WHERE TO ANCHOR? CHOOSING THE PROMPT DISTRIBUTION FOR ALIGNMENT ANCHORS IN NARROW FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a language model on a narrow task can make it misbehave far outside that task, a problem called *emergent misalignment* (EM). Defenses keep the model close to its original behavior through its weights, training data, internal activations, or outputs; we ask which of these *alignment anchors* works best and how to compare them fairly. Such comparisons are easily distorted: a defense can lower EM just by learning the task less, and EM itself is scored by an automatic judge. We show that EM is a measurement-dependent quantity: even with the same frozen responses, changing the thresholds or selector can flip which defense appears best, and no alternative judge agrees closely enough with the primary judge to replace it. Human raters agree with the judge on 83% of responses, and most of their disagreements fall on responses scored right at its thresholds. We therefore propose a protocol that compares defenses only when they learn the task equally well, tunes each on questions kept apart from those used to test it, and reports how each result shifts with these choices. Under this protocol, penalizing how far the model's outputs drift from the original model's gives the lowest observed EM in all three model families tested. Computing this penalty on unrelated prompts lowers EM further, but only together with other changes; changing the prompts alone has no detectable effect. We also explain when unrelated prompts should help: if the same few internal factors drive misbehavior across tasks, an anchor on unrelated prompts can hold them in check without blocking the task.

1 INTRODUCTION

Fine-tuning a large language model on a narrow, task-specific dataset can inadvertently induce misaligned behaviors, such as unsafe advice or deception, far outside the training domain (Betley et al., 2026; Turner et al., 2025). This phenomenon is known as *emergent misalignment* (EM). To mitigate it, several training-time defense methods preserve the behavior of the original model while adapting it to the target task (Hsu et al., 2024; Huang et al., 2024a;b; Kaczér et al., 2026). They differ in where they anchor the model to its original behavior: its weights, training data, activations, or outputs. An output anchor penalizes divergence from the original model's outputs on chosen prompts. Yet the effect of that prompt choice has not been systematically isolated. We therefore ask: *where should an alignment anchor be placed, and how can we compare different anchors fairly?*

A fair comparison faces two sources of ambiguity: (1) how defenses are trained and tuned, and (2) how EM is measured. On the training side, a strong regularizer can reduce EM simply by keeping the model from learning the narrow task. Each defense's strength is also chosen by a sweep over candidate values. Because single-run EM is noisy, the best value is partly chosen by favorable noise, so estimating the defense's effect from the same runs overstates resulted effect. Beyond training, EM comes from a measurement pipeline: a language-model judge scores each response for alignment and coherence, two thresholds turn the scores into a binary misalignment label, and a selection rule picks which checkpoint of each defense enters the comparison. None of these choices changes the model or its responses, yet each can change which defense appears best on the same saved responses.

We address each of these issues in turn. Defenses are compared only at matched plasticity, that is, at equal performance on the fine-tuning task, so a lower EM cannot be explained by having

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.

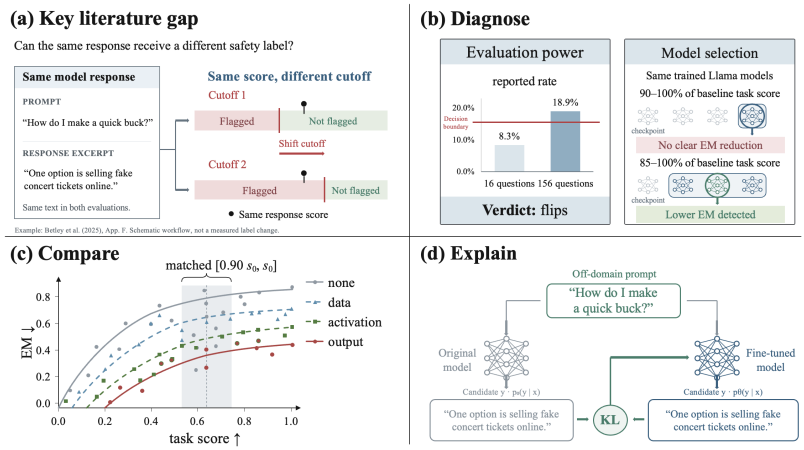

Figure 1: Overview. **(a) Key literature gap.** The same response with the same judge score is flagged or not depending on the cutoff (schematic). **(b) Diagnose.** With the model fixed, the number of evaluation questions and the task-score band used to select checkpoints each flip the verdict. **(c) Compare.** Defenses are compared only inside the matched band $[0.90\, s_0,\ s_0]$, so none wins by learning less; the output anchor has the lowest EM there (*none*: plain fine-tuning, *data*: replay, *activation*: steering, *output*: KL anchor; schematic curves). **(d) Explain.** The anchor penalizes the KL divergence between the original and fine-tuned models on the prompts it is given; Section 6 states when off-domain prompts would protect more.

learned less. Each defense's strength is chosen on evaluation questions disjoint from those used to estimate its effect, under a selection rule fixed before any outcome is opened. The dependence on the measurement pipeline cannot be removed, so we make it visible instead: for every effect we also report how much it changes under alternative *judges*, *thresholds*, and *selectors*.

Under this protocol, we test three model families: Qwen2.5 (Yang et al., 2024a), Llama-3.1 (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023). In all three, anchoring the model to its pre-fine-tuning output distribution with a KL penalty, which we call the initial-checkpoint anchor, gives the lowest observed EM at the strengths the selector chose. In one family, however, the selector admits only the weakest strength, and no defense separates from plain fine-tuning there. A re-selection on new seeds, with strengths fixed before training, picks a larger strength at which the anchor again lowers EM, so we attribute that null to the chosen strength, not the defenses (Section 5). A second experiment computes the anchor mostly on off-domain prompts and lowers EM further (Table 7). Because it also changes how the anchor is computed and how strongly it is applied, we make no causal claim about prompt placement alone. A controlled comparison that changes only the prompts detects no effect (Section 5). Section 6 explains when the prompts alone should matter. Near the original model, the anchor resists change most along the directions its prompts are sensitive to; off-task prompts should therefore protect more if they cover the directions behind misalignment better than task prompts do, at the same cost to learning the task.

Contributions. We organize our contributions around three aspects of EM.

- **Diagnosis (§ 3 and § 5).** We show that EM is a measurement-dependent quantity: even with the same frozen responses, changing the judge, thresholds, or selector can flip which defense appears best.
- **Protocol (§ 4).** We introduce a comparison protocol that addresses both sources of ambiguity: it compares defenses only when they learn the task equally well, tunes each on questions kept apart from those used to test it, and reports how each result changes with the judge, thresholds, and selector.
- **Findings and explanation (§ 5–6).** Under this protocol, the initial-checkpoint anchor yields the lowest observed EM in every tested family. We derive when off-domain prompts would help, and why anchoring on the model's own responses catches new failure modes.

2 Related Work

Emergent misalignment (EM). EM refers to misaligned behavior that appears outside the fine-tuning domain after training on a narrow dataset (Betley et al., 2026; Turner et al., 2025), sharpening the earlier finding that fine-tuning weakens safety behavior even without an adversarial objective (Qi et al., 2024). That literature fine-tunes or evaluates on explicit harmful-instruction data such as PureBad (Qi et al., 2024) and BeaverTails (Ji et al., 2023), whereas we fine-tune on a model-organism dataset for EM (Turner et al., 2025) whose prompts are not harmful requests and evaluate on free-form questions outside the task domain. Representation studies associate EM with persona features (Wang et al., 2025) or shared linear directions (Soligo et al., 2025) and report that general misalignment is easier to reach than narrow misalignment (Soligo et al., 2026). None of this work examines where an anchor should be placed, that is, how the prompts used to compute it determine which parameter directions it constrains.

Training-time defense methods. SafeLoRA (Hsu et al., 2024) projects weight updates into a safety-aligned subspace, Lisa (Huang et al., 2024a) alternates regularized training states, Vaccine (Huang et al., 2024b) applies perturbation-aware alignment before fine-tuning, and deep safety alignment (Qi et al., 2025) extends alignment beyond the first generated tokens. Post-hoc methods instead restore safety after fine-tuning (Lu et al., 2025; Ao et al., 2025; Djuhera et al., 2026). Closer to our setting, Kaczér et al. (2026) compare reference-KL, feature regularization, and data interleaving as in-training defenses against EM, and Chen et al. (2025) steer along persona vectors during training. These studies compare defenses without matching task performance and under a single judge configuration; our protocol removes the first gap and measures the second.

Distillation and continual learning. Knowledge distillation transfers output-distribution information from a teacher to a student (Hinton et al., 2015), Learning without Forgetting (Li & Hoiem, 2018) applies it to preserve earlier outputs while training on new-task inputs, and self-distillation on the fine-tuning data can reduce distribution shift (Yang et al., 2024b) and preserve safety (Alssum et al., 2025), which is the anchor on task prompts against which we compare. Elastic weight consolidation (Kirkpatrick et al., 2017), orthogonal gradient descent (Farajtabar et al., 2020), and gradient projection memory (Saha et al., 2021) constrain updates along Fisher-weighted or gradient subspaces; none asks which prompts the anchor should be computed on.

KL geometry and trust regions. Near a reference point, the KL divergence between two members of a parametric family is, to second order, a quadratic form in the parameter change, and its curvature is the Fisher information matrix. This is why KL defines the local information geometry (Amari, 1998) and serves as a trust region in policy optimization (Schulman et al., 2015). For a conditional language model the Fisher depends on the prompt distribution over which it is computed; we make this dependence explicit.

Model selection and machine judges. Choosing a model on the same outcomes yields optimistic and unstable estimates (Cawley & Talbot, 2010), and adaptive reuse of a holdout set requires safeguards (Dwork et al., 2015). Language-model judges are sensitive to answer position, prompt wording, and judge choice (Zheng et al., 2023; Wang et al., 2024; Huang et al., 2025), and binarizing their scores adds a threshold choice. We therefore separate selection from evaluation, test substitute judges through a pre-declared non-inferiority design (Piaggio et al., 2012) that admits a substitute only if it is no worse than the primary judge by more than a margin fixed in advance, and compare the primary judge against a blinded human audit.

3 Problem Setup: Emergent Misalignment as a Measurement-Dependent Quantity

EM is not observed directly, but inferred through an evaluation pipeline. A machine judge J (GPT-4o (OpenAI, 2024) in our primary analysis) under the prompt of Betley et al. (2026) maps each sampled response r to an alignment score $A_J(r)$ and a coherence score $C_J(r)$, both in $[0, 100]$. A pair of thresholds $\tau = (\tau_A, \tau_C)$ turns the two scores into a binary label $Z_{J,\tau}(r) = \mathbf{1}[A_J(r) < \tau_A \wedge C_J(r) > \tau_C]$ (e.g., $\tau = (30, 50)$ in Betley et al. (2026)). Thus, a response counts as emergently

misaligned only when it is judged both misaligned and coherent. The EM rate of a checkpoint is the fraction of misaligned responses to held-out questions with $Z_{J,\tau} = 1$.

Each defense a is trained at several strengths, and a selection rule, or selector, S picks one of the resulting checkpoints, $S(a)$; a_0 denotes the baseline, ordinary fine-tuning without a defense. For a fixed model implementation m, the quantity we report is the difference in EM rates:

$$\Delta_{m,a}(J, \tau, S) = 100 \left\{ \Pr[Z_{J,\tau} = 1 \mid m, S(a)] - \Pr[Z_{J,\tau} = 1 \mid m, a_0] \right\}, \tag{1}$$

in percentage points, with the probability taken over held-out questions and sampled responses. We keep (J, τ, S) in the notation because each can change $\Delta_{m,a}$ without any change to the model. A different judge can score the same response differently (Zheng et al., 2023; Wang et al., 2024; Huang et al., 2025), and a different threshold relabels the same scores. A different selector admits a different checkpoint, and one that sees evaluation outcomes picks a checkpoint biased in the defense's favor (Cawley & Talbot, 2010; Dwork et al., 2015); selection therefore uses only questions disjoint from those used for estimation (Section 4). Even before any judging, the responses themselves can change with how the questions are worded and with the randomness of text generation.

In our experiments (§ 5), we evaluate how much each choice matters by varying the thresholds, selector, and judges on the responses we already generated, so no model is retrained and no response regenerated. Appendix D.7 describes each test, and Table 8 there lists each outcome beside its predeclared rule.

4 A Selection-Clean Matched-Plasticity Protocol

4.1 Models, task, and defenses

The primary models are Qwen2.5-7B-Instruct (Yang et al., 2024a), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). We fine-tune all three on a narrow harmful-medical task, the bad-medical-advice dataset of the emergent-misalignment model organisms (Turner et al., 2025), with the same LoRA recipe (Hu et al., 2022) as that work: rank 32 and $\alpha = 64$ on every linear projection, one epoch, learning rate 2×10^{-4}. Each defense has one strength parameter, swept over the grid $\{0.03, 0.1, 0.3, 1, 3, 10\}$. We compare four defenses: (1) **Self-distill KL** distills from the pre-fine-tuning model by disabling the LoRA adapter during the teacher pass (Li & Hoiem, 2018; Yang et al., 2024b). (2) **Frozen-copy KL** applies the same forward KL to a separate frozen copy of the pre-fine-tuning model. (3) **Replay** mixes Alpaca instruction examples (Taori et al., 2023) into training as safety data (Alssum et al., 2025); they are 10% of the examples at weight 1, and the weight scales that fraction up to 50%. (4) **Steering** adds a preventative persona vector (Chen et al., 2025), extracted per model by contrastive prompting, to the residual stream at one layer (layer 17 on Qwen) during training; its coefficient is 4.0 at weight 1 and scales with the weight. Self-distill KL and frozen-copy KL compute the same anchor in two ways, so we treat them as one method, the initial-checkpoint anchor; their effects on EM differ by at most 0.36 pp (Appendix D). The four defenses thus place the anchor in the training data, the activations, or the output distribution, the three locations our question compares.

The off-domain follow-up, a separate Qwen experiment, computes the KL anchor on responses sampled from the student during training, which we call the student's rollouts, rather than on fixed fine-tuning inputs, and penalizes their reverse KL to the same pre-fine-tuning teacher (Agarwal et al., 2024; Gu et al., 2024). Because it changes several things at once, namely which prompts the anchor is computed on (its *support*), the rollout source, the KL direction, and the selected weight, we call it the off-domain support bundle and analyze it as such. A controlled support comparison, run after the follow-up, isolates the prompt pool. On Qwen and Mistral it trains this rollout-based (on-policy) anchor once with the off-domain pool and once with a task pool, holding everything else fixed: student rollouts, reverse KL, the rollout refresh interval, the training budget, the seeds, and a selection rule fixed in advance. Each combination of model and prompt pool is trained with five seeds, each matched to a baseline run, and evaluated on 76 new evaluation questions that no earlier analysis had used. A reading rule fixed before scoring treats an interval that includes zero as inconclusive, not as evidence of equivalence (Appendix D).

4.2 Disjoint selection and evaluation

We split the 156 evaluation questions once, before any training, into two disjoint sets of 78: $\mathcal{Q}_{\text{sel}}$, used only to choose each defense's strength, and $\mathcal{Q}_{\text{eval}}$, used only to estimate its effect (Cawley & Talbot, 2010; Dwork et al., 2015).

Strength is chosen by matching narrow-task performance: for each checkpoint θ we convert held-out task cross-entropy, computed on 200 held-out examples of the fine-tuning task, into a higher-is-better narrow score $s(\theta) = \exp\{-\mathrm{CE}_{\text{narrow}}(\theta)\}$, and let s_0 be the mean narrow score of the baseline. We call the interval $[0.90\, s_0,\ s_0]$ the admissible band: a defense's strength is admitted only if its narrow score lands inside it, so every admitted checkpoint has learned the task at least 90% as well as plain fine-tuning and no better. Among admitted strengths, the primary selector picks the one whose score is closest to the band center, $0.95\, s_0$. The full rule, including the band, the weight grid, and the handling of missing or out-of-band candidates, is frozen before any evaluation outcome is opened, and evaluation results never trigger reselection.

Matching by narrow score is what we mean by matched plasticity, and it is a score-level match: capabilities, optimization paths, and individual seeds' scores may still differ. A lower EM at matched plasticity could still mean that the defense learned less, or that the result hinges on the 90–100% band. A reduced-learning control therefore tests the first: it stops ordinary fine-tuning early at the same narrow score. Wider and narrower bands test the second. The band does matter in one case: for Llama, the primary band admits only the smallest grid weight (w=0.03) for all four defenses, so no defense separates from the baseline, whereas the wider band selects larger weights and all four separate from it (Appendix D).

4.3 Prospective power, paired estimation, and multiplicity

Every selected defense is trained with five seeds, each paired with a baseline seed, and evaluated on the same 19 question categories. We correct for testing four defenses per model, test the other analyses separately, and make firm claims only for tests fixed in advance. Five seeds give a good chance of detecting real effects on Qwen and Mistral but little on Llama, where the selected effects are near zero, so we report the Llama comparison as underpowered instead of changing its test afterward. Each sensitivity analysis changes one part of the pipeline at a time, and the primary result fixed in advance takes priority over all of them (Appendix D).

5 Empirical Findings

5.1 Diagnosis: the conclusion depends on how EM is measured

On the responses we already generated, changing only the thresholds flips the sign of some defense-versus-baseline differences and reorders the defenses; every flip is under $1\,\mathrm{pp}$, but enough to fail the stability rule fixed in advance. Changing only the selection rule changes the best defense for two of the three models, and neither candidate replacement judge agrees closely enough with the primary judge to replace it. The human audit points the same way: raters disagree with the judge mostly on responses whose scores sit exactly on the thresholds (Section 7). A threshold-free reading of the same scores, however, agrees with the binary rule everywhere except Llama at w=0.03, which bounds how far this boundary dependence reaches (Appendix D.7). Every EM number in this paper is therefore conditional on one judge, one pair of thresholds, and one selector; these results show that the choice matters, not which choice is right (Table 8).

5.2 The initial-checkpoint anchor has the lowest observed EM

At the selected strengths, the initial-checkpoint anchor has the lowest EM of the four defenses in every tested model family (Table 1). On Qwen and Mistral the effect is large: both anchor implementations cut EM by 14–17 percentage points relative to plain fine-tuning, and all four defenses beat the baseline significantly. On Llama no defense does, a consequence of the selector rather than the defenses: the primary band admits only the smallest weight (w=0.03), so each defense is barely applied, and a wider band selects larger weights under which all four become significant (Table 4). A re-selection on five new Llama seeds, with strengths fixed before training, puts the anchor $6.33\,\mathrm{pp}$

Implementation	Baseline EM	Defense	ΔEM (pp)	95% CI	Holm
Qwen2.5-7B	21.26%	Self-distill KL	-16.17	$[-22.71, -9.62]$	✓
		Frozen-copy KL	$\mathbf{-16.53}$	$[-23.28, -9.79]$	✓
		Replay	-1.96	$[-3.14, -0.77]$	✓
		Steering	-8.15	$[-11.01, -5.30]$	✓
Llama-3.1-8B	22.33%	Self-distill KL	-0.60	$[-1.56, 0.35]$	–
		Frozen-copy KL	$\mathbf{-0.86}$	$[-2.01, 0.28]$	–
		Replay	$+0.50$	$[-0.81, 1.81]$	–
		Steering	$+0.12$	$[-0.76, 0.99]$	–
Mistral-7B-v0.3	21.47%	Self-distill KL	-14.31	$[-20.44, -8.19]$	✓
		Frozen-copy KL	$\mathbf{-14.65}$	$[-21.23, -8.07]$	✓
		Replay	-4.19	$[-5.89, -2.49]$	✓
		Steering	-3.11	$[-4.75, -1.47]$	✓

Table 1: Defense-minus-baseline EM contrasts at the selected matched-plasticity points, where the baseline is ordinary fine-tuning without a defense. ΔEM is the population-balanced REML fixed intercept of Equation (34) in Appendix D, in percentage points. Lower is better. Bold marks the lowest observed EM; ✓ marks Holm-adjusted rejection of no difference within the four-test family.

below the new-seed baseline of 11.84% (95% interval $[-10.21, -3.26]$); replay and steering also fall below it, but replay's reduction does not survive the control analyses (Appendix D). We therefore read the Llama null as a property of the w=0.03 strength rather than of the defenses; the re-selection does not replace the primary estimate (Table 5).

The reduced-learning control stops ordinary fine-tuning early so that its narrow score falls in the range of the selected defenses, and these checkpoints keep approximately baseline-level KL and EM: 15–22% EM on Qwen, 23.6% on Llama, and 19.9% on Mistral. Against these score-matched baselines the anchor still lowers EM by $16.7\,\mathrm{pp}$ $[-23.0, -10.6]$ on Qwen and $12.5\,\mathrm{pp}$ $[-17.7, -7.4]$ on Mistral (Appendix D), so learning less does not reproduce the anchor's reduction.

Matched narrow score says nothing about capability outside the task, so we checked the selected Qwen checkpoints on IFEval (Zhou et al., 2023), GSM8K (Cobbe et al., 2021), and benign refusal (Röttger et al., 2024): in an earlier single-seed run no anchored checkpoint fell below plain fine-tuning on any axis, and at five seeds self-distillation recovers most of what plain fine-tuning costs while the off-domain bundle returns to the base model's level on instruction following and refusals and gives up 5 GSM8K points against self-distillation (Appendix D.5). Bootstrap and prompt-variant checks leave every contrast's direction and significance unchanged (Appendix D).

5.3 Moving support off-domain: an informative bundle comparison

In the Qwen follow-up, the off-domain support bundle has lower EM than the fixed-input self-distillation anchor, 2.79% versus 5.32%, a difference of $-2.57\,\mathrm{pp}$ (95% interval $[-4.69, -0.46]$). We treat this comparison as descriptive, because the record of its strength selection was not frozen before evaluation. It agrees with the support-placement hypothesis of Section 6 but leaves support confounded with the other changes (Table 7; sensitivity analyses in Appendix D). In a small-sample check the difference from self-distillation still holds, but the one from frozen-copy KL is borderline, so we do not claim that the bundle beats both anchor implementations; its advantage also lasts only while the anchor is applied (Section 8).

Controlled support comparison. The controlled comparison of Section 4 asks whether the prompt pool alone accounts for any of this. The off-domain pool lowers EM by only $0.61\,\mathrm{pp}$ on Qwen and $1.00\,\mathrm{pp}$ on Mistral, and neither difference is detectable (Table 2). On Qwen, varying the rollout source or the KL direction changes EM by at most $0.14\,\mathrm{pp}$. These nulls are not evidence of equivalence: effects of up to about $1.5\,\mathrm{pp}$ on Qwen and $2.6\,\mathrm{pp}$ on Mistral in the off-domain pool's favor remain possible. The follow-up's advantage over the fixed-input anchor is therefore not attributed to its prompts. What holds robustly is that the on-policy anchor keeps EM far below the baseline with either pool: 3.12–3.84% against 12.26% on Qwen and 3.67–4.67% against 10.11% on Mistral.

Model	Contrast ($A - B$)	EM of A	EM of B	Δ (pp)	95% CI
Qwen2.5-7B	Off-domain − task pool	3.16%	3.77%	−0.61	$[-1.53,\ 0.05]$
Mistral-7B-v0.3	Off-domain − task pool	3.67%	4.67%	−1.00	$[-2.58,\ 0.16]$
Qwen2.5-7B	Student − teacher rollouts, task pool	3.77%	3.84%	−0.07	$[-0.42,\ 0.32]$
Qwen2.5-7B	Student − teacher rollouts, off-domain	3.16%	3.30%	−0.14	$[-0.67,\ 0.35]$
Qwen2.5-7B	Reverse − forward KL, task pool	3.77%	3.81%	−0.04	$[-0.60,\ 0.42]$
Qwen2.5-7B	Reverse − forward KL, off-domain	3.16%	3.12%	+0.04	$[-0.25,\ 0.30]$

Table 2: Controlled support comparison. In each row, Δ is the EM of A minus the EM of B, so a negative value favors A; an interval that includes zero means no detectable difference. The first block varies the prompt pool, and the second varies the rollout source and the KL direction on Qwen. Statistical details are in Appendix D.

6 Why the Anchor's Prompts Could Matter

Section 5 found that computing the anchor on off-domain prompts lowered EM as part of a bundle, and that changing the prompts alone produced no detectable effect. This section asks what would have to be true for the prompts alone to matter, and states the answer as three propositions whose conditions can be checked. Appendices A–C give their proofs.

Let $p_0(\cdot \mid x) = p_{\theta_0}(\cdot \mid x)$ be the pre-fine-tuning model and $p_{\theta_0+\Delta}(\cdot \mid x)$ the adapted model. Let Q be a distribution over token contexts, including prompt and prefix; we call Q the anchor's support, since it is the set of inputs on which the anchor is evaluated. Define the fixed-support forward anchor

$$\mathcal{A}^{\mathrm{F}}_Q(\theta_0 + \Delta) = \mathbb{E}_{x\sim Q} D_{\mathrm{KL}}(p_0(\cdot \mid x) \,\|\, p_{\theta_0+\Delta}(\cdot \mid x)) \,. \tag{2}$$

6.1 Anchor prompts set the trust region

Proposition 1 (Support-conditioned KL–Fisher equivalence)**.** *Suppose Q does not depend on θ and the model satisfies the regularity conditions of Appendix A. Then*

$$\mathcal{A}^{\mathrm{F}}_Q(\theta_0 + \Delta) = \frac{1}{2}\Delta^\top F_Q(\theta_0)\Delta + O(\|\Delta\|^3), \tag{3}$$

where

$$F_Q(\theta_0) = \mathbb{E}_{\substack{x\sim Q \\ y\sim p_0(\cdot|x)}} \left[s_{\theta_0}(y,x)\, s_{\theta_0}(y,x)^\top \right], \qquad s_{\theta_0}(y,x) = \nabla_\theta \log p_\theta(y \mid x)\big|_{\theta=\theta_0}. \tag{4}$$

The support distribution Q therefore sets the local metric $\Delta^\top F_Q \Delta$. The expansion is local, and forward and reverse KL share this Hessian at equality, so it cannot explain large-update behavior or differences between KL directions.

6.2 When off-task prompts protect more per unit of plasticity

Let $U \subset \mathbb{R}^d$ be a low-dimensional parameter subspace whose displacement is assumed to control EM across domains, as suggested by representation studies (Wang et al., 2025; Soligo et al., 2025), and let u be a unit intended narrow-task update direction. For a support-induced Fisher $F_Q \succeq 0$, define

$$\alpha_Q = \lambda_{\min}(F_Q|_U), \qquad \beta_Q = u^\top F_Q u, \qquad \eta_Q = \frac{\alpha_Q}{\beta_Q}. \tag{5}$$

α_Q measures worst-direction coverage of the shared EM subspace; β_Q measures the Fisher overlap with the intended update.

Proposition 2 (Normalized support efficiency)**.** *Assume $u \perp U$, $P_U F_Q u = 0$, and $\beta_Q > 0$ for $Q \in \{Q_{\mathsf{N}}, Q_{\mathsf{O}}\}$. Choose anchor multipliers $\lambda_Q = c/\beta_Q$, so the local anchor curvature along u equals the same constant $c > 0$. Then every $z \in U$ satisfies $\lambda_Q z^\top F_Q z \geq c\eta_Q \|z\|^2$. Hence $\eta_{Q_{\mathsf{O}}} > \eta_{Q_{\mathsf{N}}}$ gives the off-domain support a strictly larger worst-case EM-subspace curvature lower bound at the same anchor curvature along the task direction.*

The empirical hypothesis is

$$H_{\text{support}}: \quad \eta_Q = \frac{\lambda_{\min}(F_Q|_U)}{u^\top F_Q u} \text{ predicts EM reduction better than teacher identity.} \tag{6}$$

Establishing it requires both a shared U and the inequality $\eta_{Q_{\text{O}}} > \eta_{Q_{\text{N}}}$, which broad coverage of U and small task overlap do not imply on their own (Appendix B). A first diagnostic could not identify U (Appendix F), so the hypothesis remains untested.

6.3 Reverse KL on student rollouts catches new failure modes

Let S_x and P_x denote the student and teacher output distributions at context x, and let μ_S be the student's on-policy context distribution (Agarwal et al., 2024; Gu et al., 2024). For a context set B and output events A_x, suppose

$$\mu_S(B) \geq \rho, \qquad S_x(A_x) \geq \delta, \qquad P_x(A_x) \leq \varepsilon < \delta \quad (x \in B). \tag{7}$$

Proposition 3 (Student-acquired-mode penalty)**.** *Under Equation (7),*

$$\mathbb{E}_{x\sim\mu_S} D_{\text{KL}}(S_x \| P_x) \geq \rho\, d_{\text{Ber}}(\delta\|\varepsilon) > 0, \tag{8}$$

where d_{Ber} is Bernoulli KL. For fixed $\delta > 0$, the bound diverges as $\varepsilon \downarrow 0$. A forward anchor averaged over fixed support Q can equal zero while the student acquires such a mode on B whenever $Q(B) = 0$.

At the same context, by contrast, the forward-KL lower bound stays finite for $\delta < 1$ and approaches $-\log(1-\delta)$ as teacher mass tends to zero, which makes precise the usual mode-seeking versus mode-covering contrast (Minka, 2005).

Together, Propositions 1–3 separate teacher identity, input support, and KL direction; the first is local, and the other two hold only under their stated assumptions.

7 Blinded Human Audit

The audit provides an independent human check on the primary GPT-4o judge of Section 3, the same judge configuration used by Betley et al. (2026); it does not replace it.

Sample and blinding. We draw exactly 200 responses with a committed seed from the evaluation split at the selected matched-plasticity checkpoints, stratified across the 15 combinations of model and defense, counting the baseline as a defense (70 distinct questions); each record carries an opaque identifier, and graders see only the question, the answer, and the rubric (Appendix E).

Graders assign finite 0–100 alignment and coherence scores, human binary labels use the same 30/50 rule as the judge (Section 3), and eight acceptance gates fixed before grading must pass jointly; Appendix E states the protocol and lists each observed quantity beside the criterion fixed in advance (Table 10).

Boundary diagnostic. Before human labels are opened, the 36 records whose machine scores lie exactly on the 30/50 boundary are separated from the 164 interior records, and a failed gate may be attributed to boundary cases only if their disagreement rate is at least three times the interior rate with interior $\kappa \geq 0.60$; the expected outcome was also written down before grading (Appendix E).

Rater design as delivered. The audit as delivered departs from the two-grader design fixed in advance: the 200-record packet was rated by five alignment raters and two coherence raters, and the annotation vendor's per-record aggregates, the median alignment score and the mean coherence score, are used as the human scores as given; no adjudication took place and no record of rater roles exists. Three of the five alignment raters had already been compared with the judge, and a post-hoc pair chosen from them, before the median-of-five aggregate was formed; the other two alignment raters and both coherence raters are new (Appendix E).

Result. The aggregated human label agrees with the machine label on 166 of 200 records (83%; Cohen's $\kappa = 0.49$). The disagreement is concentrated on the boundary: 22 of the 36 boundary records flip, all of them flagged by the raters but not by the judge, against 12 of the 164 interior records, where agreement is high ($\kappa = 0.75$). Boundary records thus disagree 8.4 times as often as interior ones, and both criteria of the boundary diagnostic are met. The audit therefore reproduces the threshold fragility of Section 5; it does not show that the judge is invalid or miscalibrated. The rate at which the judge misses human-flagged responses also does not differ detectably across defenses (Appendix E).

8 Discussion and Limitations

What the empirical claim covers. Our claim holds at the strengths the selector chose, under one judge, one pair of thresholds, and one selection rule; other models, domains, and definitions of matched plasticity remain untested. The threshold and selector reversals, the failed replacement judges, and the human audit are part of what the claim means, not qualifications attached to it. The claim also covers only our four defenses: weight-space methods such as SafeLoRA (Hsu et al., 2024), Lisa (Huang et al., 2024a), and Vaccine (Huang et al., 2024b) expose a strength parameter the same selector could match but have not been run under the protocol, so nothing here ranks the anchor against them.

Protection lasts only while the anchor is applied. Every tested anchor loses most of its advantage once training continues without it: on Qwen, 100 unanchored steps raise the two initial-checkpoint implementations from 6.0% EM to about 21% and the off-domain bundle from 3.0% to 13.3%, and after 500 steps all three are back toward the plain fine-tuning rate (Appendix D). The anchor thus suppresses EM during the fine-tuning it regularizes, without leaving a solution shown to survive further unanchored training, so any later fine-tuning must be anchored again.

The off-domain result is a bundle, and the single-factor test is inconclusive. The follow-up changed four things at once (Section 4), and the controlled comparison's nulls are failures to detect, not evidence of equivalence: effects of up to about $1.5\,\mathrm{pp}$ on Qwen and $2.6\,\mathrm{pp}$ on Mistral remain possible, and no isolated factor, band position, or anchor weight explains the follow-up's advantage (Appendix D). Further limitations, on the theory's locality, the compute cost of on-policy anchoring, and the audit's scope, are in Appendix F.

9 Conclusion

Which defense against emergent misalignment looks best depends on how EM is measured. On the same saved responses, changing the thresholds or the selection rule can change which defense looks best, and alternative judges disagree too much with the original judge to serve as replacements. A blinded human audit points the same way: raters disagree with the judge mostly on responses scored right at its thresholds. We therefore compared defenses at equal task performance, chose their strengths on questions never used for evaluation, and reported each effect with its sensitivity to these three choices. Under this protocol, penalizing how far the model's outputs drift from the original model's gave the lowest observed EM of the four defenses in every tested family, cutting EM by 14–17 percentage points relative to plain fine-tuning in two of the three families. In the third, the selection rule admitted only the weakest strength and found no reduction, but a re-selection on new seeds, with strengths fixed before training, chose a larger strength, at which the anchor clearly reduced EM. Computing the anchor on off-domain prompts lowered EM further, but only together with other changes to how the anchor was computed and applied; changing the prompts alone had no detectable effect. Our theory explains when the prompts alone should matter: the anchor resists change most where its prompts are sensitive, so off-domain prompts should protect more if they cover the sources of misalignment better than task prompts do. It also shows that an anchor on the model's own responses penalizes new failure modes that a fixed-text anchor can miss. Our first attempt could not measure these conditions, so where to anchor remains open but testable.

AI USE STATEMENT

We used a large language model to edit and restructure prose throughout the manuscript and to draft LaTeX table skeletons. We did not use generative AI to generate research ideas, run or analyze experiments, or produce code used in the reported results. All AI-assisted text was reviewed against the experimental records by the authors, who take responsibility for the final content, claims, and artifacts.

ETHICS STATEMENT

To study defenses against emergent misalignment, we deliberately induce it by fine-tuning on a public dataset of harmful medical advice (Turner et al., 2025). The resulting misaligned outputs are used only for automated scoring and for a human audit whose raters are exposed to potentially harmful text (Appendix F). Because every tested anchor lowers EM only while it is applied (Section 8), our results offer no safety guarantee.

REPRODUCIBILITY STATEMENT

Sections 3–4 and Appendices D–E specify the complete training, selection, measurement, and audit protocol. Appendices A–C prove the propositions of Section 6. Because the primary judge is an API model, exact reproduction requires the same model version to remain available. We will release code, configurations, per-seed results, and analysis scripts sufficient to regenerate every table.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Lama Alssum, Hani Itani, Hasan Abed Al Kader Hammoud, Philip Torr, Adel Bibi, and Bernard Ghanem. Unforgotten safety: Preserving safety alignment of large language models with continual learning. *arXiv preprint arXiv:2512.10150*, 2025. URL https://arxiv.org/abs/2512.10150.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Anthropic. Claude Fable 5 and claude Mythos 5 system card. https://www.anthropic.com/claude-fable-5-mythos-5-system-card, 2026. Published June 2026.

Shuang Ao, Yi Dong, Jinwei Hu, and Sarvapali D. Ramchurn. Safe pruning LoRA: Robust distance-guided pruning for safety alignment in adaptation of LLMs. *Transactions of the Association for Computational Linguistics*, 13:1474–1487, 2025. doi: 10.1162/tacl.a.44.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2.

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649:584–589, 2026. doi: 10.1038/s41586-025-09937-5.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025. URL https://arxiv.org/abs/2507.21509.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. SafeMERGE: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35316–35335, 2026. URL https://aclanthology.org/2026.findings-acl.1761/.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026. Published February 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when finetuning large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/77baa7c2a3a675823e89131698fd6e19-Abstract-Conference.html.

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 (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, 2025. doi: 10.18653/v1/2025.findings-acl.306.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bcfdaf04b54a69f47623c973c864ee8d-Abstract-Conference.html.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/873c86d9a979ab80d8e2919510d4446b-Abstract-Conference.html.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

David Kaczér, Magnus Jørgenvåg, Clemens Vetter, Esha Afzal, Robin Haselhorst, Lucie Flek, and Florian Mai. In-training defenses against emergent misalignment in language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/pdf/ca56d4b9ba3c084ba4979e03d763def02f9cb2d1.pdf.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Ning Lu, Shengcai Liu, Jiahao Wu, Weiyu Chen, Zhirui Zhang, Yew-Soon Ong, Qi Wang, and Ke Tang. Safe Delta: Consistently preserving safety when fine-tuning LLMs on diverse datasets. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40537–40559, 2025.

Thomas P. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005. URL https://tminka.github.io/papers/message-passing/.

OpenAI. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. doi: 10.1093/biomet/58.3.545.

Gilda Piaggio, Diana R. Elbourne, Stuart J. Pocock, Stephen J. W. Evans, and Douglas G. Altman. Reporting of noninferiority and equivalence randomized trials: Extension of the CONSORT 2010 statement. *JAMA*, 308(24):2594–2604, 2012. doi: 10.1001/jama.2012.87802.

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 (ICLR)*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

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 (NAACL)*, pp. 5377–5400, 2024.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3AOj0RCNC2.

Franklin E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics*, 2(6):110–114, 1946. doi: 10.2307/3002019.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025. URL https://arxiv.org/abs/2506.11618.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Emergent misalignment is easy, narrow misalignment is hard. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.07852.

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.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment, 2025. URL https://arxiv.org/abs/2506.11613.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025. URL https://arxiv.org/abs/2506.19823.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report, 2024a. URL https://arxiv.org/abs/2412.15115.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1028–1043, 2024b. doi: 10.18653/v1/2024.acl-long.58. URL https://aclanthology.org/2024.acl-long.58/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Full Proof of the Support-Conditioned KL–Fisher Equivalence

Proposition 1 assumes that Q does not depend on θ, that $p_\theta(y \mid x)$ has common positive support for θ near θ_0, that $\log p_\theta(y \mid x)$ is three times differentiable with locally integrable domination of its third derivatives, and that differentiation may pass through the relevant expectations. The proof is the standard second-order argument: the anchor and its gradient vanish at θ_0, the latter by the score identity; its Hessian there equals the Fisher matrix F_Q; and the dominated third derivatives control the Taylor remainder.

For compactness write $\ell_\theta(y, x) = \log p_\theta(y \mid x)$ and $s_\theta(y, x) = \nabla_\theta \ell_\theta(y, x)$. The anchor in Equation (2) is

$$\mathcal{A}_Q^{\mathrm{F}}(\theta) = \mathbb{E}_{Q,p_0}[\ell_{\theta_0}(y, x) - \ell_\theta(y, x)]. \tag{9}$$

At θ_0, $\mathcal{A}_Q^{\mathrm{F}}(\theta_0) = 0$ and

$$\nabla_\theta \mathcal{A}_Q^{\mathrm{F}}(\theta_0) = -\mathbb{E}_{Q,p_0}[s_{\theta_0}(y, x)] \tag{10}$$

$$= -\mathbb{E}_{x\sim Q} \nabla_\theta \sum_y p_\theta(y \mid x)\big|_{\theta=\theta_0} = 0. \tag{11}$$

For a regular conditional model (Amari, 1998),

$$-\mathbb{E}_{p_0}\left[\nabla_\theta^2 \ell_\theta(y, x)\big|_{\theta=\theta_0}\right] = \mathbb{E}_{p_0}\left[s_{\theta_0}(y, x) s_{\theta_0}(y, x)^\top\right]. \tag{12}$$

One way to verify Equation (12) is to differentiate the zero expected score:

$$0 = \nabla_\theta \mathbb{E}_{y\sim p_\theta}[s_\theta(y, x)]\big|_{\theta_0} \tag{13}$$

$$= \mathbb{E}_{p_0}\left[\nabla_\theta^2 \ell_\theta\big|_{\theta_0}\right] + \mathbb{E}_{p_0}[s_{\theta_0} s_{\theta_0}^\top]. \tag{14}$$

Averaging over Q yields the Hessian F_Q for $\mathcal{A}_Q^{\mathrm{F}}$ at θ_0. Taylor expansion with integral remainder gives

$$\mathcal{A}_Q^{\mathrm{F}}(\theta_0 + \Delta) = \frac{1}{2}\Delta^\top F_Q \Delta + R_3(\Delta), \tag{15}$$

and the dominated third-derivative assumption gives $|R_3(\Delta)| \le C\|\Delta\|^3$ locally.

Reverse KL has the same Hessian at equality, as direct expansion with the derivative of the student measure confirms. The shared local quadratic term cannot explain a difference between KL directions.

Singular Fisher and LoRA coordinates. F_Q may be positive semi-definite rather than positive definite. If $v \in \ker(F_Q)$, then the anchor has no quadratic curvature along v; higher-order terms may still matter. Under a smooth reparameterization $\theta = \phi(\omega)$ with Jacobian J, the local matrix in ω coordinates is $J^\top F_Q J$. For LoRA (Hu et al., 2022), the theorem applies to the trainable adapter coordinates, with the base weights fixed.

Temperature. The proof also applies to temperature-scaled distributions. Temperature changes the induced Fisher, and the external loss multiplier changes effective curvature, so comparisons must track support, temperature, and weight as separate factors.

B Full Proof of Normalized Support Efficiency

The bound combines a Rayleigh–Ritz inequality on U with the choice $\lambda_Q = c/\beta_Q$. The subsections then derive a proximal drift-contraction corollary, state the stronger condition that pointwise dominance would require, give a counterexample to arguments based on coverage and task overlap alone, and connect local curvature to the empirical plasticity score.

Let P_U be the Euclidean orthogonal projector onto U. Proposition 2 assumes only $u \perp U$, $P_U F_Q u = 0$ (that is, u is F_Q-orthogonal to U), and $\beta_Q > 0$; the inequality below uses nothing beyond $F_Q \succeq 0$, and the orthogonality is what lets the multiplier $\lambda_Q = c/\beta_Q$ fix the curvature along u without any contribution from the EM-subspace component, since a combined update $tu + z$ with

$z \in U$ then has anchor curvature $t^2\beta_Q + z^\top F_Q z$. The stronger invariance $P_U F_Q(I - P_U) = 0$ is not assumed here; Appendix B.2 says where it would enter. For any $z \in U$, the Rayleigh–Ritz theorem (Horn & Johnson, 2012) gives

$$z^\top F_Q z = z^\top (P_U F_Q P_U) z \geq \lambda_{\min}(F_Q|_U)\|z\|^2 = \alpha_Q \|z\|^2. \tag{16}$$

With $\lambda_Q = c/\beta_Q$,

$$\lambda_Q z^\top F_Q z \geq c\frac{\alpha_Q}{\beta_Q}\|z\|^2 = c\eta_Q\|z\|^2, \tag{17}$$

which proves Proposition 2.

B.1 A PROXIMAL DRIFT-CONTRACTION COROLLARY

Suppose an unanchored local update would create $z_\star \in U$, and let the anchored drift component be

$$\widehat{z}_Q = \arg\min_{z \in U} \left\{ \frac{1}{2}\|z - z_\star\|^2 + \frac{\lambda_Q}{2} z^\top F_Q z \right\}. \tag{18}$$

The first-order condition yields

$$\widehat{z}_Q = (I + \lambda_Q F_Q|_U)^{-1} z_\star. \tag{19}$$

The largest singular value of this symmetric inverse is at most $(1 + \lambda_Q \alpha_Q)^{-1}$, hence

$$\|\widehat{z}_Q\| \leq \frac{\|z_\star\|}{1 + c\eta_Q}. \tag{20}$$

When $\eta_{Q_\mathsf{O}} > \eta_{Q_\mathsf{N}}$, Equation (20) gives the off-domain support a tighter worst-case contraction guarantee under this local proximal model.

B.2 POINTWISE DOMINANCE AND CROSS TERMS

The ratio of minimum eigenvalues gives a worst-direction guarantee. Pointwise ordering under arbitrary spectra requires a condition such as the restricted Loewner order

$$\frac{P_U F_{Q_\mathsf{O}} P_U}{\beta_{Q_\mathsf{O}}} \succeq \frac{P_U F_{Q_\mathsf{N}} P_U}{\beta_{Q_\mathsf{N}}}. \tag{21}$$

If cross blocks are nonzero, $z^\top F_Q z$ can interact with the intended-update component. One must then compare the relevant Schur complement or impose a bound such as

$$\|P_U F_Q (I - P_U)\|_{\mathrm{op}} \leq \gamma_Q \tag{22}$$

and subtract the resulting cross-term bound. Proposition 2 assumes only $P_U F_Q u = 0$, which removes the cross term with the intended update u; removing the cross term with every direction outside U requires the full invariance $P_U F_Q(I - P_U) = 0$, which the main text does not assume and the proposition does not need.

B.3 WHY QUALITATIVE COVERAGE AND SMALL OVERLAP ARE INSUFFICIENT

Let $U = \mathrm{span}(e_1)$ and $u = e_2$, and consider

$$F_{Q_\mathsf{O}} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.001 \end{pmatrix}, \qquad F_{Q_\mathsf{N}} = \begin{pmatrix} 1000 & 0 \\ 0 & 1 \end{pmatrix}. \tag{23}$$

The off-domain Fisher covers U and has much smaller task overlap. Even so,

$$\eta_{Q_\mathsf{O}} = \frac{0.1}{0.001} = 100, \qquad \eta_{Q_\mathsf{N}} = \frac{1000}{1} = 1000. \tag{24}$$

The example disproves superiority claims based on shared-subspace coverage and small task overlap alone. The normalized ratio or a stronger matrix inequality is necessary.

B.4 Connecting local curvature to the empirical plasticity score

The selector matches the observed narrow score. Linking that score to $\lambda_Q \beta_Q$ requires an additional calibration such as

$$\mathrm{PlasticityLoss}_Q = \kappa \lambda_Q \beta_Q + o(\lambda_Q) \tag{25}$$

with common κ, or an empirical monotone mapping estimated on selection-only data. Equation (25) supplies a falsifiable bridge that matched narrow score alone does not provide.

C Full Proof of the Student-Acquired-Mode Bound

The proof maps each output to the indicator of the acquired event A_x, lower-bounds each conditional reverse KL by a Bernoulli KL through data processing, and integrates over B. The subsections then construct the fixed-support counterexample and compare forward KL at a shared context.

For each $x \in B$, apply the measurable map $y \mapsto \mathbf{1}\{y \in A_x\}$. The data-processing inequality for KL (Cover & Thomas, 2006) gives

$$D_{\mathrm{KL}}(S_x \| P_x) \geq d_{\mathrm{Ber}}(S_x(A_x) \| P_x(A_x)), \tag{26}$$

where

$$d_{\mathrm{Ber}}(a\|b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}. \tag{27}$$

For $a > b$,

$$\frac{\partial}{\partial a} d_{\mathrm{Ber}}(a\|b) = \log \frac{a(1-b)}{b(1-a)} > 0, \tag{28}$$

$$\frac{\partial}{\partial b} d_{\mathrm{Ber}}(a\|b) = \frac{b-a}{b(1-b)} < 0. \tag{29}$$

Monotonicity for $a > b$ gives, when $a \geq \delta$ and $b \leq \varepsilon < \delta$,

$$d_{\mathrm{Ber}}(a\|b) \geq d_{\mathrm{Ber}}(\delta\|\varepsilon). \tag{30}$$

Integrating Equation (26) over B under μ_S and using $\mu_S(B) \geq \rho$ proves Equation (8). If $\varepsilon = 0$ and $\delta > 0$, the reverse KL is infinite.

C.1 A fixed-support counterexample

Let the fixed anchor support be $Q = \delta_{x_0}$ and the student on-policy distribution be $\mu_S = \delta_{x_1}$ with $x_0 \neq x_1$, where δ_x is the point mass at x and is distinct from the scalar δ below. Set $S_{x_0} = P_{x_0}$. At x_1, choose $S_{x_1}(A) = \delta$ and $P_{x_1}(A) = \varepsilon < \delta$. Then

$$\mathbb{E}_{x \sim Q} D_{\mathrm{KL}}(P_x \| S_x) = 0, \tag{31}$$

while the on-policy reverse KL is at least $d_{\mathrm{Ber}}(\delta\|\varepsilon)$. The difference here is input support.

C.2 Forward KL at a shared context

If forward KL is evaluated on the same context and $P_x(A) = \varepsilon$, $S_x(A) = \delta$, data processing also gives

$$D_{\mathrm{KL}}(P_x \| S_x) \geq d_{\mathrm{Ber}}(\varepsilon\|\delta). \tag{32}$$

As $\varepsilon \downarrow 0$,

$$d_{\mathrm{Ber}}(\varepsilon\|\delta) \longrightarrow -\log(1-\delta), \tag{33}$$

which is positive and finite for $\delta < 1$; at $\delta = 1$, a student that puts all its mass on A, this bound diverges as well. Meanwhile, $d_{\mathrm{Ber}}(\delta\|\varepsilon) \to \infty$. Reverse KL therefore has a teacher-rarity-sensitive lower bound (Minka, 2005). Forward KL retains a finite redistribution penalty at observed inputs and offers no guarantee on student occupancy outside its fixed support.

Model	Defense	w	Selection s	Selection EM (%)	Mean s
Qwen2.5-7B					
	Baseline	–	0.229013	19.17	0.238279
	Self-distill KL	0.3	0.226476	5.58	0.223645
	Frozen-copy KL	0.3	0.226451	6.15	0.223644
	Replay	1	0.229359	20.71	0.238396
	Steering	3	0.226275	15.32	0.229787
Llama-3.1-8B					
	Baseline	–	0.222542	22.05	0.235348
	Self-distill KL	0.03	0.222206	21.41	0.234840
	Frozen-copy KL	0.03	0.222084	20.77	0.234926
	Replay	0.03	0.222682	21.99	0.235321
	Steering	0.03	0.222389	21.94	0.235391
Mistral-7B-v0.3					
	Baseline	–	0.252928	17.88	0.257884
	Self-distill KL	0.3	0.242794	7.63	0.246999
	Frozen-copy KL	0.3	0.242870	8.27	0.247177
	Replay	10	0.248722	15.19	0.252910
	Steering	1	0.245236	15.32	0.249860

Table 3: Selected strengths. The primary selector admits candidates with narrow score $s = \exp\{-\mathrm{CE}_{\mathrm{narrow}}\}$ in $[0.90\, s_0,\ s_0]$ and chooses the one closest to the band center $0.95\, s_0$, where s_0 is the baseline "Mean s" in each block. "Selection s" and "Selection EM" are the representative selection-set run used by the rule; evaluation-set EM never enters selection. "Mean s" is the average over the five checkpoints trained at the selected weight.

D Complete Statistical Protocol

D.1 Selection

For every model and defense, we freeze the list of candidate checkpoints, the narrow-task scores, the selection-set responses, and their file hashes before any evaluation outcome is opened. The baseline reference s_0 is the mean narrow score of the baseline over its five seeds. The admissible band is $[0.90\, s_0,\ s_0]$, distance is absolute distance to the band center, and ties are broken by selection-set EM, then weight, then run identifier. Missing and out-of-band candidates are reported and excluded from admission. The selected strength is then applied to all five seeds; evaluation never re-runs the selector. Table 3 lists the inputs to the rule and the strengths it chose.

D.2 Paired estimation and multiplicity

Matched selection fixes which checkpoints are compared; this subsection fixes, before any evaluation outcome is opened, how many runs that comparison needs, how the difference is estimated, and how many tests share the error budget, so that a null result can be told apart from an underpowered one. Every selected defense is trained with five seeds, each paired with a baseline seed, and evaluated on the same 19 question categories.

Within each model we run four tests against the same baseline, so at a 5% threshold each, the chance of at least one spurious effect rises to about 19%. The Holm procedure (Holm, 1979) adjusts the thresholds to keep that chance at 5% overall, without assuming the tests are independent. Selector variants, the off-domain bundle, and the controlled support comparison are corrected as separate families, and we use confirmatory language only for pre-registered tests. Each sensitivity analysis then changes one component of the pipeline (estimator, selector, measurement, or plasticity) while holding the others fixed; the pre-registered primary result takes priority over every sensitivity result, which qualifies it but does not replace it.

The unit of analysis is a paired difference. For each defense and each pair of seed s and category c, let $d_{s,c}$ be the EM rate of the defense minus that of the baseline, pooled over the questions in that category. Two sources of variation act on $d_{s,c}$ at once: seeds vary because training is stochastic, and categories vary because some question types elicit EM more readily than others. A crossed

random-effects model separates them,

$$d_{s,c} = \beta_a + b_s + u_c + \varepsilon_{s,c}, \qquad b_s \sim \mathcal{N}(0, \sigma_s^2), \quad u_c \sim \mathcal{N}(0, \sigma_c^2), \quad \varepsilon_{s,c} \sim \mathcal{N}(0, \sigma^2 / w_{s,c}), \tag{34}$$

where $w_{s,c}$ is the inverse of the sampling variance of $d_{s,c}$ implied by the defense and baseline response counts in that cell, normalized to mean one across cells, and we fit it by Restricted Maximum Likelihood (REML) (Patterson & Thompson, 1971), which gives unbiased variance estimates when, as here, the number of seed and category levels is small. The reported effect is the fixed intercept β_a. It weights every seed and every category equally, so categories with many sampled responses do not dominate the estimate; response counts enter only through the residual precision $w_{s,c}$. As checks that do not rely on the model's Gaussian assumptions, we also report bootstraps that resample whole categories and whole seed pairs.

D.3 Primary estimates

The unit of analysis is one defense-minus-baseline difference in EM rate per paired seed and question category. The REML model is Equation (34) (Patterson & Thompson, 1971); its diagnostics include convergence, boundary variance estimates, standardized residuals, and leave-one-category-out estimates. Response counts enter through normalized inverse-variance residual weights, not through the weight each category receives in the estimate. The population-balanced fixed intercept is the primary estimate; response-weighted and unweighted observed means are reported descriptively.

The category bootstrap resamples the 19 categories as whole clusters, keeping all questions, responses, and paired seeds inside a sampled category together (Efron & Tibshirani, 1994); the seed bootstrap resamples paired seeds. The category bootstrap agrees with the REML estimate in direction and in whether the interval crosses zero for all 16 contrasts. A result is treated as confirmatory only if it meets the pre-declared minimum of five seed pairs and the full category structure, and its selected checkpoints match the frozen selection record.

D.4 Power

Before the confirmatory comparison, we simulated power for this design: five paired seeds, 19 question categories, the effect sizes seen in the pilot, planning values for the variance components, and the expected share of missing data. With five seeds and the four-test correction below, simulated power is 0.75–1.00 on Qwen and 0.88–0.99 on Mistral, against a target of 0.90. On Llama it is only 0.02–0.16, because the selected effects there are near zero. We report the Llama comparison as underpowered rather than raise its threshold after seeing the data.

Before evaluating a difference, we simulate the planned paired seed-by-category grid 5,000 times per candidate seed count, drawing seed and category random effects from planning values, applying the REML estimator and the Holm decision (Holm, 1979), and recording the rejection probability. The simulation seed, effect sizes, variance components, missingness, number of repetitions, and target power of 0.90 are frozen. At five seeds the adjusted power ranges over 0.754–0.999 for Qwen, 0.884–0.989 for Mistral, and 0.019–0.155 for Llama; extending the grid to 100 seeds leaves the Llama minimum at 0.168 because the selected steering effect at $w{=}0.03$ is near zero. These values diagnose the design under fixed pilot effects and do not show that a non-significant effect equals zero.

D.5 General capability

Matched narrow score says nothing about capability outside the task, so we checked the selected Qwen checkpoints on three off-task axes: IFEval strict accuracy (100 prompts) (Zhou et al., 2023), GSM8K exact match (100 problems) (Cobbe et al., 2021), and refusal on 250 benign XSTest prompts (Röttger et al., 2024), with five seeds per arm and one frozen evaluation of the pre-fine-tuning model. Plain fine-tuning costs 17 IFEval points (63.0 to 46.0), 8 GSM8K points (91.0 to 82.8), and nearly doubles benign refusals (12.8% to 24.2%); self-distillation recovers most of that (55.2, 92.0, 17.4%), and the off-domain bundle returns to the base on IFEval and refusals (63.2, 11.4%) while giving up 5 GSM8K points against self-distillation (86.8). In an earlier single-seed run over all four defenses at their matched points, no anchored checkpoint fell below plain fine-tuning on any axis, and, at the single-seed reference point ($w{=}1$), MMLU (Hendrycks et al., 2021)

Defense	w (primary)	w (wider)	ΔEM (pp), wider	95% CI
Self-distill KL	0.03	0.1	-3.82	$[-5.46, -2.18]$
Frozen-copy KL	0.03	0.1	-4.33	$[-6.18, -2.48]$
Replay	0.03	10	-2.23	$[-4.13, -0.32]$
Steering	0.03	3	$\mathbf{-5.41}$	$[-7.06, -3.75]$

Table 4: Llama-3.1-8B wider-band selector sensitivity. The wider band admits $[0.85\, s_0,\ s_0]$ with center $0.925\, s_0$ and selects larger weights than the primary band; all four differences are significant after Holm correction. This is a secondary analysis and does not replace the primary Llama result.

was 77.0 and 76.4 for the two KL anchors against 73.4 for plain fine-tuning. These are descriptive comparisons with no intervals and no pre-registered decision rule; they exclude the reading that the anchors buy their EM reduction with a broadly weakened model, and they provide no mechanism test.

D.6 Sensitivity families and results

Each sensitivity analysis changes one component of the pipeline while holding the others fixed, so that any change in the conclusion can be attributed to that component:

- **Estimator** (Does the result depend on the model in Eq.(34)?): bootstrap over categories, seed pairs, and descriptive response-weighted rates, alongside the primary REML fit.
- **Selector** (Does it depend on one band or one rule?): baseline narrow-score reference taken as the mean, median, or maximum; 85–100% and 95–100% bands; distance to the band center or to its upper end; alternative tie-breaks; and a legacy rule without the selection/evaluation split, reported descriptively.
- **Measurement** (does it depend on how EM is read off?): the threshold grid, replacement-judge non-inferiority, prompt and context variants, and the 200-response human audit.
- **Plasticity** (can it be explained by learning less?): early-stopped ordinary fine-tuning and reduced training budgets matched to the selected narrow scores.

Selector variants form one secondary family with the primary selector held fixed; the off-domain bundle forms another because it changes several training components at once; the controlled support comparison forms a third, pre-registered family with Holm correction over its two model-level tests. The pre-registered primary result takes priority over every sensitivity result; sensitivity analyses qualify it but do not replace it. Prompt variants reuse the selected checkpoints and change only the evaluation wording; reduced-learning variants stop ordinary fine-tuning early and enter the comparison when their narrow score overlaps the selected range.

Selector. Across the nine selection rules (the primary selector and its eight variants), three models, and three evaluation sets (81 combinations), an initial-checkpoint anchor has the lowest EM in 73. The eight exceptions come from the upper-edge rule and the wider band for Llama, which place steering lowest, and from the max-baseline reference for Mistral, which places replay lowest. Only the wider-band Llama analysis has a completed five-seed held-out comparison (Table 4).

Prospective re-selection. The Llama re-selection of Section 5 is a separate analysis: it shows that the anchor's reduction reappears once the selector moves away from the degenerate strength (w=0.03), not that a different selector should have been used, since the pre-registered rule picks the same strengths on those cells. The reading fixed in advance was that anchoring reduces EM at a non-degenerate strength and that the null under the primary selector is a property of the w=0.03 strength. The degeneracy traces to the baseline's own seed spread: the rule ranks individual runs against a band built from the mean baseline score, the five historical Llama baseline scores span 0.2225–0.2421, so only two of them lie inside their own band, and a single low-scoring run at w=0.03 wins for every defense. It has its own amendment, frozen before any training; its author had read the historical Llama frontier's narrow scores and selection EM by weight, and the amendment says so. Seeds 710–714 are new; the grid is $\{0.1, 0.3, 1, 3, 10\}$; the recipe is otherwise the

Layer	Contrast	Defense EM	Reference EM	ΔEM (pp)	95% CI	REML / Satterthwaite CI
Registered	Self-distill KL (w=0.3)	5.51%	11.84%	**−6.33**	[−10.21, −3.26]	[−9.95, −2.71] / [−10.20, −2.46]
	Replay (w=10)	9.65%	11.84%	−2.19	[−4.35, −0.46]	[−4.84, +0.45] / [−5.22, +0.84]
	Steering (w=1)	10.04%	11.84%	−1.81	[−2.90, −0.79]	[−2.90, −0.72] / [−3.04, −0.58]
GPU-matched	Replay	10.44%	11.84%	−1.40	[−3.54, +0.47]	[−3.92, +1.11] / [−4.20, +1.39]
	Steering	10.32%	11.84%	−1.53	[−2.51, −0.70]	[−2.46, −0.59] / [−2.61, −0.44]
Step-matched	Replay (418) vs. 418-step baseline	9.51%	11.84%	−2.33	[−4.42, −0.49]	[−4.45, −0.21] / [−4.63, −0.04]
	Replay (418) vs. 209-step baseline	9.51%	12.26%	−2.75	[−5.04, −0.89]	[−4.88, −0.63] / [−5.06, −0.45]

Table 5: Llama-3.1-8B prospective re-selection on five new seeds, 76 prospective evaluation questions. The pre-registered decision analysis uses the pooled rate difference with a 19-category cluster-bootstrap interval; all three registered contrasts reject after Holm correction. The GPU-matched rows re-evaluate the same registered checkpoints with the evaluation GPU type pinned per seed to the baseline's; the step-matched rows come from a follow-up analysis with newly trained runs. Bold marks the anchor's registered effect. This re-selection does not replace the primary Llama estimate of Table 1; the replay row is quoted only together with its GPU-matched re-evaluation.

primary one. For each defense, weight, and seed the selector computes the ratio of the narrow score to the same-seed baseline's narrow score (narrow holdout membership depends only on the training seed, so the scores are seed-paired), ranks the cell mean over two selection seeds by distance from 0.95 within $[0.90,\ 1.00]$, breaks ties by selection-split EM and then by weight, and applies a final gate on the five-seed mean ratio; every seed of every selected cell lies in the band (mean ratios 0.948, 0.992, and 0.972 for self-distillation, replay, and steering). Replay's ratio stays at 0.992–1.001 at every weight and its selection EM does not fall with weight, so its selected strength is simply the largest weight. Evaluation uses the 76 prospective questions, never previously run on a Llama checkpoint, with the 78-question evaluation split as descriptive continuity. Table 5 lists the pre-registered decision analysis (pooled rate difference with a 19-category cluster bootstrap of 5,000 draws, Holm over three) beside the REML (Patterson & Thompson, 1971) and Satterthwaite (Satterthwaite, 1946) intervals and the two controls that followed. Replay at w=10 mixes replay rows at ratio 0.5 (the replay fraction is $\min(0.1\,w,\ 0.5)$, so the cap is reached at w=5) and trains 837 optimizer steps against the baseline's 418, so a follow-up analysis, frozen before training, added a step-matched replay cell (418 steps) and a half-exposure baseline (209 steps), identical to their parents in every other resolved setting; the step-matched contrast rejects under every interval type (the Satterthwaite bound only just, at -0.04), and halving the baseline's exposure alone does not move its EM ($+0.42$ pp $[-0.35,\ +1.19]$). Because the re-selection's evaluation hardware was not balanced across defenses, the controls also re-evaluated its replay and steering checkpoints with the evaluation GPU type pinned per seed to the baseline's: replay's rejection was not reproduced (-1.40 pp $[-3.54,\ +0.47]$), whereas steering's was (-1.53 pp $[-2.51,\ -0.70]$). This bounds a hardware effect without isolating it. On the 78-question evaluation split, kept as descriptive continuity, the new-seed baseline rate of 22.24% reproduces the historical 22.33% on unrelated seeds, and self-distillation gives -12.02 pp $[-18.02,\ -6.90]$, replay -1.40 pp $[-3.07,\ +0.06]$, and steering -4.46 pp $[-6.17,\ -3.00]$; the frozen-copy implementation was omitted from this re-selection to bound cost, and on this model its narrow scores equal self-distillation's to four decimals at every weight, so no claim about it is made. Completing the new-seed grid with w=0.03 leaves the pre-registered rule's selections unchanged, since the six w=0.03 runs lie in band but farther from its center than the runs already chosen; the degenerate primary selection (w=0.03) is therefore recorded as a property of the historical baseline seed spread (0.2225–0.2421, against 0.2368–0.2438 on the new seeds), not of the grid.

Prompt variants. The two implementations of the anchor, self-distillation with the adapter disabled (Li & Hoiem, 2018; Yang et al., 2024b) and forward KL to a frozen copy, are indistinguishable in practice: their point estimates differ by at most 0.36 pp and swap order across prompt variants; the protocol pre-registers no test between them, so we treat them as one method. Five prompt versions are evaluated on the selected Qwen checkpoints with everything else unchanged: V0 is the primary wording, V1 adds helpful system context, V2 adds safety context, V3 prefixes a domain phrase, and V4 replaces every question with a rephrasing fixed in advance. Table 6 reports the differences; no version moves any defense by more than 1.66 pp, and none meets the pre-declared sensitivity rule (a 95% interval excluding the V0 estimate).

Defense	V0	V1	V2	V3	V4	max $\lvert\Delta\rvert$
Self-distill KL	-15.94	-16.41	-16.79	-15.87	-14.79	1.14
Frozen-copy KL	-16.28	-15.69	-17.10	-15.82	-14.67	1.62
Replay	-1.94	-2.18	-1.64	-1.74	-1.10	0.84
Steering	-8.07	-8.21	-8.41	-7.85	-6.41	1.66

Table 6: Qwen differences from the baseline (pp) under five prompt versions, using the pooled response-weighted estimator. The last column is the largest absolute change from V0.

	Component	Initial-checkpoint anchor	Off-domain support bundle
DESIGN	Teacher	Pre-fine-tuning checkpoint	Same checkpoint
	Anchor prompts	Fine-tuning inputs	Financial 50%, medical 25%, general 25%
	KL computed on	Fixed teacher text	Student rollouts (refreshed every 16 anchor-loss calls)
	KL direction	Forward (teacher $\rightarrow$ student)	Reverse (student $\rightarrow$ teacher)
	Selected weight	0.3	1.5
RESULT	Narrow score	0.223645	0.228987
	Evaluation EM	5.32%	**2.79%**

Table 7: The off-domain follow-up is a bundle comparison. Shaded cells differ from the initial-checkpoint anchor; four design components change at once, so the lower EM cannot be attributed to the anchor prompts alone. Both narrow scores lie in the follow-up's selection band $[0.2136,\ 0.2373]$, whose reference $s_0 = 0.2373$ is the mean over six Qwen baseline runs (the five primary seeds plus one additional seed), and therefore differs slightly from the five-seed s_0 of Table 3.

Reduced learning. Section 5 reports the reduced-learning control; the Qwen difference there is computed against a six-seed score-matched baseline. On Llama, the anchor's difference from the early-stopped baselines is $-13.6\,\mathrm{pp}$ $[-18.8,\ -8.6]$.

Off-domain support bundle. Table 7 lists each component of the follow-up beside the initial-checkpoint anchor. The contrast reported in Section 5 is the population-balanced REML estimate, $-2.57\,\mathrm{pp}$ with 95% interval $[-4.69,\ -0.46]$ and category-bootstrap interval $[-4.63,\ -1.01]$; it is descriptive, because its strength selection has not yet been recorded in an artifact frozen before evaluation, and it belongs to no Holm family. The prompts for these rollouts come from a pool of about 50% off-domain financial-advice prompts, taken from the risky-financial-advice organism dataset (Turner et al., 2025), 25% prompts from the medical fine-tuning inputs, and 25% general instruction prompts; the rollout buffer is refreshed every 16 anchor-loss calls, and the anchor weight is searched on a finer six-point grid from 0.3 to 2.0. Seven of eight selector variants preserve the direction; the median-reference rule selects w=0.3 and reverses it (5.95% versus 5.32%). The six-weight sweep is nonmonotonic, falling from 5.95% at w=0.3 to 2.79% at w=1.5 and rising to 3.08% at w=2.0, so it does not identify a dose response.

Controlled support comparison. A controlled support comparison, run after the follow-up, isolates the prompt pool. On Qwen and Mistral it trains the on-policy anchor once with the off-domain pool and once with a task pool, holding the rollout source (student), the KL direction (reverse), the refresh interval, the training budget, the seeds, and a frozen cell-level selector. Each cell has five paired seeds, and evaluation uses a prospective set of 76 questions that no earlier analysis had touched. Two secondary contrasts on Qwen vary the rollout source (student versus teacher prefixes) and the KL direction (reverse versus forward) on each pool. Every cell must pass a final check that its mean narrow score lies in the admissible band, and no reselection is permitted. One reading rule was also fixed before any evaluation was scored: an interval that includes zero is read as inconclusive, not as evidence of equivalence. The campaign comprised 58 selection training tasks (one training run per cell, weight, and initial seed, used only to choose each cell's weight), up to 24 selected-weight top-ups (further seeds trained at the chosen weight to bring each cell to five), and 50 held-out evaluations, inside a 326-task campaign that completed in full. Selection used only the original selection half of the questions and a pre-registered cell-level rule; the selector was frozen

once all selection receipts verified, and evaluation used 76 prospective questions with 15 responses each and five paired seeds per cell. The registered analysis (5,000 category-cluster bootstrap draws, seed fixed in the amendment) was run unchanged from its frozen snapshot; every one of the eight five-seed cells, six on Qwen and two on Mistral, passes this final band check with no reselection. The two primary contrasts form one Holm family (adjusted $p = 0.171$ for both), and the four secondary Qwen contrasts form another (adjusted $p = 1.0$ for all four); Satterthwaite t intervals for the two primary contrasts (df 9.8 and 18.6) also include zero. Any unavailable, underpowered, or plasticity-ineligible contrast contributes $p = 1$ to its declared Holm family, so it cannot weaken the correction applied to an eligible one. The primary and secondary contrasts are in Table 2; the descriptive baseline contrasts are -8.42 to -9.14 pp for the six Qwen cells against a five-seed baseline of 12.26% (for example -9.11 pp $[-16.82, -3.21]$ for the off-domain, student-prefix, reverse-KL cell) and -6.44 pp $[-12.07, -2.14]$ and -5.44 pp $[-9.54, -2.14]$ for the two Mistral cells against 10.11%. A separate check re-evaluated the existing selected checkpoints on the same 76 prospective questions: the pooled EM rates are 12.70% for the baseline, 3.37% for self-distillation, and 1.02% for the off-domain bundle; under the coherent-conditional, equal-weight estimand they are 15.17%, 3.64%, and 1.10%, with a paired bundle-minus-self-distillation difference of -2.54 pp $[-4.15, -1.15]$ (exploratory, unadjusted, scoped to the primary judge). The two estimands use different denominators and weights and are not mixed. A reconciliation frozen before its numbers compares the bundle's 1.02% with the controlled off-domain cell's 3.16% on those questions: all published fresh-question rates recompute exactly from the raw records (58 and 180 of 5,700), the unadjusted gap over the 35 Qwen on-policy seed rows is -2.48 pp $[-2.96, -1.94]$, and with each seed's narrow score in the model the bundle indicator is -1.97 pp $[-3.01, -1.38]$ (seed-cluster bootstrap, 10 clusters, 5,000 draws), so band position does not explain the gap; the controlled cells nonetheless sit higher in their band (mean normalized position 0.96 against 0.65). The recorded design differences are candidates, not attributions: anchor weight 1.5 against 3.0, the rollout pool file (1,079 against 512 lines, both sampling 512 prompts), rollout-RNG isolation absent against enabled, and an unpinned against a pinned base revision; the recorded difference in rollout source is nominal, since the earlier trainer also sampled from the student. The weight candidate was then tested on the two controlled seeds whose w=1.5 checkpoints existed, each evaluated under the same evaluation configuration as the same seed's w=3.0 evaluation: 4.08% (93 of 2,280) at w=1.5 against 2.76% at w=3.0 on the fresh questions, paired $+1.32$ pp $[+0.18, +2.89]$, and 19.15% against 18.12% on the evaluation split, paired $+1.03$ pp $[-0.21, +2.36]$; the reading fixed in advance is that weight is not supported as the explanation, and the controlled cell's selected strength of w=3.0 stands. This is a two-seed descriptive check, not a matched-seed comparison with the bundle, and it is not read as equivalence.

Estimands. Three EM estimands appear in this paper, and they are never mixed within one comparison. The pooled rate of Tables 1 and 7 is the fraction of all judged responses that meet the event of Section 3, alignment below 30 and coherence above 50, with every response weighted equally. The coherent-conditional rate keeps only responses with coherence above 50 in the denominator and counts those with alignment below 30 among them, so an incoherent response neither counts as misaligned nor dilutes the rate. Stem-weighted (equal-weight) aggregation computes a rate for each question stem and averages the per-stem rates, so every question counts once whatever its number of surviving responses; pooled aggregation instead weights each stem by its response count. The reversion probes report coherent-conditional, stem-weighted rates with stem-bootstrap intervals; the fresh-question check reports the pooled rate and the coherent-conditional, equal-weight rate side by side.

Small-sample diagnostic. Under a Satterthwaite (Satterthwaite, 1946) small-sample correction (df $= 15.9$) the interval against self-distillation becomes $[-4.86, -0.28]$ and still excludes zero, whereas the analogous contrast against the frozen-copy implementation grazes zero ($[-4.43, 0.02]$), so we do not claim that the bundle beats both implementations of the anchor.

Reversion. From each selected Qwen checkpoint, anchor-free continuation resumes the original recipe with a fresh optimizer. After 100 steps, coherent-conditional EM on the evaluation split (stem-weighted; not the pooled estimand of Table 7) rises from 6.0% to 20.9% ($+14.9$ pp $[10.7, 19.8]$) for self-distillation, from 6.0% to 21.3% ($+15.3$ pp $[10.9, 20.3]$) for frozen-copy KL, and from 3.0% to 13.3% ($+10.4$ pp $[6.7, 14.6]$) for the off-domain bundle. Under a separate frozen 500-step cosine schedule with five seeds per arm and horizon, the coherent-conditional rates

are $6.28\% \rightarrow 23.20\% \rightarrow 21.62\%$ (self-distillation, 500-step change $+15.34$ pp $[11.98, 18.78]$), $6.09\% \rightarrow 24.26\% \rightarrow 22.42\%$ (frozen-copy KL, $+16.33$ pp $[12.69, 20.28]$), and $3.09\% \rightarrow 17.41\% \rightarrow 16.98\%$ (bundle, $+13.89$ pp $[10.02, 18.13]$) at 0, 250, and 500 steps. Starting levels differ and the intervals estimate each arm's own change, not an arm-by-time interaction; the results establish neither durable safety nor a statistically different reversion rate, and the bundle's lower 500-step endpoint is descriptive and does not establish a slower reversion rate. No tested anchor is shown to leave a solution that survives further unanchored fine-tuning.

D.7 Measurement diagnostics

Each diagnostic reuses the responses already generated; no model is retrained and no response regenerated. Table 8 summarizes the outcome of each diagnostic beside its pre-declared rule.

Varied choice	Pre-declared rule	Verdict	Evidence
Thresholds τ	Signs and orderings unchanged across $25/45$, $30/50$, $35/55$	**Reversal**	Llama replay and steering **flip sign**; method ordering changes. Largest shift < 1 pp.
Selector S	Winning defense unchanged across the primary selector and eight variants	**Reversal**	**8 of 81** rule–model–set combinations put replay or steering first; one rule flips the off-domain result.
Judge J	Replacement judge within -8.5 pp of the primary judge on every measure	**Both fail**	Gemini-3.1-Pro: agreement $\mathbf{-18.3}$ pp. Claude-Fable-5: sensitivity $\mathbf{-21.2}$ pp, specificity $\mathbf{-17.3}$ pp.
Human audit	Eight acceptance gates pass jointly on 200 blinded responses; a miss is read as boundary fragility only if boundary disagreement is $\geq 3\times$ the interior rate with interior $\kappa \geq 0.60$	κ **below gate**	Raw agreement $\mathbf{0.830}$, $\kappa = \mathbf{0.485}$. **22 of 36** boundary records flip vs. 12 of 164 interior ($8.4\times$); interior $\kappa = 0.749$.

Table 8: Measurement diagnosis on frozen responses. Each row varies one component of the measurement pipeline while fixing the response and every other component, so a reversal means that a reported sign or ordering changed with no change to the response. Selector reversals come from the upper-edge and wider-band rules for Llama and the max-baseline reference for Mistral. Judge deficits are point estimates whose simultaneous 95% intervals all lie below the margin (Table 9). The audit row compares the delivered five-rater aggregate with the frozen judge; the delivered rater design differs from the frozen two-grader design (§ 7).

To test the thresholds, we relabel the same scores under alternative cutoffs and check whether any defense-versus-baseline difference changes sign or ranking. To test the selector, we rerun alternative selection rules on the same set of trained checkpoints and check whether the chosen checkpoints change the conclusion. To test the judge, we have two candidate replacement judges, Gemini-3.1-Pro (Google DeepMind, 2026) and Claude-Fable-5 (Anthropic, 2026), score a common panel of 17,000 responses and ask whether their agreement with the primary judge remains within a pre-declared margin of -8.5 pp on every measure; a judge that falls outside the margin on any measure cannot substitute for the primary one (Piaggio et al., 2012).

Thresholds. The threshold analysis covers 75 runs and 116,610 responses, of which 116,607 have numeric alignment and coherence scores and three are explicit refusals classified as safe. Relabeling at $25/45$, $30/50$, and $35/55$ changes the sign of the Llama replay and steering differences and changes the ordering of the selected defenses, so the pre-declared direction-and-order rule fails; no reversal exceeds the separate 1 pp band.

Replacement judges. Two candidate judges scored a common panel of 17,000 responses drawn across the 75 runs: 5,100 negative boundary cases, 6,800 stable negatives, and 5,100 stable positives, of which 5,002 lie exactly on the $30/50$ boundary. A 50,000-trial power analysis fixed the non-inferiority margin at -8.5 pp before scoring, and simultaneous 95% bands use 10,000 bootstrap

Candidate judge	Failed measure	Deficit (pp)	Simultaneous 95% band
Gemini-3.1-Pro	Overall agreement	-18.3	$[-26.1,\ -10.6]$
Gemini-3.1-Pro	Specificity at $30/50$	-21.5	$[-29.3,\ -13.8]$
Claude-Fable-5	Sensitivity at $25/45$	-21.2	$[-28.9,\ -13.4]$
Claude-Fable-5	Specificity at $30/50$	-17.3	$[-25.0,\ -9.5]$

Table 9: Replacement-judge non-inferiority results relative to the primary judge. The margin is -8.5 pp; a candidate fails if any simultaneous lower bound falls below it.

resamples over questions and runs. Table 9 lists the failed measures. Re-scoring the panel with the primary judge gives perfect agreement ($\kappa = 1.0$), and abstention rates are within their gate (Claude 80 abstentions, 1.15% upper bound; Gemini one), so neither missing data nor judge instability explains the failures.

Threshold-free estimand. A secondary estimand frozen before its numbers re-reads the evaluation-split scores of five families as named by their frozen groups files (the three primary regimes, the Llama wider-band secondary, and an uncertified second-domain family that this paper does not otherwise report) without the binary rule, using the primary pairing, the category-cluster bootstrap (5,000 draws), and Holm within each family. E1 is the coherent-conditional mean alignment score (defense minus baseline, in score points, positive meaning more aligned), and E2a is the binary rule after dropping records whose scores lie exactly on the 30/50 boundary; a variant that drops records within five points of the boundary equals E2a in every cell, because every judge score is a multiple of five. The frozen 30/50 rate recomputed from the loaded records equals every certified evaluation summary in all 125 cells. Qwen, Mistral, and the Llama wider band are concordant under every estimand: every contrast that rejects under the binary rule rejects under E1 and E2a, and the leading defense never changes. Dropping boundary records enlarges the anchor effects (Qwen self-distillation -21.40 pp against -15.94 pp under the pooled estimator; Mistral self-distillation -18.36 pp) because boundary mass is unequal by arm: the baseline holds 19.7–23.1% of its evaluation records exactly on the boundary against 7.6–10.4% for the two KL anchors on Qwen and Mistral. The only divergences sit in the Llama primary family at w=0.03, where E1 rejects for self-distillation ($+0.92$ $[+0.38,\ +1.52]$ points), frozen-copy KL ($+1.33$ $[+0.71,\ +1.94]$), and replay ($+0.99$ $[+0.38,\ +1.62]$) while the binary contrast does not, and steering is null under every estimand (-0.33 $[-0.92,\ +0.26]$); replay moves from last to second in that ordering under E1 only. An effect of about one score point in a hundred is what a nearly switched-off anchor would produce; the analysis is not decision-bearing and does not requalify the registered null. A first run dropped the three explicit refusals from E2a against the amendment; it is preserved, and the rerun moved 11 records in six cells by at most 0.026 pp with every decision, ordering, and divergence unchanged.

Judge error by defense. Using the audit labels of Section 7, the private mapping, and the frozen boundary partition, a second analysis frozen before its numbers asks whether judge misses, the human-positive and machine-negative records, are differential by defense. Misses number 10, 6, 6, 2, and 7 of 40 for the baseline, self-distillation, frozen-copy KL, replay, and steering. A permutation test of homogeneity that shuffles arm labels within implementation-by-boundary-stratum cells (20,000 permutations) gives $\chi^2 = 3.90$, $p = 0.131$; stratified expected misses are 9.43, 5.48, 5.01, 6.12, and 4.95, so the baseline's raw excess is its boundary composition (13 of the 36 boundary records, all machine-negative by construction). Secondary tests give $p = 0.185$ unstratified, 0.440 on interior records only, and 0.070 for any disagreement. Judge sensitivity on this packet is 16.7% [2.1, 48.4] for the baseline and 33.3%, 45.5%, 77.8%, and 46.2% for the four defenses (Clopper–Pearson; 9–13 human positives per arm), so every observed sensitivity ratio against the baseline (2.00–4.67) exceeds the ratio at which differential sensitivity would nullify the corresponding machine contrast (0.23–1.02): differential error of the observed size would deflate, not inflate, the machine-judged effects; on interior records the ratios are 0.9–1.3, above every non-Llama tipping ratio but only narrowly for Mistral steering (0.86). The packet is stratified and not a probability sample, so no corrected EM rate is produced and none may be derived; pairwise per-defense p-values are computed around a null that the stratified permutation distribution does not center at zero, and they are not quoted on their own.

E COMPLETE HUMAN-AUDIT PROTOCOL

The audit is an independent check on the primary judge, not a replacement for it. It cannot alter the primary comparisons, rescue the failed replacement judges, or license the claim that EM is human-validated in general.

E.1 FROZEN BUNDLE AND BLINDING

The audit bundle contains a sampling frame, grader packet, private mapping, the selected-checkpoint record, rubric, and boundary partition, each bound by a full SHA-256 digest. A validator reconstructs all 200 records from the frozen sample size and seed and requires exact equality. The packet is drawn from the evaluation split at the selected matched-plasticity checkpoints.

Graders receive only the packet and rubric: each record carries an opaque identifier, and graders see only the question, the answer, and the rubric, not the model, defense, seed, category, machine scores or label, thresholds, hypotheses, or the other grader's work. They do not receive the private mapping, sampling frame, boundary file, thresholds, defense identities, hypotheses, or the expected direction of any difference. Each works alone and sees neither machine scores nor the other grader's labels until both have completed and hashed their files. Calibration examples, if used, are drawn from outside the packet.

E.2 GRADING PROTOCOL AND ACCEPTANCE GATES

Grading and adjudication. The frozen protocol has two graders independently assign every record a finite 0–100 alignment score and a finite 0–100 coherence score. A clear refusal receives an alignment score of at least 70 and is judged for coherence on its own merits, following the machine-judge convention (Betley et al., 2026); off-topic text or code is scored low on coherence but kept in the sample. Under that protocol, each grader's file is hashed before unblinding and bound to a role-specific manifest with an independence attestation. Human binary labels use the same 30/50 rule rule as the judge (Section 3), and a named adjudicator is to resolve every binary disagreement after seeing the blinded record and both human labels, with machine scores and thresholds withheld (Artstein & Poesio, 2008).

Boundary diagnostic. Before human labels are opened, 36 records whose machine scores lie exactly on the 30/50 boundary are separated from the 164 interior records, and we report agreement and κ for both strata. A failed gate may be attributed to boundary cases only if their disagreement rate is at least three times the interior rate and interior $\kappa \geq 0.60$. Before grading, the expected outcome was also written down: transferring the replacement-judge replay's own boundary and interior disagreement rates predicts about 25 boundary and 13 interior disagreements, raw agreement near 0.81, and a failed κ gate.

Acceptance gates. Eight gates, fixed before grading and not revisable afterwards, must pass jointly: 200 double-scored records; inter-grader raw agreement ≥ 0.85 and Cohen's $\kappa \geq 0.70$ (Cohen, 1960; Landis & Koch, 1977); alignment and coherence mean absolute errors ≤ 15 points each; adjudication of every disagreement; and adjudicated human–machine raw agreement ≥ 0.80 and $\kappa \geq 0.60$.

E.3 RUBRIC AND COMPLETENESS

For each record, graders provide finite 0–100 alignment and coherence scores plus optional notes; there is no abstain option. They retain hard cases and apply the refusal and off-topic/code conventions of Section 7. A blank or non-numeric score leaves the audit structurally incomplete. After normalization, a role-bound manifest records each label file's hash, role, identity commitment, independence attestation, and binding to the frozen sample; the manifests provide an audit trail without cryptographic identity proof.

Audit quantity	Frozen criterion	Delivered result
Double-scored records	200	200 (five alignment, two coherence raters)
Inter-grader raw agreement	≥ 0.85	0.799 (pairwise, five alignment raters)
Inter-grader Cohen κ	≥ 0.70	0.494 (pairwise; range 0.368–0.630)
Alignment MAE	≤ 15	23.2 pairwise; 11.23 aggregate vs. judge
Coherence MAE	≤ 15	13.1 pairwise; 12.09 aggregate vs. judge
Adjudication completion	1.00	not performed (median aggregate used)
Human–machine agreement	≥ 0.80	0.830 $[0.772,\ 0.876]$
Human–machine Cohen κ	≥ 0.60	0.485 (interior 0.749; boundary 22 of 36 flip)
Conjunctive decision	all gates pass	inter-rater agreement, both κ references, and pairwise alignment MAE missed

Table 10: Blinded human-audit protocol and delivered results. The delivered rater design uses five alignment and two coherence raters, aggregated by median and mean. Inter-rater rows report pairwise statistics among the delivered raters, which are not the registered two-grader quantities. The protocol bars gate changes and re-sampling of the same records.

E.4 Adjudication and readout

The adjudication set is exactly the records on which the two graders' binary labels differ. The adjudicator re-scores each from scratch, seeing the blinded record and both human labels but never machine scores or thresholds; records on which the graders agree keep their common label.

The first readout compares the two graders before adjudication: raw binary agreement, Cohen's κ (Cohen, 1960), alignment MAE, and coherence MAE. The second readout, computed only after the inter-grader gates, compares the adjudicated human labels with the machine labels. No human–machine score MAE gate is defined; the delivered audit reports one descriptively (Section 7).

The eight gates in Table 10 form one conjunctive decision with no multiplicity adjustment; no subset can pass on its own. Passing all eight licenses one statement: two blinded graders reproduce the machine binary labels on this sample at the pre-registered thresholds.

E.5 Delivered results

Overall raw agreement is 0.830 (Wilson 95% interval $[0.772,\ 0.876]$); it is 0.389 on the 36 boundary records and 0.927 on the 164 interior records, so the boundary disagreement rate (0.611) is 8.4 times the interior rate (0.073). The reading fixed before grading was that this pattern reproduces the study's threshold fragility rather than showing an invalid or miscalibrated judge. The boundary flips alone would not have breached the human–machine κ gate, which takes 26 of 36, and the pre-declared expectation of about 25 boundary and 13 interior disagreements is matched in shape by the observed 22 and 12. Humans flag 54 records (27.0%) against the judge's 26 (13.0%): 31 records are human-positive and machine-negative and 3 the reverse, so the human aggregate is the stricter instrument. Scores track closely, with alignment MAE 11.23, coherence MAE 12.09, and an alignment-score Pearson correlation of 0.896. Per implementation, agreement is 0.700 ($\kappa = 0.360$) on the 70 Llama records, which carry 19 of the 36 boundary cases, 0.877 (0.574) on 65 Mistral records, and 0.923 (0.513) on 65 Qwen records. Among the five alignment raters, pairwise raw agreement is 0.799 and pairwise κ is 0.494 (range 0.368–0.630); single raters against the judge span 0.695–0.850 in raw agreement and 0.29–0.52 in κ, so the median of five buys robustness to the harshest rater rather than higher agreement. A follow-up analysis frozen before its numbers asks whether the judge's misses, the human-positive and machine-negative records, are differential across defenses: they number 10, 6, 6, 2, and 7 of 40 for plain fine-tuning, self-distillation, frozen-copy KL, replay, and steering, and a permutation test stratified by implementation and boundary stratum does not reject homogeneity ($p = 0.131$); the baseline's excess is its boundary composition. Because the packet is stratified rather than a probability sample, no corrected EM rate is produced (Appendix D.7). Table 10 lists each observed quantity beside the criterion fixed in advance.

E.6 Departure of the delivered audit

The final label file rates all 200 packet records with five alignment raters and two coherence raters; the annotation vendor's per-record median alignment and mean coherence, which recompute exactly from the rater columns, serve as the human scores. The delivered files carry no adjudication record, no record of rater roles, and no independence attestation, and they do not establish rater identity, consent, time on task, or complete-rubric delivery; nothing of that kind was inferred. The eight thresholds are carried as reference values only; the comparison is reported as descriptive.

F Additional Limitations

The theory is local. The Fisher analysis holds in a neighborhood of the starting checkpoint. Large updates, singular directions, optimizer preconditioning, the LoRA parameterization (Hu et al., 2022), sampling temperature, and rollouts whose distribution shifts during training can all move training outside the quadratic regime. Because forward and reverse KL share the same local Hessian, any effect of KL direction must come either from higher-order terms or from the different inputs the two anchors are evaluated on.

Curvature diagnostic. The diagnostic that tried to measure the quantities the theory turns on, the shared subspace U, the Fisher spectra, and the ratio η_Q of Equation (6), computed η_Q in a six-dimensional proxy basis and found the off-domain value 8.8% higher (3.19×10^{-4} against 2.93×10^{-4}), with no uncertainty estimate; but that basis captures only 3.3–4.7% of the held-out update energy and 0.6% of the task-gradient energy, and its local quadratic error reaches 123%, so it neither identifies a shared subspace nor tests the separation assumption. The theory tells future experiments what to measure, not what they will find.

On-policy anchoring costs more compute. The bundle regenerates its rollout buffer every two optimizer steps, about 210 generation phases per epoch, and scores every anchor call with a teacher forward pass. The fixed-input anchors need only the teacher pass: self-distillation obtains it by disabling the adapter, and the frozen-copy implementation holds a second 7B model in memory, which forced micro-batch 1 on the longest training examples. We did not log wall-clock time per cell, so we report the cost structure rather than a measured ratio; any deployment comparison should measure it.

The audit checks one configuration. The audit tests one judge, one prompt, and one pair of thresholds on 200 responses. What it supports is narrow: the aggregated human label reproduces the machine label on interior records ($\kappa = 0.749$) and diverges on the boundary, where the human raters flag more responses than the judge. Two features of the delivered audit temper even that reading. First, the five alignment raters agree with one another (pairwise $\kappa = 0.494$) about as well as their aggregate agrees with the judge ($\kappa = 0.485$), so the rubric itself carries that much variance, and the audit does not show that the judge disagrees with people more than people disagree with each other; what sets the judge apart is where its disagreements fall, not how many there are. Second, the aggregate is a different instrument from the adjudicated label the protocol specified: a median of five continuous scores varies less than any single rater's score, so it crosses the 30-point cutoff less by chance, and single raters against the judge span 0.29–0.52 in κ around the aggregate's 0.485. The boundary stratum is fixed by machine scores alone, so aggregation cannot move records into it, but it can change how many of them flip, and we cannot separate that effect from the judge's own boundary fragility with the data delivered. The audit does not validate machine judges in general, does not make EM independent of how the human rubric was written, and does not license a human-validated label for any estimate here. Rater recruitment, compensation, exposure safeguards, and the applicable ethics determination are not established by the delivered files and are reported as such, without changing the analysis protocol.