FIGHTING THE DRIFT: PROBE TRANSFER ACROSS POST-TRAINING STAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear activation probes are used for monitoring large language models. Updating a probe for a new checkpoint requires re-extracting activations and may require new labels. First, we study probe transfer across SFT, DPO, and RL in OLMo-3 7B and across SFT and GRPO in Qwen3-4B. On the most disruptive SFT transfers, frozen probes achieve only about half the true positive rate of a probe refit on the target checkpoint, at 1% FPR. DPO and RL updates leave much less headroom for refitting than SFT updates. A refit on the target checkpoint can outperform the probe on its source checkpoint, so we measure transfer performance against the target refit. Label-free representation alignment can close much of this headroom but can plateau below the refit or perform worse than the frozen probe. Second, we develop an orthogonal drift correction that adjusts the old probe, with or without representation alignment, using a small number of labels. With a strong starting probe, correction achieves the best performance at most tested labeling budgets, can match the full refit, and can outperform a from-scratch refit trained on the same labeled data. Cheap probe transfer is a step toward overseeing systems that update faster than their monitors can keep up.

1 INTRODUCTION

Linear probes on internal activations are an increasingly common monitoring layer for deployed language models: they add little latency, can detect information not apparent in generated text, and have been used to detect deception, refusal, and harm (Alain & Bengio, 2017; Zou et al., 2023; Arditi et al., 2024; Goldowsky-Dill et al., 2025; Cunningham et al., 2026).

A probe fit on one checkpoint may perform worse than a refit after the model is updated. Fine-tuning can change safety-relevant behavior (Qi et al., 2024; Yang et al., 2023). Retraining the probe is limited by the cost of generating new responses, labeling them and, in some cases, extracting terabytes of per-token activations. We ask when model updates create headroom for retraining and how far cheaper interventions can close it. Our contributions:

- We measure frozen probe performance relative to the refit ceiling across SFT, DPO, and RL in OLMo-3, and across SFT and GRPO in Qwen3-4B. We compare LLM layers and study harm and refusal.
- We evaluate label-free representation alignment after full post-training stages, beyond the adapter-scale updates tested in prior work. We identify where it closes most of the headroom, where it plateaus, and where it degrades performance.
- We develop a labeled correction that can be applied to any linear probe, with or without representation alignment, and compare it against retraining from scratch across labeling budgets. With a strong starting probe, correction helps achieve the best performance at most tested budgets.

Summary of findings. On the most disruptive OLMo-3 SFT transfers, frozen probes achieve about half the target refit's true positive rate at 1% FPR; DPO and RL leave much less headroom. Label-free representation alignment depends on the concept and checkpoint pair: the ridge map, with labels used only for tuning, closes about 87% of harm AUROC headroom, but can perform

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.

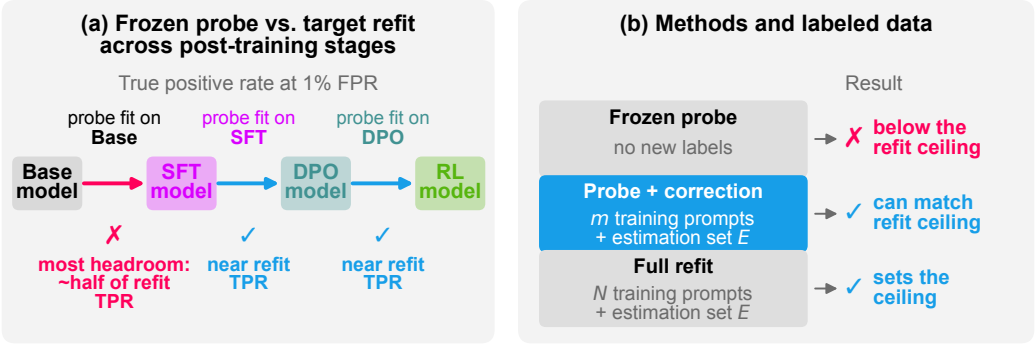

Figure 1: (a) On the most disruptive SFT transfers, frozen probes achieve about half the target refit's true positive rate at 1% FPR; DPO and RL leave much less headroom. (b) Correction updates an existing probe using labeled exchanges from m training prompts and a separate estimation set E. It can match the full refit or outperform a from-scratch refit using the same labeled data. Correction uses labeled exchanges from $m + |E|$ prompts in total.

below the frozen probe on refusal. Correction on the frozen probe has higher AUROC than the from-scratch refit at most tested refusal budgets. On harm, correction on the ridge-mapped probe can reach the full refit.

2 Related Work

Activation probes as safety monitors. Probes on internal activations (Alain & Bengio, 2017; Belinkov, 2022) have been used to study factuality (Azaria & Mitchell, 2023), latent knowledge (Burns et al., 2023), refusal (Zou et al., 2023; Arditi et al., 2024), deception (Goldowsky-Dill et al., 2025), and behavior triggered by backdoors (MacDiarmid et al., 2024). Their performance can change after a model update (Duan, 2026).

Post-training and probe fragility. Fine-tuning can compromise safety behavior (Qi et al., 2024; Yang et al., 2023). Open post-training pipelines release intermediate checkpoints (Lambert et al., 2025; OLMo Team, 2025; Team Olmo et al., 2025).

Probe staleness and maintenance. Duan (2026) finds that quantization largely preserves frozen-probe performance, while adapter fine-tuning often reduces it, and label-free representation alignment restores most of the lost performance. Their staleness measure compares the frozen probe before and after the model update. We study full SFT, DPO, and RL stages and evaluate labeled correction alongside label-free maps against a target-checkpoint refit. Gaya et al. (2026) study evaluation-awareness probes under adapter RLVR. They find little change in aggregate performance, but decision flips concentrate on borderline benchmark-style prompts, and refitting the probe gives little improvement.

Representation alignment and score combination. SVCCA compares neural representations in a shared subspace (Raghu et al., 2017); relative representations enable transfer without fitting a map (Moschella et al., 2023). FacLens transfers factuality probes using labeled source data and unlabeled target data (Wang et al., 2025). We compare ridge, Procrustes, and CCA representation alignment and use the frozen and ridge-mapped probes as correction priors.

3 Experimental Setting

Lineage and checkpoints. We use OLMo-3 7B checkpoints (Team Olmo et al., 2025) spanning its Think and Instruct post-training branches. Instruct-SFT is trained from Think-SFT, so that checkpoint pair measures a further SFT stage. We further fine-tune the final Instruct checkpoint on clinical,

code, and science data. We repeat the budget experiments on Qwen3-4B (Yang et al., 2025) checkpoints produced by SFT on SmolTalk (Ben Allal et al., 2025) followed by GRPO on OpenR1-Math (Hugging Face, 2025). Both training mixtures are filtered to remove overlap with our probe datasets. We write $a \to t$ for a *checkpoint pair*: the probe is fit on checkpoint a and deployed on checkpoint t.

Probes and data. A probe is a linear classifier on internal activations at a single layer. We study two label types. The *harm* label asks whether harm appears anywhere in the exchange. The *refusal* label asks whether checkpoint t refuses a given prompt, so refusal labels must be assigned for each checkpoint.

Off-policy versus on-policy labels. Our harm label is *off-policy*: the exchange text is fixed, so labels carry over across checkpoints. Our refusal label is *on-policy*: it depends on the checkpoint's own response, so new labels require generating and judging new responses. Representation alignment uses labels only for tuning and can reduce the labeling cost. Both tasks still require target-checkpoint activations. Re-running the full training set is expensive, especially for token-level probes (Cunningham et al., 2026). Caching token activations across LLM layers can require terabytes per checkpoint, so we compare methods that use activations from only m prompts.

Metrics and references. We draw m prompt groups from a training pool of N prompt groups at the target checkpoint. Each prompt group contains all exchanges associated with that prompt. Methods that fit on labels use the labels of every exchange in those m prompts, so m counts prompts, not response labels. Correction, maps and refits also use a separate labeled estimation set E, disjoint from training and test data. It contains $4{,}000$ prompts, except in the intermediate-SFT experiments, where it contains 800. Correction uses E for within-class score orthogonalization and to choose the blend weight. The refit and ridge map use E to choose their regularization. We write F_a for the probe's performance at its own source checkpoint, F_t for the same probe's performance at the new checkpoint, and U_t for the refit ceiling, the performance of a fresh probe trained on all N prompts at the new checkpoint. We report AUROC, standardized pAUC at $\leq 1\%$ and $\leq 5\%$ FPR, and TPR at 1% FPR. Writing R_t for the transferred probe's performance, results are expressed as the *share of headroom closed*, $(R_t - F_t)/(U_t - F_t)$. For budget-share plots, we compute this share from seed-averaged performances and estimate its standard error by bootstrapping seeds. The refit headroom is

$$\text{refit headroom} \;=\; U_t - F_t, \tag{1}$$

and it splits into two parts:

$$\underbrace{U_t - F_t}_{\text{refit headroom}} \;=\; \underbrace{(F_a - F_t)}_{\text{frozen loss}} \;+\; \underbrace{(U_t - F_a)}_{\text{ceiling drift}}. \tag{2}$$

Frozen loss measures the change in the old probe's performance across checkpoints; ceiling drift measures the change from the source probe's performance at the source checkpoint to the target checkpoint's refitted probe performance. Comparing only with the old probe's source performance (Duan, 2026) omits ceiling drift. We therefore evaluate adaptation against the target checkpoint's refit ceiling.

4 Refit headroom across post-training stages

Figure 2 shows frozen-probe transfer across thirteen OLMo-3 checkpoint pairs at layer 16. On AUROC the frozen probe achieves 93 to 100% of the refit ceiling on every checkpoint pair. On the most disruptive SFT transfers, frozen probes achieve only about half the true positive rate of a probe refit on the target checkpoint, at 1% FPR: base→Think-SFT achieves 48.0% and Think-SFT→Instruct-SFT 51.9% of the refit's TPR, while AUROC reaches 96.3% and 93.3% of the respective ceilings. The frozen probes can be close to the refit in AUROC but far below it in TPR at 1% FPR.

Most headroom comes from SFT. We measure refit headroom ($U_t - F_t$) at every layer for each stage (Figure 3). SFT produces positive headroom at every layer; DPO and RL have headroom close to the unchanged-weight control at layer 16, though the Instruct-branch DPO checkpoint pair shows modest headroom at layer 31 ($+0.0037$). A probe fit after SFT needs little adjustment for subsequent DPO or RL on these pipelines. Qwen3-4B shows the same contrast between SFT and subsequent RL (Appendix Table 4).

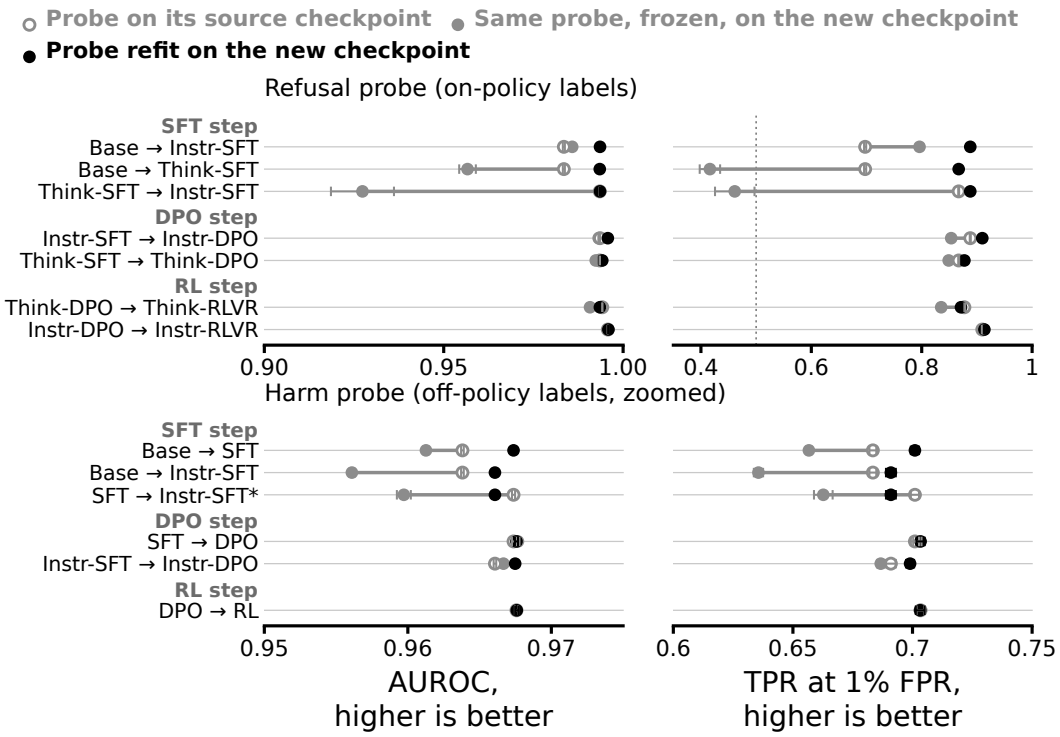

Figure 2: **On the most disruptive SFT transfers, frozen probes achieve about half the refit's true positive rate at 1% FPR, despite reaching at least 93% of its AUROC.** DPO and RL leave much less headroom. Left: AUROC; right: TPR at 1% FPR. Each row is one OLMo-3 checkpoint pair at layer 16; the harm panels are zoomed to their narrower ranges. Whiskers are ±1 SE over 8 seeds.

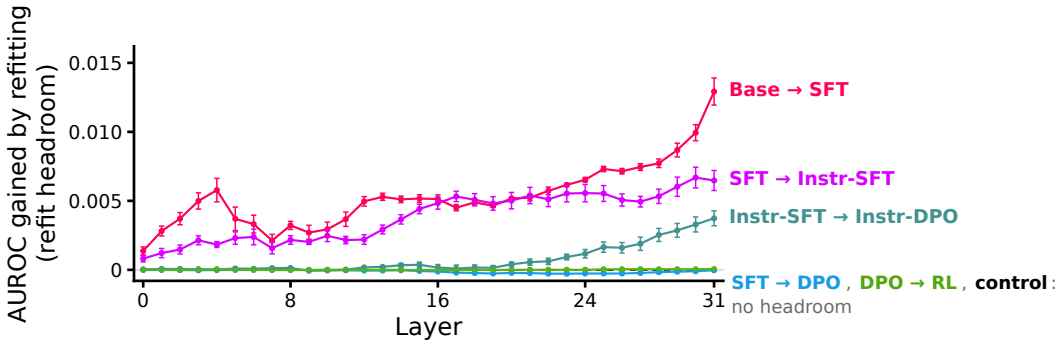

Figure 3: **SFT creates the refit headroom at every layer. DPO and RL add much less.** AUROC gained by refitting the harm probe on the new checkpoint instead of freezing it, at each layer of OLMo-3 7B (dots: mean over 5 seeds at each layer; whiskers: ±1 SE). The control recomputes activations without changing the model weights. Values in Appendix Table 5.

5 Label-free activation mapping

The map. Given activations from the same prompts at both checkpoints, a ridge regression maps target activations into the source checkpoint's activation space. The frozen probe then scores the mapped activations, with no refit at the target. Our map is

$$\mathcal{N}_\lambda = \arg\min_{\mathcal{N}} \|\mathbf{Z}_t\mathcal{N} - \mathbf{Z}_a\|_F^2 + \lambda\|\mathcal{N}\|_F^2 = \left(\mathbf{Z}_t^\top \mathbf{Z}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{Z}_t^\top \mathbf{Z}_a, \tag{3}$$

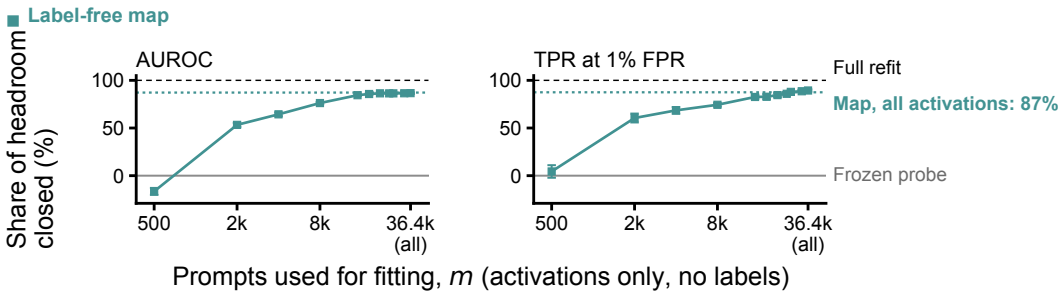

Figure 4: **The label-free map closes most of the headroom without labels for fitting, stops short of the full refit, and is below the frozen probe on AUROC at the smallest budget.** Harm probe, OLMo-3 base→DPO (N=36,400, 8 seeds): share of the headroom closed by the ridge map fit on m new prompts, on AUROC (left) and TPR at 1% FPR (right). The dotted line is the same map fit on all N activations.

which yields a single scoring direction $\tilde{\mathbf{w}} = \mathcal{N}_\lambda \mathbf{w}_a$ in target space. Fitting the map uses no labels. For the map-only baselines, we choose λ for each seed using a small, fixed labeled estimation set E. It is fit either on the m correction prompts alone (the realistic setting for off-policy labels) or on all of the probe's training set (useful for on-policy labels when labeling is the dominant cost; §3). Other representation alignment methods (rotations, subspace, low-rank) are compared in Appendix A. The ridge map is generally the strongest label-free representation alignment method on harm; on refusal, Procrustes and the frozen probe can each outperform it on specific checkpoint pairs. For simplicity we focus on the ridge map and the frozen probe as label-free transfer baselines throughout.

Map performance. When fit on all available activations, the map closes about 87% of the AUROC headroom on harm base→dpo, using labels only for tuning (Figure 4), and 28%, 49%, and 61% of the AUROC headroom on refusal base→Think-SFT, Think-SFT→Think-DPO, and Think-DPO→Think-RLVR, respectively. On OLMo harm base→DPO, the map remains below the full refit on both metrics. On Qwen3-4B harm base→SFT and base→RL, it can exceed the full refit in TPR (Appendix Figure 26). The map can also degrade performance. When fit on only m prompts, it overfits at small m: at m/N=1.4% it falls below the frozen probe (-16% on base→dpo, -136 to -441% of the TPR headroom on the domain-SFT checkpoint pairs). Even with all activations, the map by itself performs below the frozen probe on some refusal checkpoint pairs (-0.037 AUROC on base→Instruct-SFT). We next test whether labeled correction closes the remaining headroom (§6).

6 Orthogonal drift correction

Setup. Correction uses a small labeled sample to improve an existing linear probe, whether frozen or obtained through representation alignment. It takes any prior direction $\mathbf{w}_0 \in \mathbb{R}^d$ and fits one correction direction $\hat{\mathbf{g}} \in \mathbb{R}^d$ by regularized logistic regression on the labeled exchanges from m training prompts. The corrected probe is confined to their span,

$$\mathbf{w} \;=\; b_0\,\mathbf{w}_0 + b_1\,\hat{\mathbf{g}}, \qquad s(x) \;=\; b_0\,u(x) + b_1\,v(x), \quad u = \mathbf{z}^\top \mathbf{w}_0,\; v = \mathbf{z}^\top \hat{\mathbf{g}}, \tag{4}$$

so the final blend uses two scalar weights regardless of the activation dimension d. A from-scratch refit can choose any direction in $\mathbb{R}^d$, while correction restricts the final probe to the span of the prior and fitted correction direction. We compare their performance at small label budgets in §6.1. Correction accepts any linear probe expressed in the target activation space as its prior: the frozen probe, a probe obtained through ridge or Procrustes representation alignment, or another linear detector. For simplicity, we examine the same two priors on both concepts: the frozen probe and the ridge-mapped probe (§5). We use the same correction procedure with each prior. We do not evaluate Procrustes followed by correction.

Orthogonalization and search. We orthogonalize the correction scores against the prior scores after centering both within each class on E. Let $\mathbf{p} = \mathbf{w}_0/\|\mathbf{w}_0\|$ and define the scores on E as

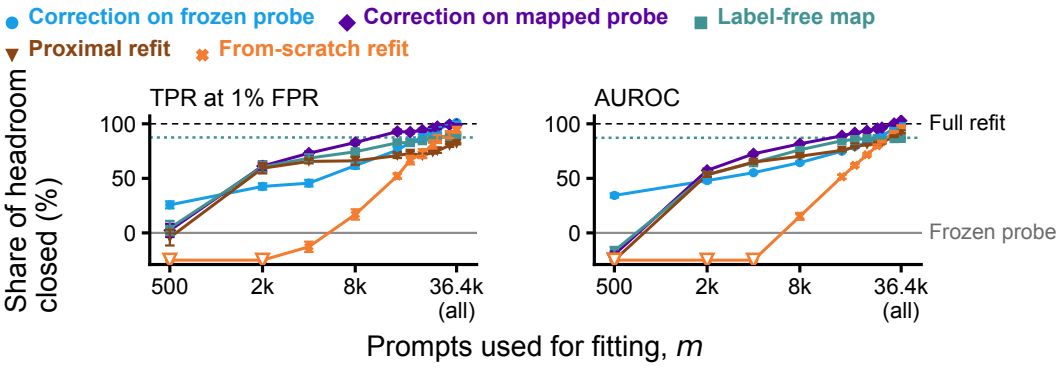

Figure 5: **Correction leads the methods using the same m prompts: correction on the frozen probe is strongest at the smallest budget, while correction on the ridge-mapped probe performs best at most larger budgets.** Harm probe, OLMo-3 base→DPO (N=36,400, 8 seeds, ± 1 SE): share of headroom closed against the number of training prompts m processed at the target checkpoint, on TPR at 1% FPR (left) and AUROC (right). 0% is the frozen probe and 100% a refit on all N prompts; the dotted line is the label-free map fit on all N activations; points below -25% are drawn at the bottom edge as open triangles. Values in Appendix Tables 6 to 7.

$\mathbf{u} = \mathbf{Z}\mathbf{p}$ and $\mathbf{v} = \mathbf{Z}\hat{\mathbf{g}}$. Writing $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ for the class-centered scores, we compute $\beta = \tilde{\mathbf{u}}^\top \tilde{\mathbf{v}} / \tilde{\mathbf{u}}^\top \tilde{\mathbf{u}}$ and set $\hat{\mathbf{g}}_\perp = (\hat{\mathbf{g}} - \beta \mathbf{p}) / \|\hat{\mathbf{g}} - \beta \mathbf{p}\|$. Orthogonality here is between the class-centered scores, rather than the weight vectors under the Euclidean inner product. The corrected probe is then a one-parameter blend,

$$\mathbf{w}(\alpha) = \mathbf{p} + \alpha\, \hat{\mathbf{g}}_\perp, \qquad \alpha^* = \underset{\alpha \in [0,100]}{\arg\max}\ \mathcal{M}(y;\, \mathbf{u} + \alpha\, \mathbf{Z}\hat{\mathbf{g}}_\perp)\,, \tag{5}$$

approximately optimized by bounded Brent search (Brent, 1973). We choose α using AUROC on the estimation set E in all reported experiments and deploy the selected blend directly. Once the correction direction is fitted, the final blend has one free parameter.

We compare the correction and map against two baselines that use the same m labeled prompts: a *from-scratch refit*, which trains a new probe with no information from the source checkpoint, and a *proximal refit*, which runs gradient descent in $\mathbb{R}^d$ initialized at the prior and regularized toward it.

6.1 Off-policy label: harm

Figure 5 compares methods on harm base→dpo. The correction on the frozen prior closes a third of the AUROC headroom at the smallest budget (m/N=1.4%) and consistently improves AUROC and TPR. At each tested budget, one of the two corrected probes leads the methods using the same m prompts. Correction on the frozen prior is strongest at the smallest budget; correction on the ridge-mapped probe performs best at most larger budgets and can reach the full refit ceiling. Correction can improve performance beyond mapping alone.

6.2 On-policy label: refusal

Figure 6 compares the methods on six refusal checkpoint pairs (TPR in Appendix Figure 19). Correction on the frozen prior consistently improves TPR and has higher AUROC than the from-scratch refit at most tested label budgets. A mapped prior can further improve performance on some checkpoint pairs (Appendix Figure 20), while the frozen prior is more effective where ridge mapping performs poorly. The performance of the correction method therefore depends on the prior as well as the label budget.

Decoupled budgets. When labeling is the dominant cost, we can fit the map on activations from all N prompts while fitting the correction on labeled exchanges from only m training prompts. The map needs paired activations for the same prompts at both checkpoints. Generating responses and

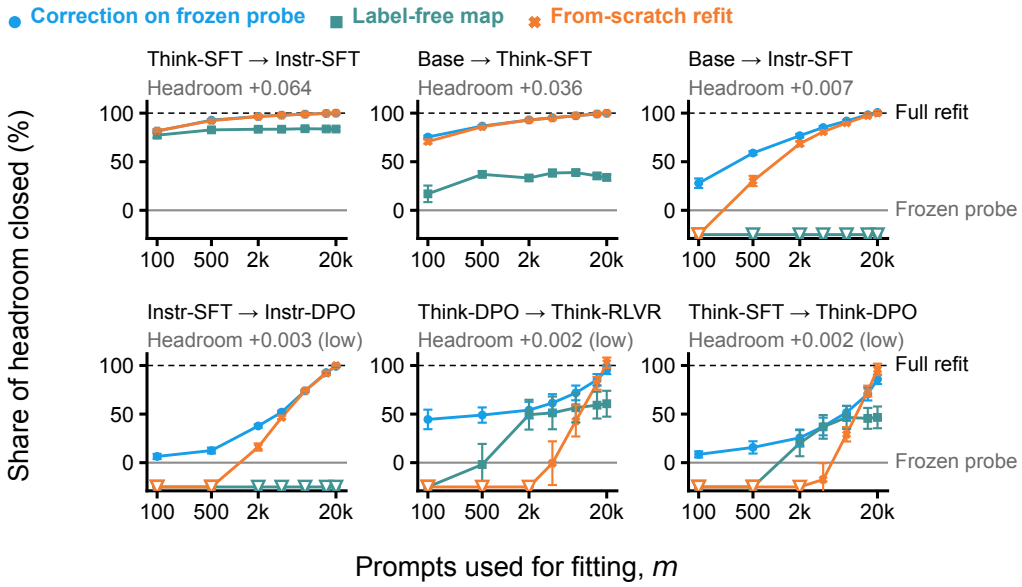

Figure 6: **On refusal, correction on the frozen probe has higher AUROC than the from-scratch refit at most tested label budgets, with the largest advantage at small budgets.** Six OLMo-3 checkpoint pairs at layer 16, ordered by how much a refit would gain (gray subtitles; N=20,000, 8 seeds, ± 1 SE): share of AUROC headroom closed against the number of new labeled prompts m, for the correction on the frozen probe, the label-free map fit on the activations of those m prompts, and the from-scratch refit. Points below -25% sit at the bottom edge as open triangles. TPR in Appendix Figure 19.

extracting activations still have a cost; the saving is in the number of responses that need labels. Figure 7 shows this decoupled budget on six refusal checkpoint pairs. At most tested budgets, one of the two corrected probes has the highest AUROC. Which correction performs better, on the frozen prior or on the ridge-mapped prior, depends on the checkpoint pair. Correction can approach the refit ceiling using fewer labeled training prompts.

Choosing the prior. Correction with a strong prior often gives the best performance at a given labeling budget. We study the same two starting probes across both concepts: a ridge-mapped probe is generally strong on harm, while the frozen probe is competitive on refusal. Other label-free methods, including Procrustes, can also supply the starting probe. Candidate probes can be corrected and compared on E. We evaluate correction only on the frozen and ridge-mapped probes; standalone Procrustes results are in Appendix A.

6.3 Generalization across domains, layers, and model families

Domain SFT. We perform additional SFT to specialize the final Instruct checkpoint for clinical, code, and science domains (Figure 8) and test our probe transfer and adaptation methods on these checkpoints. Across all three domain SFT updates, correction achieves the highest AUROC at most tested label budgets.

Layer robustness. Figure 9 shows the method ordering at every layer on harm base→dpo at two budgets. The from-scratch refit has the lowest AUROC at every layer. Correction on the ridge-mapped probe has the highest AUROC at most layers with 8,000 training prompts; with 2,000, the label-free map is best at most layers.

Qwen3-4B replication. On Qwen3-4B, correction improves on its priors and can outperform refits trained on the same labeled data (harm: Appendix C.11; refusal: Appendix C.8).

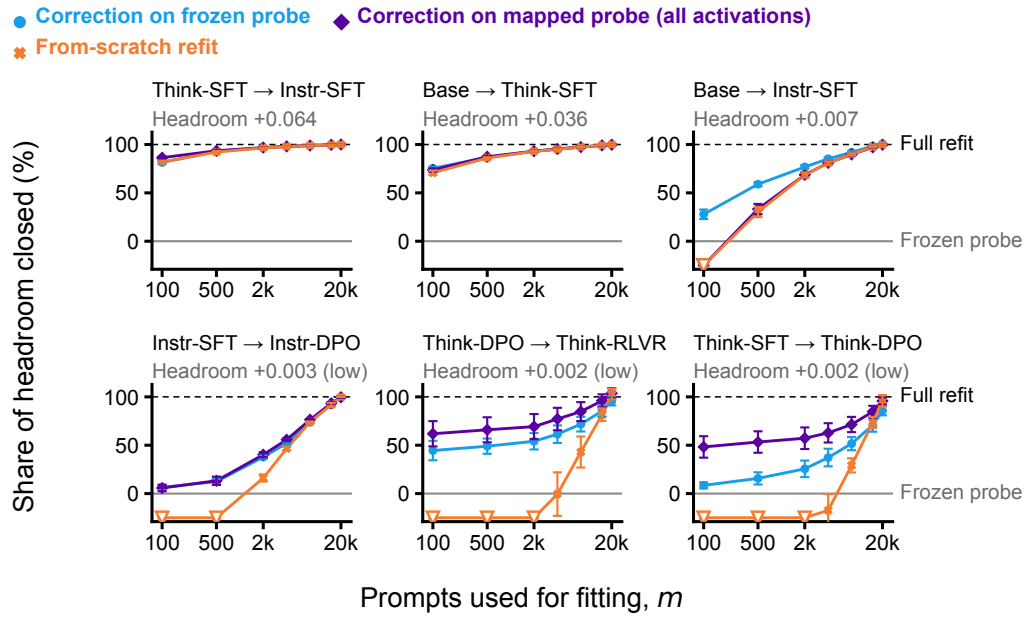

Figure 7: **With activations for all N prompts but labeled exchanges from only m training prompts, one of the two corrected probes has the highest AUROC at most tested budgets. Which correction performs better, on the frozen prior or on the ridge-mapped prior, depends on the checkpoint pair.** Six OLMo-3 refusal checkpoint pairs (8 seeds, ± 1 SE): the map is fit on all N activations (when labeling is the dominant cost); the correction and the from-scratch refit use labeled exchanges from only m training prompts; the correction on the frozen probe is repeated from Figure 6 for reference. Conventions as in Figure 6.

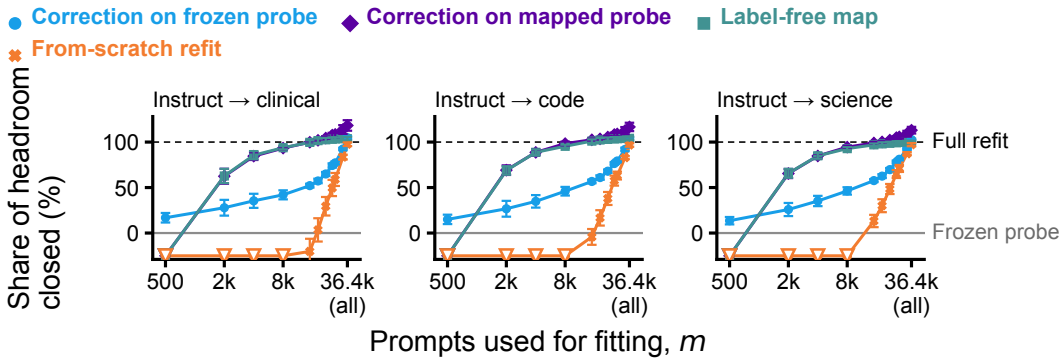

Figure 8: **Across all three domain SFT updates, correction achieves the highest AUROC at most tested label budgets.** Harm probe, Instruct→clinical, code and science SFT (N=36,400, 8 seeds, ± 1 SE): share of the AUROC headroom closed against new prompts m; conventions as in Figure 5. Values above 100% mean the method outperforms the full refit, which need not be the best possible probe. Raw AUROC in Appendix Figure 24.

7 Discussion

For these post-training pipelines, fitting the probe after SFT leaves little need for later adjustment. When an update is needed, candidate priors can be corrected and compared on E. Choosing the training budget from estimated headroom is a direction for future work.

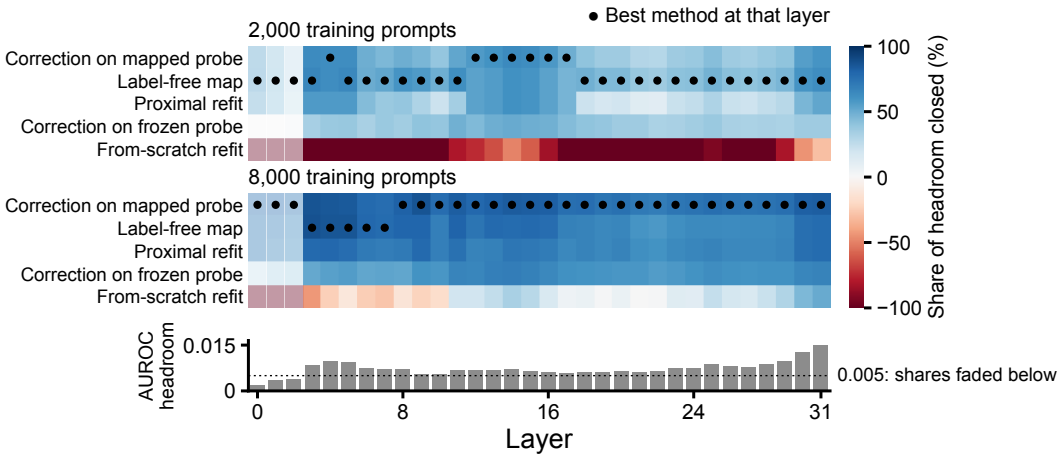

Figure 9: **The from-scratch refit has the lowest AUROC at every layer. Correction on the ridge-mapped probe is best at most layers with 8,000 training prompts, and the label-free map with 2,000.** Share of the AUROC headroom closed on harm base→DPO, per layer and method, with 2,000 (top) and 8,000 (bottom) training prompts (8 seeds). A dot marks the best method at each layer. Colour is clipped at $\pm 100\%$: with 2,000 training prompts the refit falls to $-1{,}945\%$ at layer 0, that is, far below the frozen probe. The strip below gives this checkpoint pair's refit headroom per layer, the denominator of every share; the three faded layers are below 0.005 AUROC, too little to measure shares reliably. The same data as line plots: Appendix Figure 25.

8 Limitations

We test on two model families (OLMo-3 7B and Qwen3-4B) and two probe concepts (exchange-level harm and refusal). We evaluate realistic post-training pipelines using standard SFT, DPO, and RL objectives. We do not test adversarial updates designed to evade the probe, such as RL with a penalty on probe scores. Such updates could reduce even the performance of a fully refitted probe. Whether probes remain maintainable under such threat models is an open question.

9 Conclusion

SFT creates most of the refit headroom in our experiments, while later DPO and RL stages add little. Label-free representation alignment can close much of this headroom, but the best method depends on the checkpoint pair. Correction can update either the frozen probe or a probe obtained through representation alignment. With a strong starting probe, it gives the best performance at most tested labeling budgets and can match the full refit. These methods offer cheaper ways to keep probes current as models change.

Reproducibility statement

We use publicly available OLMo-3 7B (Team Olmo et al., 2025) and Qwen3-4B (Yang et al., 2025) starting checkpoints. We train the additional domain-SFT and Qwen post-training checkpoints described in Section 3. Experimental procedures and evaluation metrics are described in Sections 3, 5, and 6.

Ethics statement

This work is defensive: we study how to update safety probes after model changes so they can continue to support oversight. We measure their performance against refits on the target checkpoint. The harm and refusal datasets are derived from existing corpora and model-generated text; we release no new personal data.

AI ASSISTANCE DISCLOSURE

We used AI assistance substantially in this work, in the categories below. Anthropic Claude models were used throughout, both as a coding and writing assistant and as a component of the data pipeline itself.

Research ideation and execution. LLM coding assistants were used throughout the implementation: the data pipeline, the code for probe fitting and transfer methods, experiment orchestration, and figure-generation pipeline. LLMs were also used to flesh out research ideas proposed by the authors. All experimental design decisions were made by the authors.

Generating synthetic datasets. We use model-generated response texts and LLM-generated labels. For refusal, each checkpoint generates its own responses to the prompt pool. For harm, the exchange texts are generated in advance and held fixed across checkpoints. Refusal and compliance labels are produced by LLM judges (Claude Sonnet 5, with Claude Haiku 4.5 as fallback). The source prompts for refusal and exchanges for harm are drawn from existing corpora, some of which may themselves contain AI-generated data.

Retrieval and discovery. We used AI tools to find related work. The authors checked every resulting citation against the primary source.

Drafting and writing. AI assistance was used to draft sections of this paper and to restructure and copy-edit prose. All claims, numbers and citations were verified by the authors against the underlying artifacts.

Responsibility. The authors take full responsibility for the entire contents of this paper, including all text, code and results produced with AI assistance, and for the correctness of all claims.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL https://aclanthology.org/2022.cl-1.7/.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Clémentine Fourrier, Hugo Larcher, Ben Burtenshaw, Haojun Zhao, Caleb Fahlgren, Mathieu Morlon, Agustín Piqueres Lajarín, Xuan-Son Nguyen, Cyril Zakka, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a fully open small language model. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H.

Richard P Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall, 1973. URL https://maths-people.anu.edu.au/brent/pub/pub011.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.03827.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Andy Dau, Alek Dimitriev, Logan Howard, Yijin Hua, Rob Gilson, Mu Lin, Christopher Liu, Vladimir Mikulik, Rohit Mittapalli, Clare O'Hara, Jin Pan, Nikhil Saxena, Alex Silverstein, Yue Song, Giulio Zhou, Jan Leike, Jared Kaplan, Ethan Perez, and Mrinank Sharma. Constitutional classifiers++: Efficient production-grade defenses against universal jailbreaks. In *International Conference on Learning Representations*, pp. 48038–48054, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4f697305ef1f868ad77c3c0027989a6f-Abstract-Conference.html.

Evan Duan. Do activation monitors survive model updates? Benchmarking, predicting, and repairing activation-monitor staleness. *arXiv preprint arXiv:2606.15980*, 2026. URL https://arxiv.org/abs/2606.15980.

Sana Gaya, Imran Nasrullah, Cade Justad-Sandberg, and Girish Gupta. When is a probe monitor safe to act on? A transferability stress test for eval-awareness probes under RLVR. In *Workshop on Actionable Interpretability (AIW) at COLM 2026*, 2026. URL https://openreview.net/forum?id=sQHKfxuDbP.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19755–19786. PMLR, 2025. URL https://proceedings.mlr.press/v267/goldowsky-dill25a.html.

Hugging Face. Open R1: A fully open reproduction of DeepSeek-R1. https://github.com/huggingface/open-r1, 2025. URL https://github.com/huggingface/open-r1#citation.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic Alignment Science, April 2024. URL https://www.anthropic.com/research/probes-catch-sleeper-agents.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

OLMo Team. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/83b7da3ed13f06c13ce82235c8eedf35-Abstract-Conference.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Yanling Wang, Haoyang Li, Hao Zou, Jing Zhang, Xinlei He, Qi Li, and Ke Xu. FacLens: Transferable probe for foreseeing non-factuality in fact-seeking question answering of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18563–18582. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.937. URL https://aclanthology.org/2025.emnlp-main.937/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023. URL https://arxiv.org/abs/2310.02949.

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

A LABEL-FREE REPRESENTATION ALIGNMENT

Figure 10 compares three label-free representation alignment methods: the ridge map, Procrustes rotation, and CCA, with fixed whitening regularization or with whitening regularization tuned on E. On harm base→dpo at $m = N$, the ridge map closes 87% of the AUROC headroom versus 45% for Procrustes. CCA closes 73% with fixed whitening regularization and 77% with whitening regularization tuned on E. On refusal, relative performance depends on the checkpoint pair. On base→Instr-SFT (low headroom), Procrustes has the highest AUROC among these representation alignment methods, although all remain below the frozen probe. On base→Think-SFT (high headroom) at $m = N$, Procrustes closes 56% of the AUROC headroom and 37% of the TPR headroom, outperforming the ridge map (35% AUROC, 21% TPR) and the frozen probe. The best label-free method depends on the checkpoint pair. For simplicity we focus on the ridge map and the frozen probe as the two label-free transfer baselines in the main text; the correction (§6) can take any of these as its prior, including Procrustes.

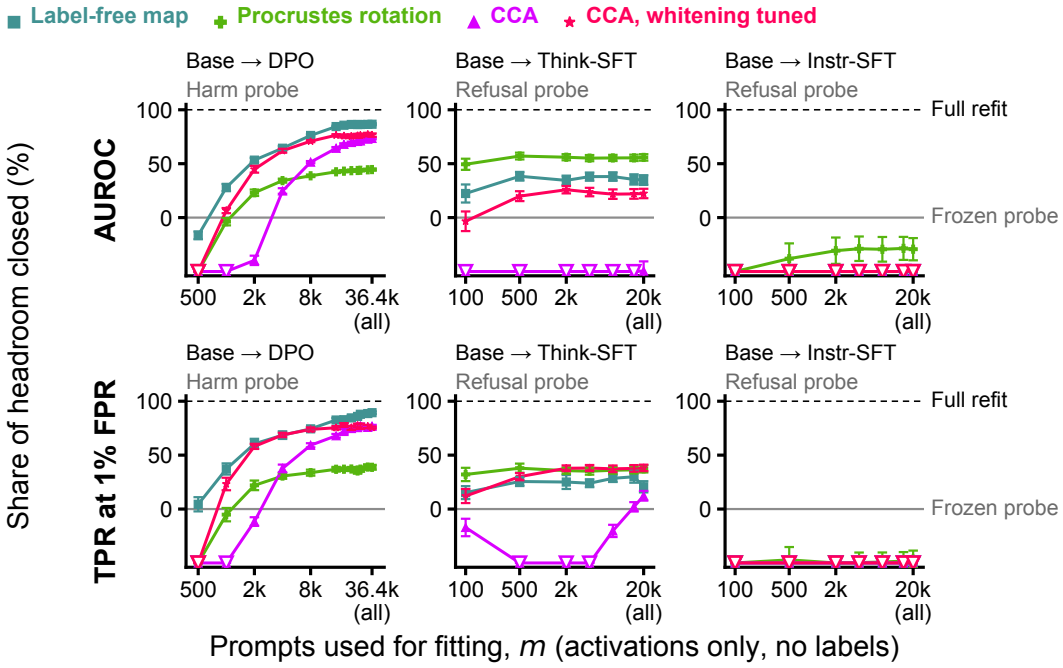

Figure 10: **The relative performance of label-free representation alignment methods depends on the checkpoint pair. Ridge is generally strongest on harm, Procrustes is competitive on refusal base→Think-SFT, and the frozen probe leads the representation alignment methods on refusal base→Instr-SFT.** Share of headroom closed (AUROC top, TPR at 1% FPR bottom) against the number of new prompts whose activations each method reads; 8 seeds, ±1 SE. Points below −50% sit at the bottom edge as open triangles.

B ADDITIONAL EXPERIMENTS

Source checkpoint. Table 1 compares probes fit at the base checkpoint and after SFT. When the probe is fit after SFT (sft→dpo), the refit headroom is −0.0001 (effectively zero), leaving essentially no advantage for refitting over the frozen probe. Fitting the probe after SFT therefore leaves little need for adjustment on this checkpoint pair.

Table 1: Source checkpoint: base→DPO versus post-SFT→DPO. The map uses activations from all N training prompts; correction on the frozen probe uses m=8,000 training prompts. Fitting the probe after SFT leaves effectively zero refit headroom on this DPO checkpoint pair.

Source → target	Headroom	Map share (%)	Correction share (%)	Frozen
base → dpo	+0.0059	87.2 ± 1.0	64.4 ± 1.0	0.9636
sft → dpo	−0.0001	null	null	0.9696

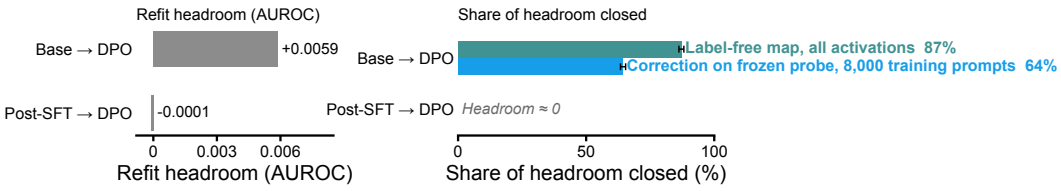

Figure 11: **Fitting the probe after SFT leaves effectively zero refit headroom on this DPO checkpoint pair.** Left: refit headroom in AUROC for the harm probe on base→DPO versus post-SFT→DPO. Right: the share of headroom closed by the label-free map (all activations) and the correction on the frozen probe (8,000 labeled prompts); for the probe fit after SFT, headroom is effectively zero, so we do not report a share. Whiskers: ±1 SE over 8 seeds.

Intermediate SFT checkpoints. Figure 12 shows the map and refit on two intermediate SFT checkpoints, at steps 15,000 and 30,000. The map fit on all N prompts closes 87–90% of the AUROC headroom. At small budgets, the from-scratch refit has lower AUROC than the frozen probe. Its mean AUROC remains below the map at every tested budget, including the largest, m = 24,000.

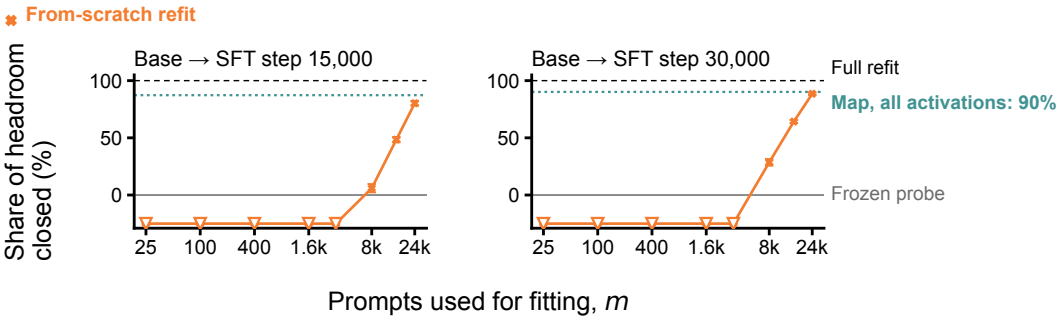

Figure 12: **At the two intermediate SFT checkpoints, the map fit on all N prompts closes 87–90% of the AUROC headroom. The from-scratch refit has lower mean AUROC than the map at every tested budget.** Harm probe, base→SFT at steps 15,000 and 30,000 (N=30,000, $|E|$ = 800, 8 seeds, ±1 SE): share of the AUROC headroom closed against new prompts m. Points below −25% sit at the bottom edge as open triangles.

Budget sizing $m^*(\tau)$. Table 2 gives the first tested training budget m whose mean AUROC closes at least τ% of the headroom. These are crossings of the measured curves; we do not interpolate between tested budgets. The table excludes the fixed estimation set E; Figure 15 includes it and shows both correction priors. On refusal Think-SFT→Instr-SFT, both correction on the frozen probe and the from-scratch refit close 80% of the AUROC headroom with m = 100 prompts. On the harm checkpoint pair (headroom +0.006), $m^*(90\%)$=32,760.

Table 2: Budget $m^*(\tau)$: smallest m that closes $\tau\%$ of the AUROC headroom, by checkpoint pair and method.

Correction (frozen)				
Checkpoint pair	Headroom	80%	90%	95%
base → Think-SFT	+0.036	500	2,000	4,000
Think-SFT → Instr-SFT	+0.064	100	500	2,000
base → Instr-SFT	+0.007	4,000	8,000	15,000
harm base → dpo	+0.006	21,840	32,760	32,760
harm Instruct → clinical	+0.003	32,760	32,760	36,400
harm Instruct → code	+0.003	32,760	32,760	36,400
harm Instruct → science	+0.003	27,300	32,760	36,400
Refit				
Checkpoint pair	Headroom	80%	90%	95%
base → Think-SFT	+0.036	500	2,000	8,000
Think-SFT → Instr-SFT	+0.064	100	500	2,000
base → Instr-SFT	+0.007	4,000	15,000	15,000
harm base → dpo	+0.006	27,300	32,760	36,400
harm Instruct → clinical	+0.003	32,760	36,400	36,400
harm Instruct → code	+0.003	32,760	36,400	36,400
harm Instruct → science	+0.003	32,760	36,400	36,400

C Additional figures and value tables

C.1 Performance against total labeled prompts

The main budget plots vary the number of training prompts m while keeping the estimation set E fixed. Figures 13 and 14 show the same results against the total number of prompts whose exchanges are labeled. Correction and refits use $m+|E|$ prompts in total. The standalone ridge map uses labels only from E, so we plot the map fit on all N prompts at a total budget of $|E|$ labeled prompts. Here $|E| = 4{,}000$. These plots change the budget accounting, not the fitted probes or the size of E.

C.2 Zero-shot AUROC

Figure 16 is the AUROC companion to main-text Figure 2. Frozen probes achieve at least 93% of the refit's AUROC on every checkpoint pair, despite achieving only about half the refit's TPR on the most disruptive SFT transfers.

C.3 OLMo-3 7B headroom by post-training stage

Table 3 reports refit headroom by stage at the peak layer and at layer 16.

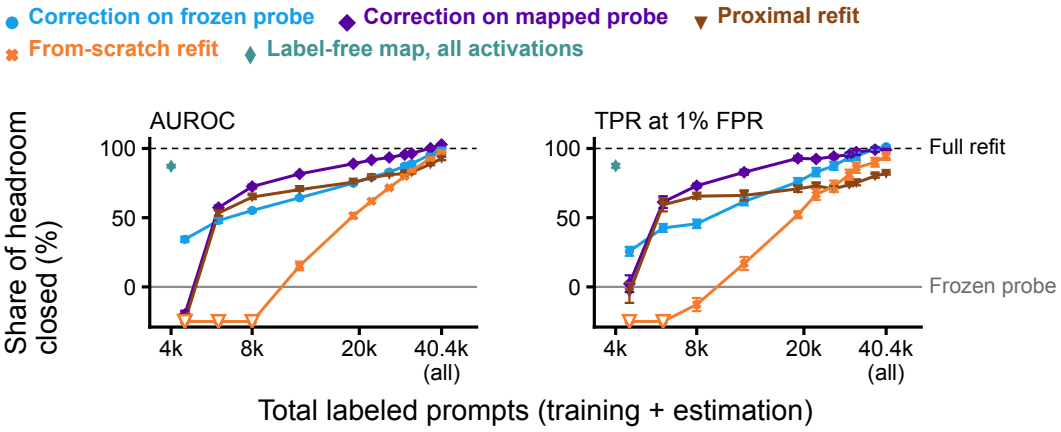

Figure 13: Harm performance against total labeled prompts, including the estimation set. OLMo-3 base→DPO ($N = 36{,}400$, $|E| = 4{,}000$): share of AUROC headroom closed (left) and share of TPR headroom closed at 1% FPR (right). Correction and refits use labeled exchanges from $m + |E|$ prompts. The standalone map uses all N prompts' activations and labels only from E. Mapped correction and proximal refit use maps fit on m prompts. Error bars show ± 1 SE; downward triangles mark values below -25%.

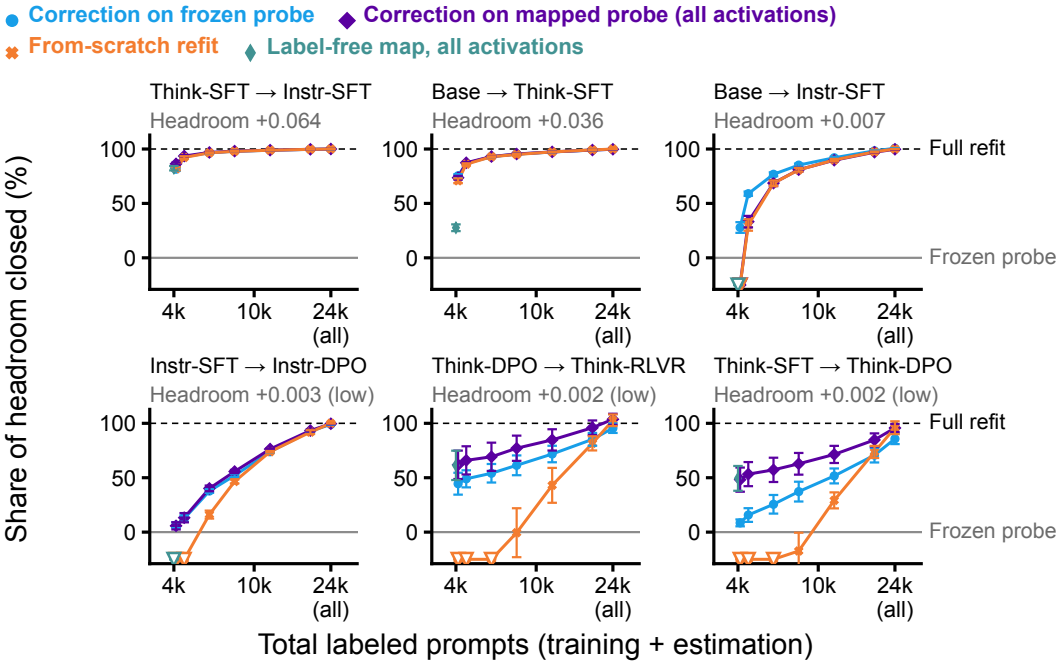

Figure 14: Refusal AUROC against total labeled prompts for six OLMo-3 checkpoint pairs ($N = 20{,}000$, $|E| = 4{,}000$). Correction and the from-scratch refit use labeled exchanges from $m + |E|$ prompts. The standalone map uses labels only from E. Both the standalone map and the mapped correction use activations from all N prompts. Error bars show ± 1 SE; downward triangles mark values below -25%.

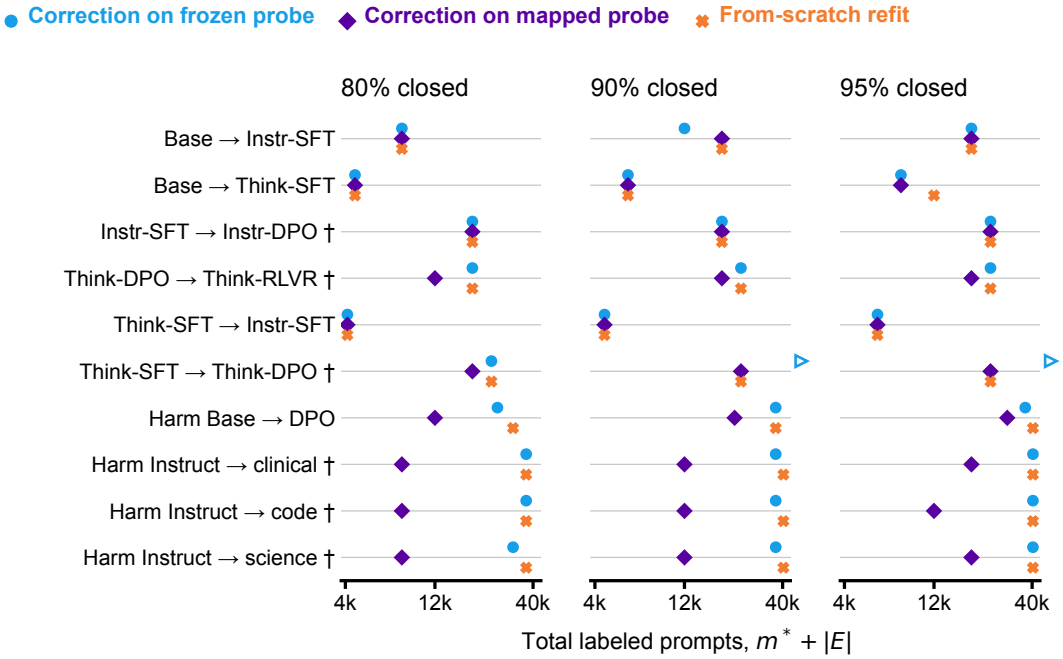

Figure 15: Total labeled prompts needed to close 80%, 90%, and 95% of AUROC headroom, across six refusal and four harm checkpoint pairs. Each point is the first tested training budget m whose mean AUROC reaches the target, plus the 4,000 estimation prompts; no interpolation is used. We show correction on the frozen probe, correction on the ridge-mapped probe, and refitting from scratch separately. The mapped prior uses activations from m prompts for harm and all N prompts for refusal. Open arrows mark targets not reached at any tested budget.

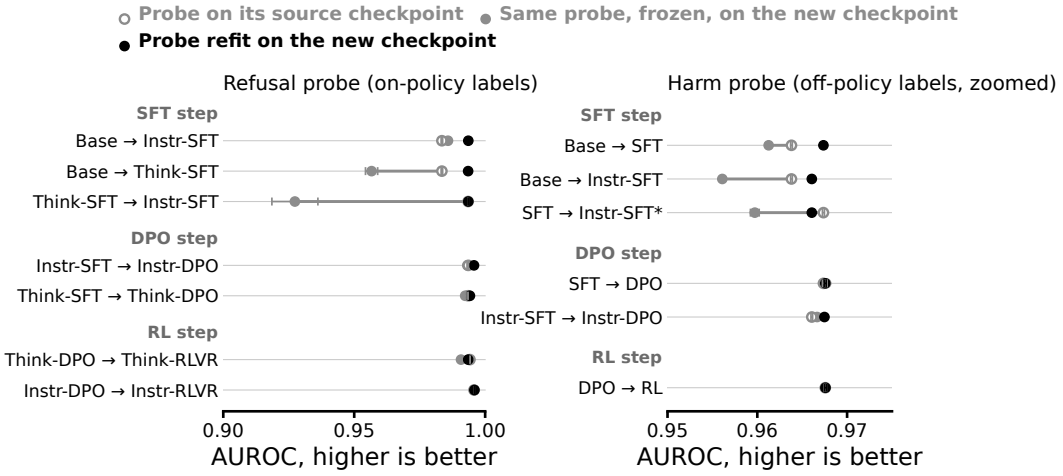

Figure 16: **Frozen probes achieve at least 93% of the refit's AUROC on every checkpoint pair.** The companion of Figure 2: AUROC of the probe on its source checkpoint (hollow), frozen on the new checkpoint (gray), and refit on the new checkpoint (black), per OLMo-3 checkpoint pair at layer 16 (± 1 SE, 8 seeds); the harm panel is zoomed to its narrower range. Even on SFT transfers where frozen probes achieve only about half the refit's true positive rate at 1% FPR, they remain close to the refit in AUROC.

Table 3: Refit headroom in AUROC by stage, OLMo-3 7B, harm. Values are means ±1 SE over five seeds. The peak layer has the largest mean headroom. SFT creates nearly all headroom; DPO and RL leave little headroom at layer 16.

Stage	Checkpoint pair	Peak layer	Headroom at peak layer	Headroom at layer 16
SFT	base → sft	31	$+0.012922 \pm 0.000981$	$+0.005126 \pm 0.000303$
SFT	sft → isft	30	$+0.006686 \pm 0.000751$	$+0.004832 \pm 0.000437$
DPO	sft → dpo	6	$+0.000023 \pm 0.000012$	-0.000125 ± 0.000022
DPO	isft → idpo	31	$+0.003733 \pm 0.000538$	$+0.000165 \pm 0.000221$
RL	dpo → rl	31	$+0.000056 \pm 0.000021$	-0.000009 ± 0.000017
Control	same weights	8	$+0.000005 \pm 0.000007$	-0.000013 ± 0.000003

C.4 Qwen3-4B headroom by post-training stage

Figure 17 shows harm refit headroom at each layer of Qwen3-4B; Table 4 gives values at the peak layer and at layer 18. SFT creates nearly all the refit headroom; RL adds little.

Table 4: Qwen3-4B refit headroom by stage (harm, ℓ_2=1,000). We report means ±1 SE over five seeds at the peak layer and at layer 18.

Checkpoint pair	Peak layer	Headroom at peak layer	Headroom at layer 18	Seeds
base → rl	34	$+0.006398 \pm 0.000405$	$+0.001056 \pm 0.000207$	5
base → sft	34	$+0.006429 \pm 0.000392$	$+0.001053 \pm 0.000189$	5
sft → rl	35	$+0.000239 \pm 0.000050$	-0.000033 ± 0.000017	5

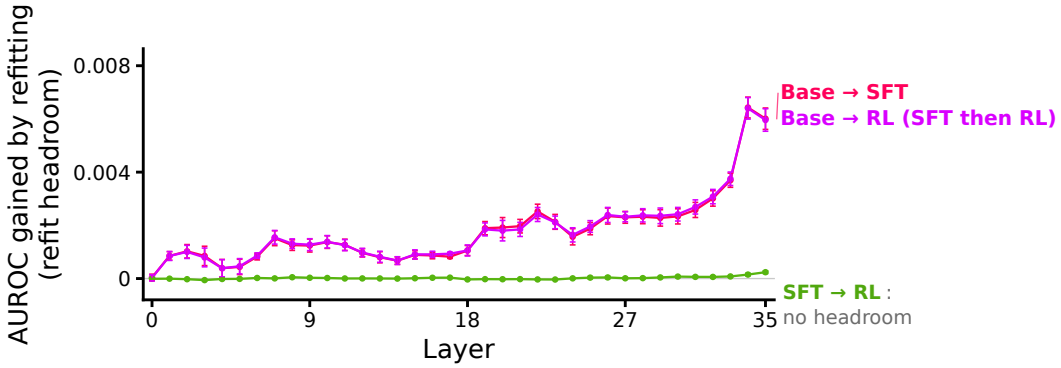

Figure 17: **On Qwen3-4B, SFT creates nearly all the refit headroom; RL adds little.** AUROC gained by refitting the harm probe at each layer (dots: mean over 5 seeds; whiskers: ±1 SE): base→SFT and base→RL nearly overlap, and SFT→RL stays near zero, replicating Figure 3.

C.5 Refit headroom at selected layers

Table 5 shows representative layer values from Figure 3.

Table 5: Refit headroom at selected layers, OLMo-3 7B, harm, 5 seeds.

Checkpoint pair	Layer	Headroom (mean)	Headroom (SE)	F_t	U_t
base → sft	0	0.001363	0.000294	0.881757	0.883120
base → sft	8	0.003219	0.000296	0.943101	0.946320
base → sft	16	0.005126	0.000303	0.957463	0.962590
base → sft	24	0.006525	0.000172	0.949782	0.956307
base → sft	31	0.012922	0.000981	0.939223	0.952145
sft → dpo	16	−0.000125	0.000022	0.962963	0.962838
sft → dpo	31	−0.000032	0.000074	0.953343	0.953311
dpo → rl	16	−0.000009	0.000017	0.962882	0.962873

C.6 Harm AUROC and partial AUC by prompt budget

Figure 18 shows the pAUC $\leq 1\%$ and $\leq 5\%$ FPR companions to main-text Figure 5. Tables 6 and 7 give the numeric AUROC and pAUC$\leq 5\%$ values.

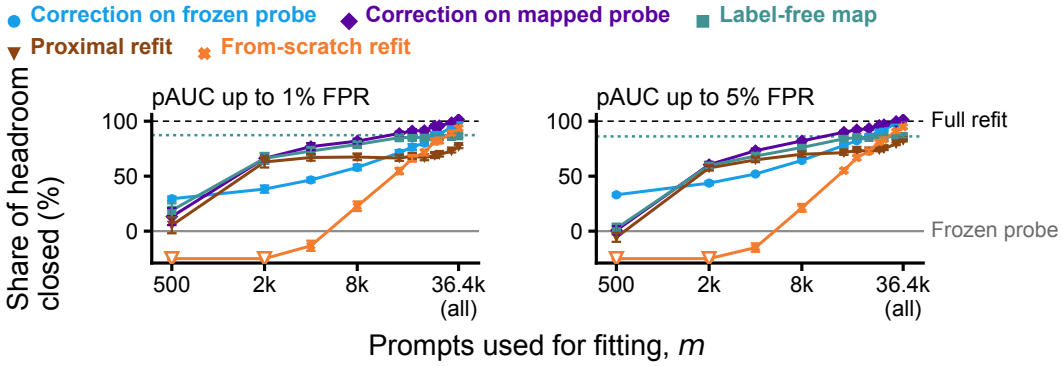

Figure 18: **At small label budgets, correction has higher pAUC than the from-scratch refit.** Harm probe, base→DPO (8 seeds, ±1 SE): share of the headroom closed on standardized pAUC up to 1% FPR (left) and up to 5% FPR (right); conventions as in Figure 5.

Table 6: Share of AUROC headroom closed by prompt budget, harm base→dpo (nested, N=36,400, 8 seeds). Frozen = 0.9636, ceiling = 0.9695, headroom = +0.0059.

Method	1.4%	5.5%	11%
Map on m	-16.5 ± 3.6	53.3 ± 2.1	64.3 ± 2.0
Correction (frozen)	34.4 ± 1.9	47.9 ± 1.7	55.2 ± 1.1
Correction (map on m)	-19.9 ± 2.7	57.3 ± 1.6	72.5 ± 1.5
Proximal refit (map)	-30.2 ± 3.1	53.3 ± 2.0	64.9 ± 1.5
From-scratch refit	-273.4 ± 11.8	-83.9 ± 5.8	-26.5 ± 4.2
Method	**22%**	**50%**	**100%**
Map on m	76.2 ± 1.1	85.7 ± 0.8	86.6 ± 0.9
Correction (frozen)	64.4 ± 1.0	79.2 ± 0.8	100.0 ± 0.6
Correction (map on m)	81.6 ± 0.8	91.7 ± 0.7	102.9 ± 0.4
Proximal refit (map)	70.3 ± 0.9	78.8 ± 0.8	92.8 ± 1.2
From-scratch refit	15.2 ± 3.1	61.9 ± 1.3	98.2 ± 0.6
Map on all N	87.2 ± 1.0 (budget-invariant)		

Table 7: Share of standardized pAUC headroom closed by prompt budget, harm base→dpo ($\leq 5\%$ FPR). Frozen = 0.8610, ceiling = 0.8801, headroom = $+0.0191$.

Method	1.4%	5.5%	11%
Map on m	2.5 ± 4.3	59.7 ± 2.5	68.2 ± 2.9
Correction (frozen)	33.0 ± 1.3	43.7 ± 1.8	52.0 ± 1.0
Correction (map on m)	0.8 ± 3.8	60.5 ± 2.6	73.2 ± 2.0
Proximal refit (map)	-5.6 ± 4.1	57.5 ± 2.4	64.8 ± 2.1
From-scratch refit	-204.6 ± 11.6	-63.8 ± 4.1	-14.9 ± 3.8
Method	**22%**	**50%**	**100%**
Map on m	76.1 ± 1.9	85.0 ± 1.4	86.1 ± 1.2
Correction (frozen)	64.3 ± 1.6	82.3 ± 1.2	101.0 ± 0.4
Correction (map on m)	82.0 ± 1.5	92.2 ± 1.0	101.5 ± 0.5
Proximal refit (map)	69.9 ± 2.3	73.5 ± 1.5	82.5 ± 0.9
From-scratch refit	21.2 ± 3.6	66.7 ± 1.9	94.9 ± 1.7
Map on all N	86.1 ± 1.4 (budget-invariant)		

C.7 Refusal on OLMo-3 7B

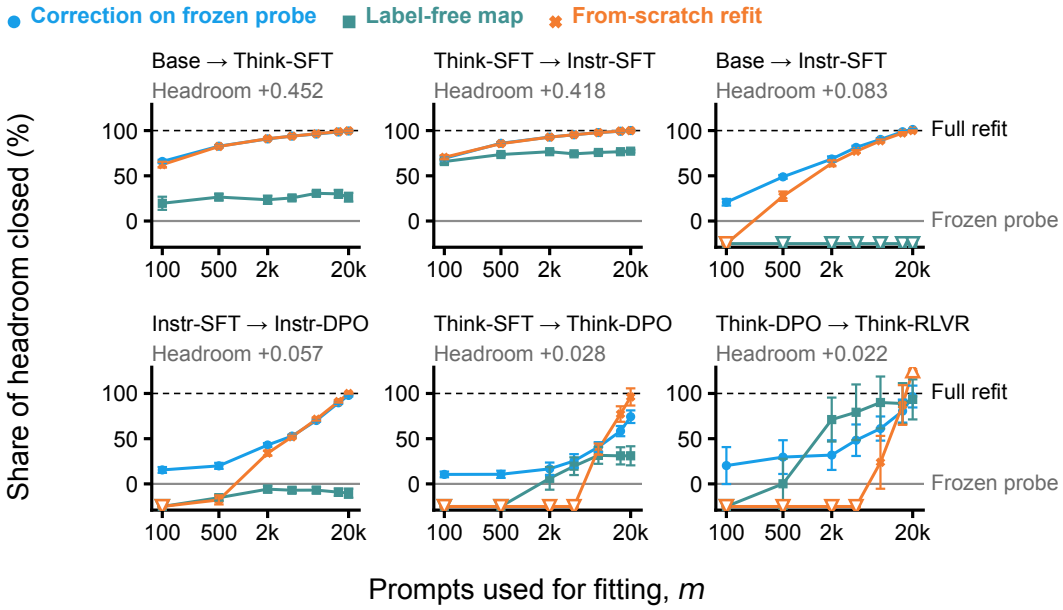

Figure 19: **At the 1% FPR operating point, correction on the frozen probe generally leads the from-scratch refit at small budgets.** Six OLMo-3 refusal checkpoint pairs, ordered by TPR headroom: share of the TPR headroom closed against new labeled prompts m (8 seeds, ± 1 SE); conventions as in Figure 6.

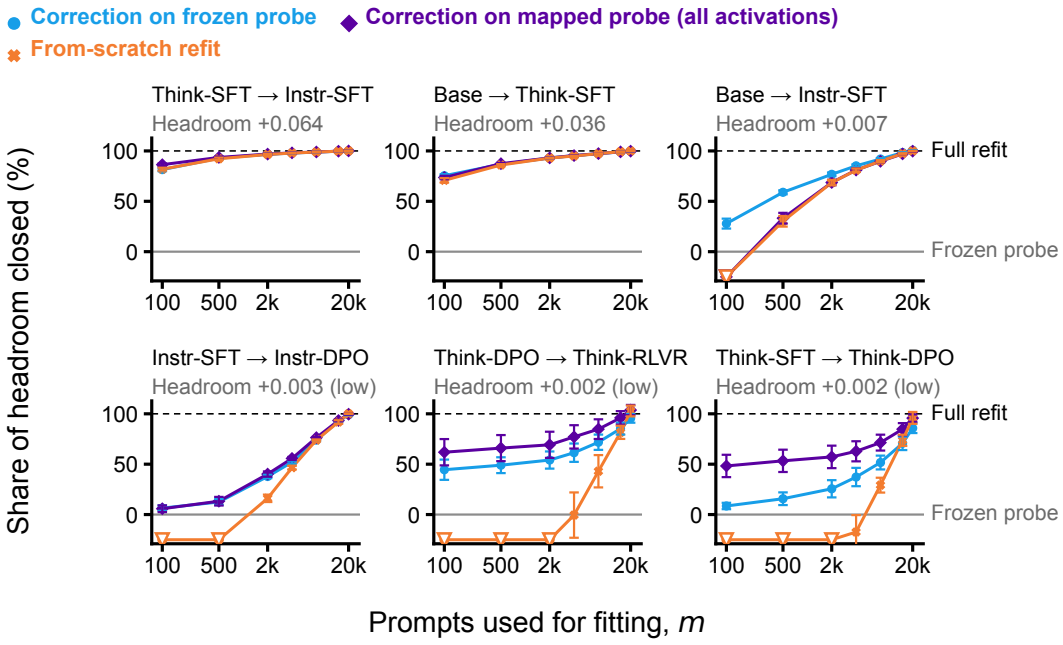

Figure 20: **With activations for all N prompts but labeled exchanges from only m training prompts, one of the two corrected probes has the highest AUROC at most tested budgets. Which correction performs better, on the frozen prior or on the ridge-mapped prior, depends on the checkpoint pair.** Six OLMo-3 refusal checkpoint pairs (8 seeds, ± 1 SE): share of AUROC headroom closed against the number of labeled training prompts m for correction on the probe mapped using all N activations, the correction on the frozen probe, and the from-scratch refit; conventions as in Figure 6.

C.8 REFUSAL ON QWEN3-4B

Figures 21 and 22 show refusal AUROC and TPR across prompt budgets on the three Qwen3-4B checkpoint pairs. Figure 23 shows the decoupled-budget comparison.

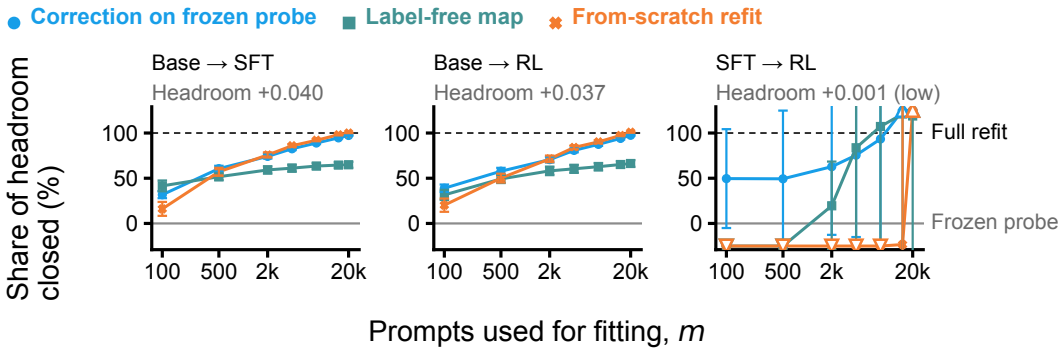

Figure 21: **On Qwen3-4B refusal, correction on the frozen probe has higher AUROC than the from-scratch refit at small label budgets. On base→SFT and base→RL, the label-free map remains below the full-refit ceiling.** Three checkpoint pairs (N=20,000): share of the AUROC headroom closed against new labeled prompts m; conventions as in Figure 6. Results use eight seeds, except base→SFT correction at m=100, which uses six. Error bars show ±1 SE. Headroom on SFT→RL is small, so small AUROC differences produce large differences in the share of headroom closed.

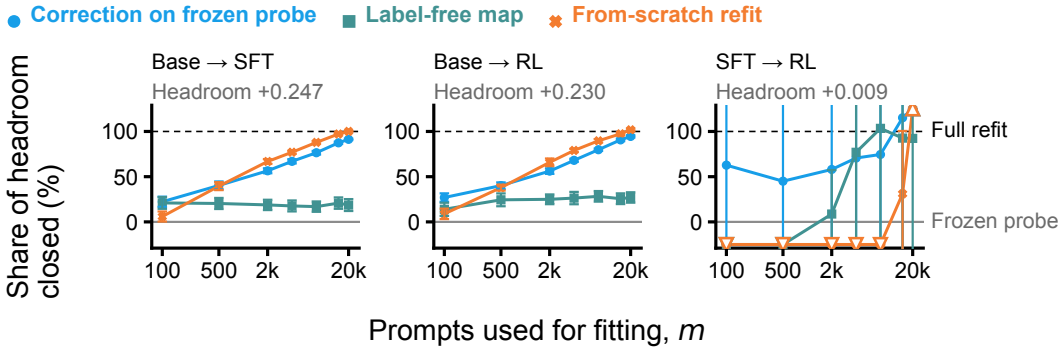

Figure 22: **Correction on the frozen probe consistently improves TPR on Qwen3-4B refusal.** Share of TPR headroom closed at 1% FPR against new labeled prompts m, for three checkpoint pairs. Results use eight seeds, except base→SFT correction at m=100, which uses six. Error bars show ±1 SE.

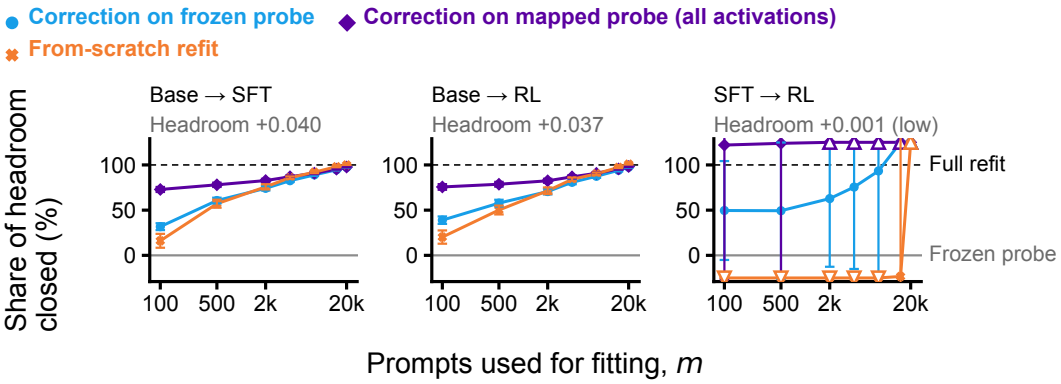

Figure 23: **On Qwen3-4B refusal, correction can use a map fit on all N prompts while fitting the correction on labeled exchanges from m training prompts.** Three checkpoint pairs; the relative performance of correction on the mapped and frozen probes depends on the checkpoint pair and budget. Results use eight seeds, except base→SFT correction at m=100, which uses six for both priors. Error bars show ±1 SE. Conventions as in Figure 20.

C.9 Domain SFT

Figure 24 shows the raw AUROC (not share of headroom closed) for the three domain-SFT checkpoint pairs.

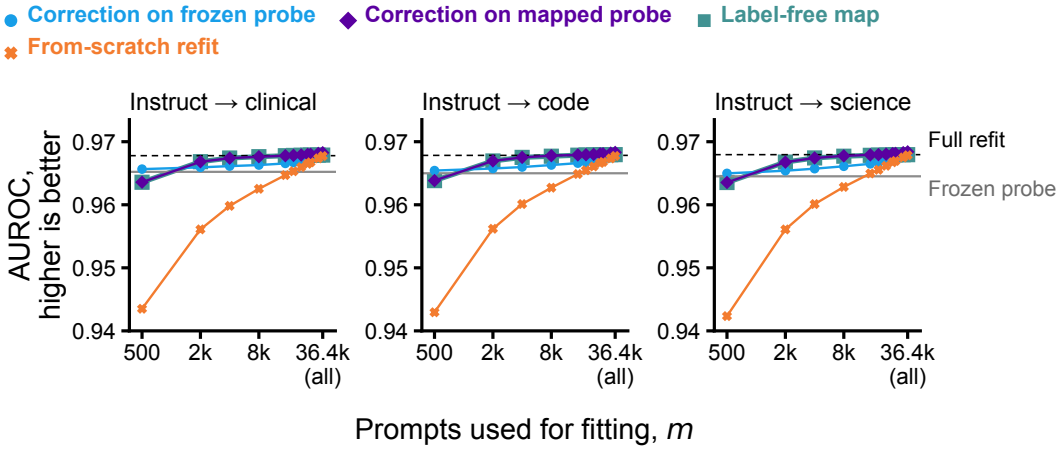

Figure 24: **Refit headroom is approximately 0.003 AUROC in each of the three domain-SFT checkpoint pairs.** Harm probe, Instruct→clinical, code and science (N=36,400): AUROC of each method against new prompts m, with the frozen probe (gray line) and the full refit (dashed) as references; the companion of Figure 8.

C.10 Layer sweep

Figure 25 compares methods across layers at two prompt budgets.

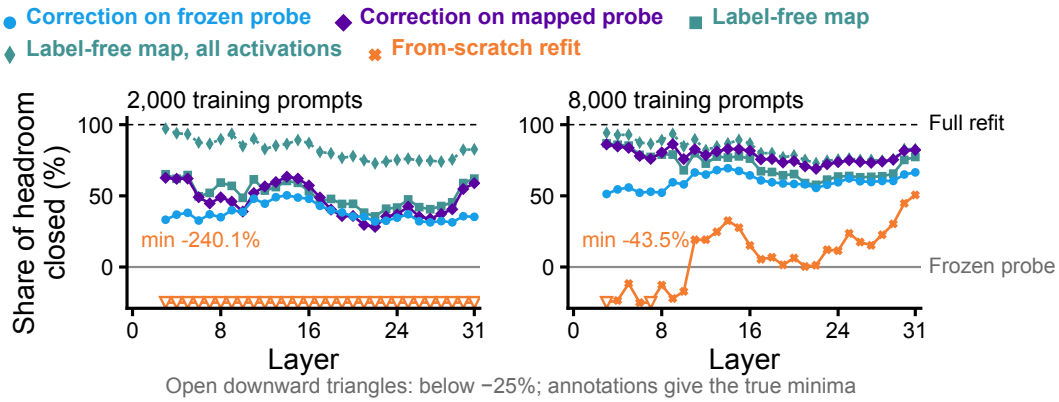

Figure 25: **Layer by layer, the from-scratch refit stays below the frozen probe with 2,000 training prompts and below every other method with 8,000.** Share of the AUROC headroom closed on harm base→DPO at each layer, the line version of Figure 9; layers 0 to 2 are omitted for lack of headroom, and the dotted line with diamond markers is the map fit on all N activations. Open downward triangles mark shares below -25%; annotations report the minimum at each budget.

C.11 Harm on Qwen3-4B

Figure 26 compares harm AUROC and TPR across prompt budgets for Qwen3-4B base→SFT and base→RL.

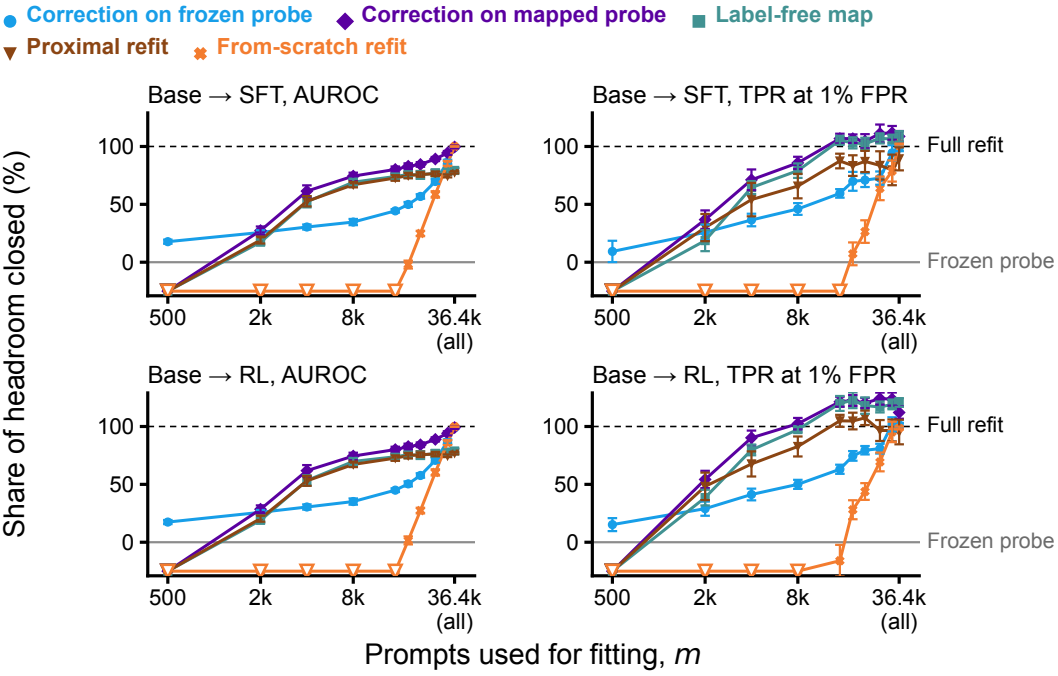

Figure 26: **On Qwen3-4B harm, correction on the frozen probe performs best at the smallest budget; correction on the ridge-mapped probe performs best at most larger budgets.** base→SFT (top) and base→RL (bottom), N=36,400, 8 seeds, ±1 SE, on AUROC (left) and TPR at 1% FPR (right); conventions as in Figure 5. SFT→RL has negligible headroom and is not shown.

Table 8 reports paired AUROC differences across eight seeds and ten prompt budgets on Qwen3-4B. Averaged across these seeds and budgets, correction on the frozen prior has higher AUROC than both the frozen probe and the from-scratch refit on the two checkpoint pairs that include SFT.

On SFT→RL, correction on the frozen probe has nearly the same AUROC as the frozen probe and higher AUROC than the from-scratch refit.

Table 8: Mean paired AUROC differences on Qwen3-4B across eight seeds and ten prompt budgets. The SE and t are descriptive summaries of the 80 combinations of seed and budget. Each seed is used at all ten budgets.

Comparison	Mean	SE	t	Cells
Checkpoint pair: base → rl				
Correction/frozen − map all N	−0.00061	0.00007	−9.3	80
Correction/frozen − frozen	+0.00114	0.00006	+17.7	80
Correction/frozen − refit	+0.00494	0.00081	+6.1	80
Correction/map m − prox. refit/map m	+0.00022	0.00002	+14.1	80
Checkpoint pair: base → sft				
Correction/frozen − map all N	−0.00060	0.00006	−9.6	80
Correction/frozen − frozen	+0.00109	0.00006	+17.7	80
Correction/frozen − refit	+0.00498	0.00081	+6.1	80
Correction/map m − prox. refit/map m	+0.00021	0.00001	+13.7	80
Checkpoint pair: sft → rl				
Correction/frozen − map all N	+0.00003	0.00001	+5.9	80
Correction/frozen − frozen	+0.00001	0.00001	+1.9	80
Correction/frozen − refit	+0.00599	0.00086	+7.0	80
Correction/map m − prox. refit/map m	+0.00004	0.00001	+5.0	80

D DATASETS, POST-TRAINING, AND ACTIVATION EXTRACTION

Harm dataset. The harm label is binary (harm present/absent in the exchange) and off-policy: the text is fixed across checkpoints, so labels carry over without relabeling. The main harm experiments use N=36,400 training prompts and $|E| = 4{,}000$ estimation prompts for both OLMo-3 and Qwen3-4B. The intermediate-SFT experiments use $N = 30{,}000$ and $|E| = 800$. The estimation set is disjoint from both training and test data.

Refusal dataset. The refusal label depends on whether the checkpoint's own response refuses the prompt (on-policy). Refusal labels are assigned again after each update. Training pool size is N=20,000 for OLMo-3 checkpoint pairs; the Qwen3-4B refusal pool is N=20,000.

Qwen3-4B post-training. We post-train Qwen3-4B ourselves: SFT on SmolTalk (Ben Allal et al., 2025), then GRPO on OpenR1-Math (Hugging Face, 2025). Both mixtures are filtered to remove overlap with our probe datasets. This repeats the analysis on a second model family with different post-training datasets.

Activation extraction. All activations are extracted from the residual stream at the specified layer. Each exchange yields one pooled activation vector. All exchanges associated with the same prompt belong to one prompt group. For the harm label, all exchanges are pre-generated and fixed across checkpoints. For the refusal label, the model's own response is generated per checkpoint.