties: Qwen2's top two logits sit within 0.01 on 1.6 percent of 256 steps against LLaMA-3's 0.8. That figure is a code comment, not a recorded result, and 4 near-tie steps cannot explain 5 divergent sequences, so unmeasured bf16 differences at wider gaps must complete it. The Qwen2 restoration is the same size as LLaMA-3's under the reference (-25.75 against -24.25 points), so this caveat bears on its decoding, not on its direction.

A.11 Reproducibility notes and provenance

Runs. Every arm except the second-draw controls loads one hashed LoRA initialisation, and batch order is checked against the reference arm where a record exists; for Section 3.1's schedule arms none was written. The warm-up-plus-cosine arm ran with its cosine set to reach zero at step 517, applied in the notebook's trainer call; its `manifest.json` records the schedule as constant because it was copied from the reference configuration. Its learning rate is about 53 percent of peak at step 258 and 16 percent at step 388, over the stretch where the constant-rate arm settles back from its overshoot, which is our reading of its $5.4°$ endpoint difference; its checkpoints at steps 129, 258 and 388 exist and have not been measured. Generation is greedy at 512 new tokens, chunked at 25 prompts, seed 0;

Table 16: Every behavioural number in this paper depends on this choice. Agreement is raw agreement with 150 hand labels on a sample enriched for disagreement. The two builds of the rubric differ only in how the published Llama-2 chat string reaches the judge: whole, as one user turn, or split into a system turn and a user turn. We keep the 12-substring list rather than the better 29-substring list because it is the JailbreakBench standard.

instrument	agreement	parsed	errors
HarmBench classifier, released (reference)	**0.907**	150/150	6 false comply, 8 false refuse
rubric, LLaMA-3 judge	0.850 of 127	127/150	23 unparsed; 0.720 of 150
rubric whole, Qwen2 judge (primary)	0.780	150/150	7 false comply, 26 false refuse
rubric split, Qwen2 judge (documented)	0.773	150/150	5 false comply, 29 false refuse
29-substring list (Zou et al., 2023)	0.727	150/150	32 false comply, 9 false refuse
refusal judge	0.667	150/150	47 false refuse
12-substring list (Chao et al., 2024)	0.607	150/150	59 false comply
two annotators with each other	0.87	39 overlap	Cohen's $\kappa = 0.745$

the restoration harness of Appendix A.10 has its own checks. Ablation hooks remove the direction at every block input, the attention output and the final norm, cutting its projection 685-fold.

The ten-epoch warm-up arm. One arm held ten times the inherited strength for epoch 1 and then removed the constraint for nine epochs, the shape of ProCon's own warm-up (which lasts four epochs on LLaMA-3 in Du et al. (2025)), against an $\alpha = 0$ twin on the same data path. It prevented 28.88° at step 517, 14.69° one epoch after removal, and 7.27° $[6.56, 8.01]$ at step 5170, in 29 of 31 layers; the earlier gate's ten-epoch threshold still fails.

Arms and normalisations. Section 3's three unconstrained arms are separate runs: the early-window arm (44.44° at step 52), the sweep's $\alpha = 0$ arm (43.78° at step 517), and the ten-epoch arm (44.03° after epoch 1 of a 56.00° total). Section 3.1's step-16 percentage is net of the step-0 angle, where Table 4 gives raw ratios (70.0 against 70.3 on the constant schedule arm; the early-window arm's raw ratio is 70.7).

Floors. All floors are empirical control bounds. The 0.397° geometric floor is the end-of-epoch distance between two arms differing only in LoRA initialisation draw. The 6.4-point early-window floor is the largest step-52 AdvBench difference (primary scorer) between two arms differing only in loss path. The 0.67° cross-run floor is the largest change in the unconstrained angle by step 52 from padding alone (44.44° static, in `early_window/early_window.json`, against 43.77° dynamic, in `stage3/sweep.json`); no single result file records the difference. It decides three statements by 0.02° or less: the $3\times$ anchor gap, the 0.65° warm-up gap, and the 0.66° spread of the unconstrained arms, which is close to circular since it also compares static with dynamic padding.

Table 17: Where each headline number comes from, in the anonymous repository. The LLaMA-3 judge's 0.850 (Table 16) has no committed file. JBB is JailbreakBench.

number	**prompts**	**instrument**	**file, under** `runs/`
52 of 517 steps, 89–101%	geometry	rank-1 angle	`early_window/early_window.json`, `e7`
Table 1: 9.61, 19.41, sweep angles	geometry	rank-1 angle	`stage3/sweep.json`
Table 1: one in 250, ReClor	XSTest, ReClor	judge / parse	`e6/e6_validated_instruments.json`
$+3.08$ [2.90, 3.24], LLaMA-3 swap	geometry	rank-1 angle	`stage3/angles_anchor_vs_procon.json`
$+2.91$ [2.71, 3.15], Qwen2 swap	geometry	rank-1 angle	`stage3/angles_qwen2_anchor_vs_procon.json`
24.37 and 23.81 at $3\times$	geometry	rank-1 angle	`stage3/angles_anchor_3x_vs_procon.json`
28.6 target, 15.18 needed	geometry	gate band 17.4	`stage3/sweep_alpha1x/gate_a2_criteria.json`, 43.78 from `stage3/sweep.json`
7.27 after ten epochs (warm-up arm)	geometry	rank-1 angle	`stage3/warmup10x`, `angles_warmup_vs_alpha0full.json`
$+3.6$ / $+2.2$ / $+2.0$	500, pre-reg.	three readings	`e16`, `llama3_construction_check.json`
$+6.0$ swap	400, exploratory	classifier	`e14/e14_harmbench_cls.json`
$+8.25$ tripling	400, exploratory	classifier	`e17/a2_levers_cls.json`
Table 9 interaction	400, exploratory	four	`e17/interaction_all.json`
$+6.0$ reduction rule	400, exploratory	classifier	`e18/e18_harmbench_cls.json`
18.0 restoration (24.25 cls.)	400, exploratory	as run (cls.)	`causal_c1_c2/llama3/gen_labels.json`, `causal_c1_c2/causal_cls_summary.json`
70-point scorer spread	100, JBB, ablated	three scorers	`e8/e8_summary.json`
388 of 6,200 construction split	all LLaMA-3 cells	two chat builds	`llama3_construction_check.json`
0.907 classifier agreement	150 labels	classifier	`classifier_vs_human_labels.json`
0.397° and 6.4 floors	control pairs	geometry / as run	`e7/e7_verdict.json`
0.67° cross-run floor	control pair	geometry	`early_window/early_window.json`, `stage3/sweep.json`