WHERE CAN ADAPTER COMPOSITION HELP? BASE-MODEL MARGINS LOCALIZE LORA COMPLEMENTARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

When is it worth composing LoRA adapters instead of using the base model alone? For multiple-choice evaluation with a pool of adapters trained on the same base model, an oracle that may fall back to the base model gains only where the base model errs; we ask *which* base errors a pool can fix, and whether they can be found from the base model before any adapter is run. On all 14,042 MMLU test items, the Qwen3.5-9B base model is more accurate than each of its four adapters, yet the pool holds 8.28 points of oracle headroom, and the share of base errors that some adapter fixes falls from 49.5% in the lowest base-margin quartile to 2.0% in the highest. The same decline appears for independently trained Llama-3-8B adapters and for a new pool with a different domain mix, and on five tested workloads, where adapters change base answers, observable without labels, tracks where their rescues lie. A threshold frozen from 400 labels before the new pool was trained keeps 91.7% and 94.0% of its rescues on MMLU and MMLU-CF while sending about a third of the items to the adapters. Refitted post hoc after eight numeric calibration labels are mapped to letters, the threshold keeps 90.0% and 92.1% of these rescues with fewer passes. These rescues are the opportunity left to any selector; an exact decomposition of router improvement into this ceiling, missed gains and damage, together with a margin ceiling and a perturbation condition, serves as the accounting.

1 INTRODUCTION

Should you compose your LoRA adapters? Weight-space merging, recipe search, and input-conditioned routing—Task Arithmetic (Ilharco et al., 2023), TIES-Merging (Yadav et al., 2023), DARE (Yu et al., 2024), Model Soups (Wortsman et al., 2022), LoraHub (Huang et al., 2024), Twin-Merging (Lu et al., 2024), and PHATGOOSE (Muqeeth et al., 2024)—all answer *how* to combine adapters. Before choosing a method, a practitioner faces a prior question: does the adapter pool contain any correct answers that the base model misses on this workload, and where are they? We give a cheap, workload-level answer for multiple-choice (MCQ) logprob evaluation and test it on held-out items and independent benchmarks.

The problem. Consider a base language model f_0 and N domain-specific LoRA adapters producing models $f_1, \ldots, f_N$. For item x, the *oracle gain* $g^+(x) \in \{0, +1\}$ is $+1$ if some adapter answers correctly where the base model fails. Its average is the ceiling for any per-item selector over the pool. We ask whether this helpable set can be located from a small labeled calibration sample plus unlabeled base-model margins on the target workload.

Our approach. Because an adapter can only help where the base model errs, the tail complementarity $C_>(\tau) = \mathbb{E}[g^+(x) \mid m(x) \geq \tau]$ factors into the base model's selective risk above margin τ and the rescue rate h, the share of those base errors that the pool fixes. Section 3 makes this precise and links the pool's benefit to selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) and cascades (Jitkrittum et al., 2023; Dekoninck et al., 2025). The selective-risk factor alone bounds the benefit; we study the second factor: how does the share of fixable errors depend on the base margin, and can that dependence be used before any adapter is evaluated? Two properties of

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.

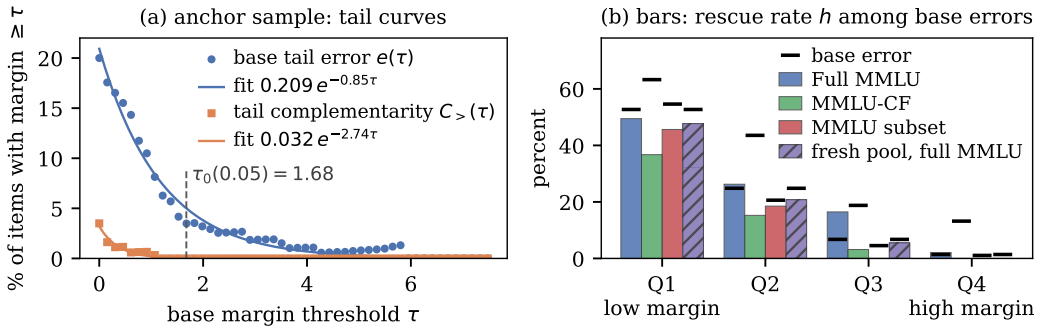


Figure 1: (a) With the original labels of the 400-item science+medical anchor, the base tail error $e(\tau) = \mathbb{P}(\text{base wrong} \mid m \geq \tau)$ and the tail complementarity $C_>(\tau)$, the oracle gain among items with margin $\geq \tau$, each with its exponential fit; the dashed line is the screening threshold of Section 3, the margin at which the fitted tail error reaches 5%. (b) Per base-margin quartile on three workloads not used for fitting with the original four adapters, and in hatched bars on full MMLU with the freshly trained pool of Section 4.5: bars show the rescue rate h, the share of base errors that at least one adapter answers correctly, and black ticks the base error rate. Low-margin errors are both more frequent and far more often fixable.

this setting shape the answer. On full MMLU the base model is the most accurate single model under our zero-shot protocol, as Table 6 shows, so the value of the pool lies entirely in per-item complementarity rather than in a stronger model to defer to. And by Lemma 6, any score perturbation, an adapter's included, can change the base answer only if it shifts some option by at least half the base margin, so moving a confident base answer takes a large score shift.

Contributions.

1. **Base confidence predicts which errors a pool can fix.** As Figure 1 and Tables 2 and 7 show, the rescue rate h falls from 36.7–49.5% in the lowest-margin quartile to at most 2.0% in the highest on three workloads, and from 47.7% to 0.0% for a freshly trained pool with a different domain mix. The decline holds within MMLU subjects, with independently trained Llama-3-8B adapters, and under a 5-shot prompt, and it tracks how close the gold option already is.
2. **A label-free view of where rescues lie.** Where the pool changes base answers, observable without labels, follows the margin profile of its rescues: on the five tested workloads of Section 4.2, the share of answer changes below the screening threshold lies 0.1–9.0 points below the share of rescues there, and it orders the margin quartiles like h.
3. **A base-only screen that transfers to a new pool.** A threshold set from 400 labeled items needs no adapter passes and, frozen before a new pool was trained, keeps 91.7% and 94.0% of that pool's rescues on full MMLU and MMLU-CF at 2.3–2.6 instead of five candidate passes per item; for the original pool it keeps 86–95% of the rescues on four workloads. Section 4.6 tests calibration resampling, overlap removal, pool growth and Llama-3-8B transfer; Section H separates the original frozen results from a post-hoc label correction.
4. **Accounting tools.** A margin ceiling, the exact identity $\mathrm{Imp}(r) = \mathrm{OC} - \mathrm{Miss} - \mathrm{Damage}$, and a perturbation condition, all in Section 3, put screening and routing in the units of the oracle ceiling; for an item-disjoint supervised router, the identity splits its distance to the ceiling into missed gains and damage.

2 Related Work

Selective prediction and cascades. The tail error $\mathbb{P}(\text{base wrong} \mid m \geq \tau)$ is the base model's selective risk (El-Yaniv & Wiener, 2010); Geifman & El-Yaniv (2017) choose such thresholds with a risk guarantee, which we use as a baseline in Table 11. Cascades defer to a second model when confidence is low: Jitkrittum et al. (2023) show that the optimal rule compares the two models' correctness probabilities and give conditions under which confidence thresholding is optimal, namely

that it orders inputs like that difference, which specialist second models can violate. LLM cascades and routers (Gupta et al., 2024; Chen et al., 2024; Dekoninck et al., 2025; Kotte, 2026) trade cost against quality, and Dekoninck et al. (2025) treat routing and cascading in one framework. Remark 4 ties the oracle benefit of a pool to this selective risk. Our estimand differs from a deferral rule between two models: for a fixed pool of same-base adapters, we measure how the rescue rate among base errors depends on the base margin, and screen with the base model alone before any adapter runs.

Routing and composition. Routers across LLMs (Hu et al., 2024; Ong et al., 2025), LLM ensembling by pairwise ranking (Jiang et al., 2023), and routers across adapters or experts (Pfeiffer et al., 2021; Wu et al., 2024; Muqeeth et al., 2024; Ostapenko et al., 2024; Lu et al., 2024) learn which candidate to use; SLIM (Han et al., 2025) routes between LoRA experts and an identity branch so that inputs the adapters would harm keep base behavior. We estimate beforehand, from the base model alone, where a same-base pool can help, and account for what a router realizes with Proposition 5. Task Arithmetic (Ilharco et al., 2023), TIES-Merging (Yadav et al., 2023), DARE (Yu et al., 2024), Model Soups (Wortsman et al., 2022), LoRA Soups (Prabhakar et al., 2025) and LoraHub (Huang et al., 2024), which fixes one composition per task, apply the same model to every input, so in our accounting their improvement is their gains minus their damage on base-correct items.

When composition works. Merging works better from strong base models (Yadav et al., 2025). Rahamim et al. (2026) study example-level mergeability through the base model's gap between its top and gold answers; we show that this gold gap predicts whether a pool of unmerged adapters rescues a base error, as Section B shows, and use the label-free top-two margin, which lower-bounds that gap on every base error, to screen a target workload for likely rescues before any adapter is run. Similarity and gradient signals predict merged-model quality or operator choice (Bolton et al., 2026; Zhou et al., 2026), Tam et al. (2026) evaluate whether merging achieves compositional generalization, Chen et al. (2026) question whether reuse composes knowledge, and Liu et al. (2026) find that random auxiliary LoRAs match recycled ones when a trained target-task LoRA is merged too.

3 Setup and Accounting Framework

We define the quantities measured in Section 4. The central one is the rescue rate h, the share of base errors that the pool can fix, and its dependence on the base margin; three accounting tools state where gains can exist, how much a router realizes, and how large a score shift must be to change a base answer: a margin ceiling, an exact decomposition and a perturbation condition.

3.1 Quantities

Consider a base model f_0 and N domain-specific LoRA (Hu et al., 2022) adapters. All $\arg\max$ operations resolve ties, which finite-precision scores can produce, to the first maximizing option. Conditional expectations are defined only when $\mathbb{P}(m(x) \geq \tau) > 0$.

Definition 1 (Score vectors and margin)**.** *Let $f_0(x) \in \mathbb{R}^C$ denote the base model's logprob vector over C answer options for input x, and let $f_i(x)$ denote the score vector when adapter i is applied. The* base margin *is*

$$m(x) = f_0(x)[\hat{y}] - \max_{c \neq \hat{y}} f_0(x)[c], \qquad \hat{y} = \arg\max_c f_0(x)[c], \tag{1}$$

the logprob gap between the top-1 and top-2 predictions, computable without labels.

Definition 2 (Oracle gain and helpable set)**.** *For item (x, y) with true label y, the set of adapters that would help is*

$$A_+(x) = \{\, i \in \{1, \ldots, N\} : f_i(x) \text{ correct and } f_0(x) \text{ wrong} \,\}.$$

The oracle over base and all adapters gains $g^+(x) = \mathbf{1}[A_+(x) \neq \emptyset] \in \{0, 1\}$ over the base model, and its mean $\mathrm{OC} = \mathbb{E}[g^+(x)]$ is the adapter oracle ceiling.

Definition 3 (Tail error, tail complementarity and rescue rate)**.** *For a margin level τ, the* tail error $e(\tau) = \mathbb{P}($*base wrong* $\mid m(x) \geq \tau)$ *is the base model's selective risk, the* tail complementarity *is $C_>(\tau) = \mathbb{E}[g^+(x) \mid m(x) \geq \tau]$, and the* rescue rate *is $h_>(\tau) = \mathbb{P}(A_+(x) \neq \emptyset \mid$*

base wrong, $m(x) \geq \tau$). *Because* $g^+(x) = 1$ *only on base errors,*

$$C_>(\tau) = e(\tau) \cdot h_>(\tau). \tag{2}$$

The same factorization holds within any margin bin such as a quartile; h *without subscript denotes the rescue rate within a bin.*

Remark 4 (Margin ceiling)**.** *Since* $h_>(\tau) \leq 1$*, Equation* (2) *gives* $C_>(\tau) \leq e(\tau)$ *for all* τ*: the oracle benefit of any pool is bounded by the base model's selective risk. The ceiling would hold with* h *constant in the margin; Section 4.2 measures how* h *itself varies.*

Threshold rules. A screen evaluates the adapters only on items with $m(x) < \theta$. To place θ from n labeled items, we estimate the tail error by the least-squares curve $\hat{e}(\tau) = \delta e^{-\alpha\tau}$ on a grid of empirical tail errors and set

$$\tau_0(\epsilon) = \frac{1}{\alpha} \ln \frac{\delta}{\epsilon}, \tag{3}$$

the margin at which the estimated tail error reaches ϵ; by the ceiling it estimates the margin above which the residual oracle gain is at most ϵ. The curve is a trend estimate used to place a threshold, not a finite-sample bound; its residuals are in Section J.2, and Section 4.4 places thresholds from the same labels by selection with guaranteed risk and by a margin quantile. The original 400 anchor labels gave $\alpha = 0.8537$, $\delta = 0.2092$ and the frozen $\tau_0(0.05) = 1.68$; see Section H for the label correction. The screen gives up only the oracle gain above θ; if a fraction p of items lies below θ, $\mathrm{OC} \leq p + (1-p)\, e(\theta)$, so the threshold localizes headroom rather than certifying that the workload has none.

3.2 Router Improvement Decomposition

Proposition 5 (Router Improvement Decomposition)**.** *For any router* $r : \mathcal{X} \to \{0, 1, \ldots, N\}$ *that selects which model to use per input, with* 0 *denoting the base model,*

$$\mathrm{Imp}(r) = \mathrm{OC} - \mathbb{E}[\mathrm{Miss}(r, x)] - \mathbb{E}[\mathrm{Damage}(r, x)], \tag{4}$$

where $\mathrm{Imp}(r)$ *is the router's accuracy minus the base accuracy and*

$$\mathrm{Miss}(r, x) = \mathbf{1}[A_+(x) \neq \emptyset] \cdot \mathbf{1}[r(x) \notin A_+(x)], \tag{5}$$

$$\mathrm{Damage}(r, x) = \mathbf{1}[f_0(x) \textit{ correct}] \cdot \mathbf{1}[f_{r(x)}(x) \textit{ wrong}]. \tag{6}$$

The identity, proved in Section O, puts missed helpable items and routing damage in the units of the oracle ceiling; a perfect oracle has $\mathrm{Miss} = \mathrm{Damage} = 0$. A composed predictor g, such as a merge, a soft mixture or a fixed choice of one adapter, is a new candidate: $\mathrm{Imp}(g) = \mathrm{Gain}_{\mathrm{in}}(g) + \mathrm{Gain}_{\mathrm{out}}(g) - \mathrm{Damage}(g)$, where $\mathrm{Gain}_{\mathrm{in}} \leq \mathrm{OC}$ counts base errors in the helpable set that g fixes and $\mathrm{Gain}_{\mathrm{out}}$ those that no single adapter fixes.

3.3 Stability of Confident Base Answers

Lemma 6 (Perturbation stability)**.** *Let* $\delta_M(x) = f_M(x) - f_0(x)$ *be the score perturbation from applying an adapter or a merge* M*. A prediction flip, that is, a change of* $\arg\max$ *including a tie resolved against* $\hat{y}$*, requires* $\|\delta_M(x)\|_\infty \geq m(x)/2$*. For* $\tau, w > 0$*, let* $B_{\tau,w} = \{$*base correct,* $m \in [\tau, \tau + w]\}$ *have positive probability. Markov's inequality gives:*

$$\mathbb{P}(\textit{flip} \mid B_{\tau,w}) \leq \frac{4\,\mathbb{E}[\|\delta_M(x)\|_\infty^2 \mid B_{\tau,w}]}{\tau^2}, \tag{7}$$

provided the conditional second moment is finite.

The lemma is a necessary condition, since the conditional moment can itself change with τ; as Table 2 shows, on full MMLU an adapter changes a top-quartile base answer at a mean rate of 0.04% when the base model is correct and 0.5% when it is wrong, against 29.7% and 37.7% in the lowest-margin quartile.

4 Experiments

We test margin-based rescue prediction, threshold transfer across workloads and pools, calibration sensitivity, and realized selector gains.

4.1 SETUP

Base model and adapters. We use the public Qwen3.5-9B checkpoint (Qwen Team, 2026) from the Qwen model family (Yang et al., 2025) with four LoRA adapters of rank 16, $\alpha = 32$ and dropout 0.05, trained for 2 epochs in bf16. The science adapter is trained on the 11,679 examples of the SciQ training split (Welbl et al., 2017) and the medical adapter on 19,765 Medical Meadow medical flashcards (Han et al., 2023); the history and philosophy adapters are trained on MMLU validation and dev questions of those subjects and are evaluated only on MMLU test rows. Target modules and training details are in Section N.

Evaluation. Candidates predict the argmax mean per-token logprob over option letters; the margin is the top-1/top-2 score gap. Science items come from ARC-Challenge (Clark et al., 2018), medical items from MedMCQA (Pal et al., 2022), and history and philosophy items from the corresponding MMLU (Hendrycks et al., 2021) test subjects. Each domain pair is evaluated on 400 items, 200 per domain drawn with a sample seed, and the six-pair matrix is repeated for six seeds, giving 36 pair-seed traces and 14,400 records. Because seeds and pairs resample the same benchmark questions, these records contain 3,361 distinct items; every holdout analysis below splits by item. The records contain 16 ARC items whose numeric gold labels do not match the letter predictions; removing them moves the mean oracle gap from 6.97 to 7.00pp. Section 4.6 rescores ARC-Challenge; Section H corrects the anchor labels using its original predictions.

4.2 ORACLE GAIN CONCENTRATES AT LOW MARGIN

Stratified by base-margin quartile, the seed-1 matrix of Tables 26 and 28 in the appendix shows that in every pair the lowest-margin quartile carries 12.9–25.5 points of oracle gain, while the highest carries at most 2 points and exactly zero in five pairs. The pattern holds over all 36 pair-seed traces, with an overall gap of $+6.97 \pm 3.07$pp, $+19.2 \pm 5.5$pp in Q1 and $+0.39 \pm 0.97$pp in Q4 in Table 29; for the four-adapter oracle over the 3,361 distinct items, with gap $+9.28$pp, $+25.76$pp in Q1 and 0.00pp in Q4 in Table 1; and in an all-pair Llama-3-8B (Grattafiori et al., 2024) replication with independently trained adapters, with mean gap $+12.13 \pm 4.31$pp and Q1/Q4 gaps of $+30.61/0.00$pp after label correction in Table 35.

Table 1: Four-adapter oracle over the 3,361 distinct items of the six-pair records. Each item appears in all three pairs that contain its domain, so all four adapters are evaluated on it; candidate scores and correctness agree across seeds without a single conflict. Quartiles are by base margin.

Items	Base	Science / medical / history / philosophy adapter	Oracle	Gap	Q1 gap	Q4 gap
3,361	0.768	0.760 / 0.756 / 0.728 / 0.762	0.861	+9.28	+25.76	0.00

On the science+medical pair the base model is the most accurate single model, at 0.802 against 0.788 and 0.780 for the adapters, and the oracle reaches 0.840; on all 14,042 MMLU test rows of Table 6 base accuracy is 78.81%, the strongest single adapter 78.64%, and the oracle over base plus four adapters 87.09%: the complementarity is invisible in aggregate accuracy.

Margin predicts rescue, not only error. By Equation (2), $C_>$ is base error times the rescue rate h, and the ceiling of Remark 4 would hold with h constant. It is not constant: as Figure 1b and Table 2 show, h falls with margin, on full MMLU from 49.5% [47.2, 51.8] in the lowest quartile to 2.0% in the highest, and on MMLU-CF from 36.7% to 0.0% although 13.2% of top-quartile MMLU-CF items are base errors, 66 errors against 49 in the full-MMLU top quartile. Adapters change a top-quartile base answer at a mean rate of at most 0.5%, so confident base errors are shared by the whole pool. Among base errors, a low margin predicts rescue with AUC 0.68–0.75 on the three MMLU workloads and 0.60 [0.56, 0.64] on the two-adapter pair traces; restricted to comparisons within the same MMLU subject it is 0.68 and exceeds 0.5 in 56 of 57 subjects, as detailed in Section B. The decline recurs with an independently trained Llama-3-8B pool and with a different prompt in Table 7, and with the freshly trained Qwen pool of Section 4.5.

What the margin tracks is how close the gold option already is. For a base error the gold option is not on top, so the gap between the top and the gold option is at least the margin; this label-aware gold gap

Table 2: Base margin predicts not only base error but whether the adapter pool can fix it. Quartiles of the base margin, Q1 lowest. h is the share of base errors that at least one adapter answers correctly, so that $C_>$ factors as base error times h, with 95% Wilson intervals; the pair traces use item-cluster bootstrap intervals, since each item appears in up to three pairs, and the Wilson interval over distinct items for a zero count. Chg_W and Chg_C are the mean rates, over the four adapters, at which an adapter's prediction differs from the base prediction on base-wrong and base-correct items. AUC is the probability that a rescued base error has a lower base margin than an unrescued one, with a 95% bootstrap interval clustered by item for the pair traces, whose records share items across pairs. The three MMLU workloads use base plus four adapters; the pair traces use base plus the two pair adapters. With Mann–Whitney pairs formed only inside each MMLU subject and a subject-cluster bootstrap, the AUC is 0.68 [0.58, 0.73] on full MMLU, above 0.5 in 56 of 57 subjects, and 0.63 [0.56, 0.76] on the 1,500-item subset subjects.

Workload	Q	Base err. (%)	#wrong	h (%) [95% CI]	Chg_W (%)	Chg_C (%)
Full MMLU, 14,042 AUC 0.68 [0.67, 0.70]	Q1	52.7	1795	49.5 [47.2, 51.8]	37.7	29.69
	Q2	24.8	895	26.3 [23.5, 29.2]	13.0	6.36
	Q3	6.7	237	16.5 [12.3, 21.7]	6.1	0.90
	Q4	1.4	49	2.0 [0.4, 10.7]	0.5	0.04
MMLU-CF, 2,000 AUC 0.75 [0.70, 0.79]	Q1	63.3	305	36.7 [31.5, 42.3]	34.3	36.30
	Q2	43.6	223	15.2 [11.1, 20.6]	8.6	7.18
	Q3	18.8	95	3.2 [1.1, 8.9]	0.8	0.67
	Q4	13.2	66	0.0 [0.0, 5.5]	0.0	0.00
1,500-item subset AUC 0.68 [0.62, 0.74]	Q1	54.6	195	45.6 [38.8, 52.6]	37.2	29.17
	Q2	20.6	81	18.5 [11.6, 28.3]	8.0	4.49
	Q3	4.5	17	0.0 [0.0, 18.4]	0.0	0.63
	Q4	1.1	4	0.0 [0.0, 49.0]	0.0	0.07
Pair traces, 10,083 AUC 0.60 [0.56, 0.64]	Q1	53.8	1338	34.7 [30.7, 38.6]	–	–
	Q2	29.0	741	25.6 [20.5, 31.0]	–	–
	Q3	9.6	243	28.0 [18.8, 37.7]	–	–
	Q4	0.7	18	0.0 [0.0, 39.0]	–	–

Table 3: Answer changes, items on which any adapter differs from base, and rescues. Below: percentage below the primary threshold. The 400-item change-share estimate is median [10th, 90th percentile] across 200 draws. Quartile profiles are in Table 10.

	Count		Below (%)		400-item
Workload, pool	Changes	Rescues	Changes	Rescues	estimate (%)
MMLU, original	3454	1163	82.1	87.1	81.6 [77.1, 86.7]
MMLU-CF, original	465	149	91.8	91.9	91.8 [88.0, 95.2]
MMLU subset, original	317	104	86.8	95.2	86.6 [82.8, 90.5]
MMLU, fresh	3354	1055	82.6	91.7	82.6 [77.6, 86.8]
MMLU-CF, fresh	441	151	90.5	94.0	90.7 [87.1, 94.4]

predicts rescue with AUC 0.76–0.86 on the three zero-shot workloads, and pooling Mann–Whitney comparisons formed within gold ranks, a low margin still predicts rescue with AUC 0.72–0.81 in Table 8. The margin thus serves as a label-free proxy for the distance an adapter must bridge. Section C shows that the decline holds for each of nine adapters on its own, with median relative score shifts from 0.8 to 4.8, and for a control that keeps each adapter's per-item shift sizes but assigns them to random options; we therefore read the margin as locating where shifts of the adapters' size can change the base answer. What the adapters learned shows in realized accuracy: selectors on the real pools exceed the same selectors on this control by 0.41–3.21 points with averaging, vote or confidence, and by $+0.48$ $[+0.17, +0.79]$ points with the screened logistic selector on the original pool in Table 43.

Answer changes locate rescues without labels. A rescue requires some adapter to change the base answer, and whether an adapter changes it needs no label. Below the primary threshold of Section 4.4 lie 82.1, 91.8 and 86.8% of the items on which the original pool changes a base answer on full MMLU, MMLU-CF and the 1,500-item subset, and 82.6 and 90.5% for the fresh pool, against 87.1, 91.9, 95.2, 91.7 and 94.0% of the rescues in Table 3. On each of these workloads the label-free share

Table 4: Screening with the threshold $\tau_0(0.05) = 1.68$ frozen from the 400-item anchor fit. "Above" items would skip adapter evaluation. Oracle gaps are over base plus the four adapters, and over base plus the two pair adapters for the six-pair records, with one record per distinct item and pair: 10,083 records of 3,361 items.

Workload	n	Base acc.	Oracle gap	% above	Gap above	Gap below	Rescues below
Six-pair records, item and pair	10,083	0.768	+7.16	67.3%	+2.51	+16.73	76.5%
Full MMLU test, 57 subjects	14,042	0.788	+8.28	67.1%	+1.59	+21.94	87.1%
MMLU-CF	2,000	0.655	+7.45	60.6%	+0.99	+17.36	91.9%
MMLU subset, 51 subjects	1,500	0.802	+6.93	66.7%	+0.50	+19.80	95.2%

lies 0.1–9.0 points below the rescue share and orders the margin quartiles as h does, an empirical regularity of the tested pools; 400 unlabeled items estimate the change share to within about 5 points.

4.3 Calibration Fit and Frozen Holdout

Fitting $\mathbb{P}(\text{base wrong} \mid m \geq \tau) = \delta e^{-\alpha\tau}$ on the original labels of the 400-item science+medical anchor, a separate evaluation of the seed-1 questions, gave $\delta = 0.2092$, $\alpha = 0.8537$, $R^2 = 0.982$ and $\tau_0(0.05) = 1.68$, shown in Figure 1a. With these original labels, realized oracle gain decays faster than the ceiling requires, with $\delta_C = 0.032$, $\alpha_C = 2.74$ and $R^2 = 0.958$, an empirical property of the pool. The exponential, logistic, and stretched-exponential forms of Table 24 imply $\tau_0(0.05)$ values within 0.032 of each other on each dataset we fit.

Frozen-parameter holdout. Fitted on a random half of the anchor items, the frozen curve predicts the other half's tail error with median $R^2 = 0.892$ over 200 halves, with 10th and 90th percentiles 0.684 and 0.952; split by item on the 3,361 all-pair items, the median is 0.953. Across four Qwen3.5 scales, Qwen3-8B and Llama-3-8B the tail error has the same monotone shape with setting-dependent parameters, as Tables 22 and 23 show.

4.4 Screening Transfers to Independent Workloads

We retain the originally frozen $\tau_0(0.05) = 1.68$ and apply it without refitting to three workloads scored with the same base model and the four adapters and summarized in Table 4: all 14,042 MMLU test rows, 2,000 MMLU-CF items (Zhao et al., 2025), and a separately scored sample of 1,500 MMLU test items from 51 subjects that omit six of the ten history and philosophy training subjects; as a subset of full MMLU it is a subject-restricted check rather than an independent workload, and without its 88 items from the other four training subjects 95.8% of its rescues still lie below the threshold. On each workload, 61–67% of items lie above the threshold, and those items hold at most $+1.59$ points of oracle gain, below the $\epsilon = 5$-point level used to set the threshold, compared with $+17.4$ to $+21.9$ points below it; 87–95% of all oracle-rescued items fall below the threshold. Evaluating the four adapters only below the threshold cuts candidate forward passes per item from five to 2.3–2.6 while giving up at most 1.1 points of overall oracle gain. Cost is counted in candidate passes and benefit in oracle rescues, the opportunity left to the selectors of Section 4.7.

The screen does not depend on our parametric curve: from the same 400 labels, selection with guaranteed risk (Geifman & El-Yaniv, 2017) at target risk 0.10, with bound 0.098 for the original recorded labels only, and the anchor's 30% margin quantile screen comparably, keeping 82–92% of rescues in Table 11; and at equal coverage the margin ranks rescues at least as well as the maximum option probability or the entropy in Table 12.

4.5 Prospective Transfer to a Freshly Trained Pool

The analyses above reuse the pool whose records informed the method. We therefore registered a new pool before training it, with the threshold, baselines and endpoints frozen as described in Section I: four adapters with a domain mix disjoint from the original one, namely law with 3 MMLU subjects, economics and business with 7, mathematics and statistics with 5, and computer science and engineering with 5, each trained with the recipe of the history and philosophy adapters on the 83–209 MMLU validation and dev questions of its subjects, with one seed and the final checkpoint,

Table 5: Cost–headroom tradeoff at the original frozen threshold $\tau_0(0.05) = 1.68$, before anchor-label correction. Passes: out of five; kept: percentage of oracle rescues, with 95% item-bootstrap intervals; loss: overall oracle gain in points. All workloads and the 200-half calibration variability appear in Table 13.

Workload, pool	Passes	Rescues kept (%)	Gain lost
Fresh MMLU	2.32	91.7 [89.9, 93.3]	0.63
Fresh MMLU-CF	2.58	94.0 [90.4, 97.7]	0.45
Original MMLU	2.32	87.1 [85.2, 89.2]	1.07
Original MMLU-CF	2.58	91.9 [87.5, 96.2]	0.60

and scored on all MMLU test and MMLU-CF items. Again the base model is the most accurate single model, with 78.8% against 75.9–78.3% on MMLU, the new pool holds 7.51 points of oracle headroom on MMLU and 7.55 on MMLU-CF, and 327 of its 1,055 MMLU rescues are items that the original pool does not fix. Its rescue rate among base errors falls from 47.7% [45.4, 50.0] in the lowest-margin quartile to 0.0% in the highest on MMLU and from 39.0% to 0.0% on MMLU-CF, as the hatched bars of Figure 1b show, and a low margin predicts rescue among base errors with AUC 0.72 [0.70, 0.74] and 0.77 [0.73, 0.81]. The frozen threshold $\tau_0(0.05) = 1.68$ sends 32.9% and 39.5% of the items to the adapters, 2.32 and 2.58 of five passes, and keeps 91.7% [89.9, 93.3] and 94.0% [90.4, 97.7] of the new pool's rescues, forgoing 0.63 and 0.45 points of oracle gain in Table 5; this meets the effect sizes registered as the target: at least 85% kept with at most 40% of items evaluated and 1–1.5 points lost. These base rows come from the original evaluation; in the registered re-scoring of Section I on all items in the new adapters' configuration, the base model changes 0.9% and 0.75% of its answers, all with margins of at most 0.25, and the target still holds with 91.5% [89.8, 93.2] and 94.0% [90.3, 97.4] kept and 1,024 of 1,055 and 146 of 151 rescues shared. Refitted on 200 halves with the original labels, it never gives a poor operating point, and it keeps at least 80% of the rescues in 46 of the 50 MMLU subjects with five or more. Matched-coverage maximum probability and entropy, SGR and the 30% quantile keep 86.0–94.7% of the rescues in Table 19, so the transfer rests on base confidence rather than on the fitted curve. Pooling the new and the original adapters, eight in total, raises the MMLU headroom to 10.61 points, and the same threshold keeps 86.4% of the 1,490 rescues at 3.63 of nine passes.

4.6 Stability of the Cost–Headroom Tradeoff

Calibration sample. With the original labels, thresholds from 200 random anchor halves applied to the original pool's workloads give a 10th percentile of rescues kept of at least 81.9% and a 90th percentile of gain lost of at most 1.50 points in Table 13; Section G reports other budgets and rules.

Post-hoc correction of eight numeric anchor labels recovers six correct original base predictions and lowers the fitted threshold to 1.536. On the fresh pool it keeps 90.0% and 92.1% at 2.25 and 2.50 passes; Section H reports all nine targets under both label versions alongside the original frozen experiment.

Training-text overlap. All 14,400 records match the gold answers of their reconstructed source rows, as the provenance audit of Section L shows. SciQ and Medical Meadow share no 13-gram with any MMLU or MMLU-CF question; the questions behind the history and philosophy adapters share an exact question or a 13-gram with 1,124 MMLU test items, 895 of them via the fixed moral-scenarios template. Removing them leaves 7.32 points of headroom, a rescue AUC of 0.72, and 93.3% of rescues kept.

Rescored ARC-Challenge. Rescored over each item's full option set as described in Section L, the base model again leads the single models on ARC-Challenge and the oracle gains 2.30 points; on the 972 items outside the calibration questions, a low margin predicts rescue with AUC 0.75 [0.60, 0.87], and the frozen threshold keeps 19 of the 22 rescues at 1.39 passes per item, and 24 of 27 on all 1,172 items.

Another model family. On Llama-3-8B with its own four adapters, a threshold fitted on the 400 science and medical items and fixed keeps all 91 rescues on the 400 distinct history and philosophy

items, and the threshold fitted on history and philosophy keeps all 51 rescues on the science and medical items, sending 86% and 57% of the items to the adapters at base accuracies of 61–65%, after positional label correction; Section K separates this post-hoc reanalysis from the original thresholds.

Pool growth. Adding code and math adapters on the 1,500-item subset raises the rescues from 104 to 123; the unchanged threshold keeps 117 of them at 3.0 of seven passes, including 18 of the 19 new ones.

4.7 Realized Selectors Under the Screen

A random-forest router on score features, trained on the six-pair records under 5-fold cross-validation grouped by item as described in Section J.7, improves accuracy by $+1.02$ points, 95% CI $[0.28, 1.81]$, of the 7.16-point ceiling; Proposition 5 assigns the remainder to 3.49 points of missed gains and 2.65 of damage in Table 30.

Fixed selectors with and without the screen. In registered comparisons on full MMLU in Tables 40 to 42, a per-candidate logistic selector trained on the labels of the other folds, with 5-fold cross-validation that keeps repeated questions in one fold, gains $+0.46$ $[+0.14, +0.78]$ points with all five candidates and $+0.43$ $[+0.11, +0.74]$ when it runs only below the frozen threshold, at 2.32 passes per item, and $+0.39$ and $+0.36$ with item-level folds; screened, it exceeds the best single adapter by $+0.60$ $[+0.30, +0.91]$, and over the eight-adapter union it gains $+0.56$ $[+0.24, +0.87]$ at 3.63 of nine passes. The candidate with the best training-fold accuracy is the base model in every fold, so these gains are also gains over a training-selected fixed candidate in Table 44; by Proposition 5, the screened selector realizes 1.94 of the 8.28 points of oracle gain on the original pool, with 1.52 points assigned to damage in Table 45. On the fresh pool, skipping the adapters above the threshold raises the accuracy of score averaging and plurality vote by $+0.15$ $[+0.06, +0.23]$ and $+0.43$ $[+0.29, +0.57]$ points, avoiding damage on 30 and 79 items and forgoing gains on 9 and 19. Over 20 registered question-group fold assignments in Table 46, the screened selector gains $+0.32$ to $+0.46$ points on the original pool and $+0.46$ to $+0.59$ on the union, with the interval above zero each time, and the screen costs at most 0.06 points on all three pools. Trained on only 1,000 labels, it gains $+0.32$ $[+0.05, +0.59]$ and $+0.41$ $[+0.13, +0.69]$ and exceeds a fixed candidate selected from them by $+0.54$ and $+0.64$ in Table 47. With all training labels, a one-candidate subject lookup gains $+0.63$ and $+0.61$ in Table 42, a strong inexpensive comparator; from measured pass ratios in Section J.9, estimated scoring-only costs on the original pool are 2.51 base passes per item screened, 5.59 unscreened and 1.09 for the lookup.

5 Discussion and Conclusion

Base confidence localizes adapter complementarity in the tested MCQ pools, including a pool trained after the screening threshold was fixed.

Practical recipe. Label a small calibration sample, set a threshold from the fitted curve, SGR or a quantile, and evaluate adapters only below it. The original frozen screen keeps 86–95% of rescues, 92–94% on the fresh pool, at 1.4–2.6 instead of five candidate passes, and the threshold refitted after the label correction keeps 90.0% and 92.1% of the fresh pool's rescues at 2.25 and 2.50 passes. The label budget is best set against explicit targets for retention and lost gain; with 200 or more original labels, the median retention of refitted thresholds equals that of the full anchor on all three workloads in Table 14. The screen lowers the cost of evaluating a pool relative to running every candidate, and costs the selectors behind it at most 0.06 points in Table 46; where subject metadata exist, a one-candidate subject lookup is a strong inexpensive comparator. Calibration labels are spent per base model and prompt, separately from selector-training labels; a new adapter pool reuses the screen.

Scope. We test MCQ logprob scoring with a fixed letter interface and same-base pools on Qwen3.5-9B and Llama-3-8B, not free-form generation. Adapter relative score shifts are 0.8–4.8 and MMLU accuracies within 3.5 points of base; larger shifts or in-domain gains may relocate rescues. Each fresh adapter is trained with one seed, and the evidence across pools comes from two Qwen pools and one Llama pool. The label-free change profile can guide a labeled check. Merges and soft compositions are further candidates as in Section 3.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to support qualitative and thematic data analysis, and formulating questions for surveys or interviews and transcribing recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and no personal data. All experiments use publicly released models, namely Qwen3.5 at four scales, Qwen3-8B and Llama-3-8B, and public datasets: MMLU, MMLU-CF, ARC-Challenge and MedMCQA for evaluation, and SciQ, Medical Meadow medical flashcards, CodeAlpaca-20k, MathInstruct and MMLU validation and development questions for adapter training. We release per-item scores and model outputs together with analysis code, with selected public ARC questions and options for verifying label mappings; full datasets are obtained from their original releases. The medical adapter and the medical evaluation items serve only as a benchmark domain; nothing in this work is intended for clinical use, and the adapters are not evaluated for safety or fairness in any deployment setting. The method screens where an existing adapter pool can change a model's answers and reduces evaluation cost; we see no direct harmful application beyond those of the underlying models, whose known limitations, including biases inherited from their training data, apply unchanged.

REPRODUCIBILITY STATEMENT

The formal setup, the margin ceiling, the router decomposition and the perturbation condition are stated in Section 3, with complete proofs in Section O. The models, adapters, workloads and scoring protocol are described in Section 4 and Section N, the oracle quantities in Section M, the question provenance and overlap audit in Section L, the label corrections in Sections H and K, and the registered analyses with their endpoints in Section P. The anonymous supplementary material provides the per-item option scores of every candidate, the analysis code that recomputes the reported statistics from them under the original and the corrected labels, a table index, the registration documents, and the training, scoring and data-preparation code of the freshly trained pool.

REFERENCES

Oliver Bolton, Aakanksha, Arash Ahmadian, Sara Hooker, Marzieh Fadaee, and Beyza Ermis. SimMerge: Learning to select merge operators from similarity signals. *arXiv preprint arXiv:2601.09473*, 2026.

Sahil Chaudhary. Code alpaca: An instruction-following LLaMA model for code generation. `https://github.com/sahil280114/codealpaca`, 2023.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=cSimKw5p6R`.

Mei-Yen Chen, Thi Thu Uyen Hoang, Michael Hahn, and M. Saquib Sarfraz. When does LoRA reuse work? theoretical limits and mechanisms for recycling LoRAs without data access. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=lVqUJlsnRy`.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12987–13010, 2025.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KgaBScZ4VI.

Jiayi Han, Liang Du, Hongwei Du, Xiangguo Zhou, Yiwen Wu, Yuanfang Zhang, Weibo Zheng, and Donghong Han. SLIM: Let LLM learn more and forget less with soft LoRA and identity mixture. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4792–4804, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.246. URL https://aclanthology.org/2025.naacl-long.246/.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. MedAlpaca – an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

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*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *Conference on Language Modeling*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792.

Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems*, volume 36, pp. 9891–9906, 2023. doi: 10.52202/075280-0431.

Varun Kotte. UCCI: Calibrated uncertainty for cost-optimal LLM cascade routing. *arXiv preprint arXiv:2605.18796*, 2026.

Haokun Liu, Gyung Hyun Je, Marco Ciccone, Zhenlin Xu, Prasanth YSS, and Colin Raffel. The appeal and reality of recycling LoRAs with adaptive merging. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. In *Advances in Neural Information Processing Systems*, volume 37, pp. 78905–78935, 2024. doi: 10.52202/079017-2505.

Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. Learning to route among specialized experts for zero-shot generalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36829–36846, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Oleksiy Ostapenko, Zhan Su, Edoardo Ponti, Laurent Charlin, Nicolas Le Roux, Lucas Caccia, and Alessandro Sordoni. Towards modular LLMs by building and reusing a library of LoRAs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38885–38904, 2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.39. URL https://aclanthology.org/2021.eacl-main.39/.

Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. LoRA soups: Merging LoRAs for practical skill composition tasks. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 644–655, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.55/.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 2026-05-06.

Adir Rahamim, Asaf Yehudai, Boaz Carmeli, Leshem Choshen, Yosi Mass, and Yonatan Belinkov. Will it merge? On the causes of model mergeability. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26551–26570, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1322. URL https://aclanthology.org/2026.findings-acl.1322/.

Derek Tam, Yash Kant, Brian Lester, Igor Gilitschenski, and Colin Raffel. Realistic evaluation of model merging for compositional generalization. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=j7ye0nXvEm.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on*

Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *The Twelfth International Conference on Learning Representations*, pp. 47302–47318, 2024. URL https://openreview.net/forum?id=uWvKBCYh4S.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115, 2023. doi: 10.52202/075280-0310.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=9sbetmvNpW.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 2024.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13371–13391, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.656.

Luca Zhou, Bo Zhao, Rose Yu, and Emanuele Rodolà. Demystifying mergeability: Interpretable properties to predict model merging success. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uFQwehcepH.

Table 6: Zero-shot results on the 14,042 uniquely indexed test rows of all 57 MMLU subjects. Fixed-arm accuracy is recomputed from dataset gold labels. Oracle rows use per-item correctness labels and do not represent a deployed router.

Candidate or rule	Correct / 14,042	Accuracy (%)	Δ (pp)
Base	11,066	78.8064	+0.0000
History adapter	10,583	75.3668	−3.4397
Medical adapter	11,042	78.6355	−0.1709
Philosophy adapter	10,868	77.3964	−1.4101
Science adapter	10,990	78.2652	−0.5412
Oracle: four adapters	12,163	86.6187	+7.8123
Oracle: base + four adapters	12,229	87.0887	+8.2823

A Full-MMLU Complementarity

Table 6 evaluates Qwen3.5-9B and its four adapters zero-shot on all 14,042 MMLU test rows across 57 subjects. The base model answers 11,066 rows or 78.81% correctly; the strongest single adapter, the medical one, answers 11,042 or 78.64%. An oracle over the four adapters alone answers 12,163 rows or 86.62%, and including the base model among the candidates yields 12,229 or 87.09%, a gain of $+8.28$ points that rescues 1,163 base errors. The base model is the most accurate single model in aggregate accuracy, and the pool contains substantial per-item complementarity. These are label-aware ceilings; the row count reflects benchmark coverage, which includes some repeated content. Ties in option scores are resolved to the first maximizing option in the order A–D; 985 of the 70,210 candidate–item records have zero margin, 244 of them for base.

B Rescue Rate Among Base Errors

Table 2 factors the tail complementarity into base error and the rescue rate h, using the per-item scores of the base model and the four adapters on the three workloads of Table 4 and of the two pair adapters on the six-pair traces. This analysis of the original pool is exploratory; Section 4.5 tests it prospectively on a freshly trained pool. Three observations support the reading in Section 4.2. First, h decreases from the lowest to the highest margin quartile on every workload, so a confident base error is rarely one that the pool can fix. Second, the effect is not a subject-mix artifact: when the Mann–Whitney comparison is restricted to pairs of items from the same MMLU subject, the AUC for predicting rescue from a low margin remains 0.68 [0.58, 0.73] with a subject-cluster bootstrap interval and exceeds 0.5 in 56 of 57 subjects. Third, the Chg columns show that adapters rarely move a confident base prediction at all, consistent with the necessary condition of Lemma 6 that an adapter changes the answer only when its score shift reaches half the base margin: in the top quartile of full MMLU, one base error is rescued and five base-correct items are broken by at least one adapter. The same columns show that exposure to damage is concentrated at low margin as well: in the lowest quartile the mean adapter changes 29.7% of base-correct answers, which is why Proposition 5 scores a selector over the pool by its damage as well as its gains. Concentration is stronger for fixable errors than for errors: the lowest-margin 30% of items hold $71/54/73\%$ of base errors but $85/84/94\%$ of rescues on full MMLU, MMLU-CF, and the 1,500-item MMLU subset. On the two-adapter pair traces the item-clustered AUC is 0.60 [0.56, 0.64].

Independent pool and prompt. Table 7 repeats the analysis on the position-corrected seed-1 records of the Llama-3-8B six-pair replication, whose four adapters were trained independently of the Qwen adapters, with $2{,}400$ records of 800 items, and on full MMLU scored with a 5-shot prompt by the same Qwen3.5-9B base model and adapters. In both, h falls monotonically from the lowest to the highest margin quartile: from 48.3% to 0.0% for the Llama pairs with AUC 0.71 [0.67, 0.76], from 67.2% to 0.0% for the four-adapter Llama oracle, and from 55.9% to 6.2% with AUC 0.68 under the 5-shot prompt.

What the margin measures. For a base error, the gold option is not the top option, so the gold gap, the top score minus the gold score, is at least the margin. Table 8 shows that this label-aware gap predicts rescue: on full MMLU h is 49.2% when the gold option is the base model's runner-up

Table 7: Rescue rate h by base-margin quartile, Q1 lowest, on an independently trained adapter pool and under a different prompt. Llama-3-8B: original and position-corrected seed-1 records of the six-pair replication of Table 35; the four-adapter row pools each item's three pairs. Qwen3.5-9B 5-shot: the same base model and adapters as the main workloads with a 5-shot prompt. AUC: low margin predicting rescue among base errors, with 95% bootstrap intervals clustered by item for the pair records.

Pool and prompt	Records	Oracle gap (pts)	h (%) Q1 / Q2 / Q3 / Q4	AUC [95% CI]
Llama six pairs, original	2,400	12.0	48.3 / 23.2 / 12.5 / 0.0	0.72 [0.67, 0.77]
Llama four adapters, original	800	17.6	67.2 / 37.4 / 20.8 / 0.0	0.75 [0.69, 0.80]
Llama six pairs, corrected	2,400	12.12	48.3 / 24.9 / 13.3 / 0.0	0.71 [0.67, 0.76]
Llama four adapters, corrected	800	17.75	67.2 / 39.6 / 22.2 / 0.0	0.74 [0.69, 0.80]
Qwen full MMLU, 5-shot	14,042	10.2	55.9 / 33.9 / 18.0 / 6.2	0.68 [0.66, 0.70]

Table 8: What the margin measures among base errors. Rescue rate h in % by gold rank, the position of the gold option in the base ranking with 2 for the runner-up; a gold option tied at the top counts as rank 1 and is not shown. AUC for predicting rescue from a low margin, from a small label-aware gold gap, the top minus the gold score, and from the margin pooling Mann–Whitney pairs formed within gold ranks, with 95% item-bootstrap intervals.

Workload	Base errors	h by gold rank 2/3/4	AUC margin	AUC gold gap	AUC margin within rank
Full MMLU	2976	49 / 26 / 16	0.68	0.76	0.72 [0.70, 0.74]
MMLU-CF	689	34 / 10 / 2	0.75	0.86	0.81 [0.77, 0.85]
1,500-item subset	297	45 / 16 / 14	0.68	0.79	0.74 [0.68, 0.80]
Full MMLU, 5-shot	3223	55 / 36 / 16	0.68	0.75	0.69 [0.67, 0.71]

and 16.1% when it is ranked last, and the gold gap alone predicts rescue with AUC 0.76–0.86 on the three zero-shot workloads. Pooling Mann–Whitney comparisons formed within gold ranks, the margin still predicts rescue with AUC 0.72–0.81, and margin and gold gap are strongly correlated among base errors, with $r = 0.85$ on full MMLU. Adapters thus tend to fix base errors whose correct answer the base model already nearly prefers, and the margin is a label-free proxy for that distance.

C Adapter Shift Size and Per-Adapter Rescue Profiles

Table 9 describes each adapter by the size of the score shift it induces and repeats the rescue analysis of Table 2 for each adapter on its own on full MMLU. For item x and adapter i, let $\delta_i(x) = f_i(x) - f_0(x)$; the relative shift $D_i(x) = \max_c \delta_i(x)[c] - \min_c \delta_i(x)[c]$ is the largest change an adapter can make to the gap between two options, so by Lemma 6 an adapter can change the base answer only where $D_i(x) \geq m(x)$. The nine adapters span training sets from 83 MMLU questions, repeated eight times, to about 20,000 examples, and their median relative shifts differ by a factor of six, from 0.81 for code to 4.75 for mathematics and statistics. For every adapter, the share of base errors it fixes falls from 16.6–26.5% in the lowest to 0.0–2.0% in the highest margin quartile, and a low margin predicts its rescues among base errors with AUC 0.62–0.77. The localization therefore does not require small shifts: adapters whose median shift exceeds the primary threshold of 1.68 show the same decline.

Magnitude-matched shift control. To separate the size of an adapter's shift from its direction, the control keeps every item's shift vector $\delta_i(x)$ but assigns its four coordinates to the options by a uniformly random permutation, drawn independently per item and adapter in each of 20 draws. The control has exactly the adapters' per-item shift sizes, but its shifts are not aligned with the options. Its rescue rate among base errors declines across the margin quartiles as well: on full MMLU from 54.8% to 4.4% for the original pool and from 62.7% to 4.8% for the fresh pool, and on MMLU-CF from 48.2% to 3.6% and from 56.3% to 7.4%. The profile of h over the margin is thus what shifts of the adapters' sizes produce, whatever their direction, and the margin locates where such shifts can change the base answer. This is the property the screen uses; it makes no assumption about what an adapter has learned, and the registered comparison of rescue counts between the adapters and this control is listed in Section P.

Table 9: Per-adapter shift size and rescue profile of Qwen3.5-9B adapters on all 14,042 MMLU test items, with base accuracy 78.81%. D: relative score shift $\max_c \delta_c - \min_c \delta_c$ against the base scores, median and 90th percentile over items. h: share of base errors that this adapter alone answers correctly, by base-margin quartile from Q1, the lowest, with 1,795, 895, 237 and 49 base errors. AUC: low margin predicting the adapter's rescues among base errors. Rows 1–4 form the original pool, rows 5–8 the freshly trained pool of Section 4.5; the code adapter is used in the six-adapter pool of Section 4.6.

Adapter	Training data	Acc. (%)	Median D	P90 D	Rescues	h Q1	h Q2	h Q3	h Q4	AUC
Science	SciQ train, 11,679	78.3	1.12	2.06	376	19.1	3.8	0.0	0.0	0.76
Medical	Medical Meadow, 19,765	78.6	1.00	1.75	336	17.7	2.1	0.0	0.0	0.77
History	145 MMLU val.+dev q.	75.4	4.38	6.75	578	25.6	12.5	2.5	2.0	0.66
Philosophy	229 MMLU val.+dev q.	77.4	3.50	5.50	653	26.5	15.9	14.8	0.0	0.62
Law	209 MMLU val.+dev q.	75.9	2.87	4.44	533	23.7	10.9	3.8	0.0	0.67
Economics/business	194 MMLU val.+dev q.	76.3	3.63	5.69	551	25.8	9.4	1.3	0.0	0.70
Mathematics/statistics	140 MMLU val.+dev q.	78.3	4.75	6.87	435	21.7	5.0	0.4	0.0	0.74
CS/engineering	83 MMLU val.+dev q.	76.6	3.88	5.62	499	23.6	8.4	0.4	0.0	0.69
Code	CodeAlpaca-20k	78.0	0.81	1.56	324	16.6	2.9	0.0	0.0	0.76

D Answer Changes and Rescues

A rescue is a base error that some adapter answers correctly, so every rescued item is also an item on which some adapter predicts a different option than the base model. Whether an adapter changes the base answer needs no label, so the margin profile of these answer changes can be read from adapter outputs on unlabeled items. Table 3 compares, at the primary threshold $\tau_0(0.05) = 1.68$, the share of answer changes below the threshold with the share of rescues below it, which are the rescues kept by the screen, and the change rate by base-margin quartile with the rescue rate h. On all five workloads the share of changes below the threshold is lower than the share of rescues, by 0.1 to 9.0 points, and the change rate falls across the quartiles in the same order as h, with a Spearman correlation of 0.95–1.00 over the four quartiles. That rescues are a subset of changes does not by itself order the two shares, so the lower change share is an empirical regularity of these pools, observed on every workload we measured. Computed from 400 randomly drawn unlabeled items in each of 200 draws, the change share stays within about 5 points of its full-workload value between the 10th and 90th percentiles; this is the sampling spread of the change share, and where the two full-workload shares nearly coincide, as for the original pool on MMLU-CF, a single draw can lie on either side of the rescue share. For a new pool, adapter outputs on a few hundred unlabeled items thus show, before any label is collected, where its answer changes lie; on the tested pools this profile placed the rescues at least as far below the threshold.

Table 10: Quartile profiles for the workloads of Table 3. Change rate is the percentage of all items whose answer differs for some adapter; h conditions on base errors.

Workload, pool	Change rate (%) Q1/Q2/Q3/Q4	h (%) Q1/Q2/Q3/Q4
MMLU, original	70.7 / 24.5 / 4.5 / 0.2	49.5 / 26.3 / 16.5 / 2.0
MMLU-CF, original	69.9 / 22.3 / 2.8 / 0.0	36.7 / 15.2 / 3.2 / 0.0
MMLU subset, original	67.2 / 17.0 / 2.4 / 0.3	45.6 / 18.5 / 0.0 / 0.0
MMLU, fresh	68.8 / 23.8 / 4.3 / 0.1	47.7 / 20.8 / 5.5 / 0.0
MMLU-CF, fresh	66.8 / 19.1 / 4.0 / 0.2	39.0 / 13.0 / 3.2 / 0.0

E Threshold Rules Under a Common Label Budget

This section preserves the original-label analysis; corrected-label thresholds are in Section H. Table 11 compares the fitted-curve threshold with two alternatives computed from the same 400 recorded anchor labels: selection with guaranteed risk (Geifman & El-Yaniv, 2017), abbreviated SGR, which binary-searches for a margin threshold whose base selective risk is certified below a target at confidence 0.95, run at target 0.10, and the anchor's 30% margin quantile. SGR returns margin 1.37 with certified bound 0.098, practically the same threshold as the 30% quantile. All three rules localize headroom on the three workloads: the fitted threshold is slightly more conservative, skipping fewer

items and keeping more rescues, and the lower thresholds skip more adapter passes at the cost of 0.2–0.7 additional points of oracle gap above the threshold.

Exact operating points. Every rule sends an item to the adapters if and only if its base margin is strictly below the threshold; an item whose margin equals the threshold is skipped. Option scores are computed in bf16, so margins take a small set of values, and floating-point differences between scores make nominally equal margins differ in the last digits. We therefore state the thresholds at full precision: $\tau_0(0.05) = 1.6767057312$, which no margin in our data equals; the SGR threshold $\theta_{\mathrm{SGR}} = 1.3749999702$; and the 30% quantile $\theta_{q30} = 1.3749999404$. On full MMLU, 239 items have a margin within 10^{-6} of 1.375; 30 of them lie in $[\theta_{q30}, \theta_{\mathrm{SGR}})$ and are skipped by the quantile rule but evaluated by SGR, against 3 on MMLU-CF and 1 on the 1,500-item subset, which accounts for the small differences between the SGR and quantile rows of Tables 11 and 19. With the values above, the released code reproduces every operating point in these tables exactly.

Scope of the SGR certificate. Under the i.i.d. calibration assumption, SGR gives a 0.95-confidence bound of 0.098 on selective risk measured against the original recorded labels. Those labels mark 15 of the 279 items above θ_{SGR} as errors. Because eight anchor labels were mismatched, this original certificate is not presented as a verified guarantee for corrected ground truth; Section H recomputes the rule after correction. The certificate says nothing about other distributions: above the same threshold the base error is 9.8% on full MMLU, 21.4% on MMLU-CF and 8.0% on the 1,500-item subset. We use SGR only as an alternative way to place a threshold from the same labels and evaluate all rules by the rescues they keep on the target workloads. When SGR certifies no threshold on a calibration sample, which happens on 19 of the 200 halves, we evaluate the adapters on every item; over all 200 halves SGR then keeps a median 87.4% [81.8, 98.5], 91.9% [85.2, 100.0] and 95.2% [92.3, 100.0] of the rescues on the three workloads at a median 2.34–2.61 passes in Table 15, and gives a poor operating point on 11.0% of the halves on full MMLU, against 5.0% for the fitted threshold, and on none on the other two.

Table 11: Three margin thresholds chosen from the same 400 labeled anchor items and applied, frozen, to three workloads with base plus four adapters. SGR is selection with guaranteed risk (Geifman & El-Yaniv, 2017) at confidence 0.95, run at target risk 0.10; it returns a threshold with certified bound 0.098. Columns as in Table 4.

Workload	Rule	Threshold	% above	Gap above	Gap below	Rescues below
Full MMLU, 14,042	Fitted curve, $\tau_0(0.05)$	1.68	67.1%	+1.59	+21.94	87.1%
	SGR, certified risk 0.098	1.37	71.9%	+2.06	+24.20	82.1%
	Anchor 30% margin quantile	1.37	72.1%	+2.07	+24.33	81.9%
MMLU-CF, 2,000	Fitted curve, $\tau_0(0.05)$	1.68	60.6%	+0.99	+17.36	91.9%
	SGR, certified risk 0.098	1.37	66.5%	+1.65	+18.98	85.2%
	Anchor 30% margin quantile	1.37	66.7%	+1.65	+19.07	85.2%
1,500-item subset	Fitted curve, $\tau_0(0.05)$	1.68	66.7%	+0.50	+19.80	95.2%
	SGR, certified risk 0.098	1.37	71.7%	+0.74	+22.59	92.3%
	Anchor 30% margin quantile	1.37	71.7%	+0.74	+22.64	92.3%

F Screening Signals at Equal Coverage

Table 12 compares the top-1/top-2 margin with other base-confidence scores computed from the same option scores: the maximum option probability and the option entropy. At matched coverage the three signals retain similar shares of rescues, and the margin's AUC for predicting a rescue is at least as high as the others' on every workload: 0.855–0.890 on the three MMLU workloads, against 0.846–0.890 for the maximum probability and 0.824–0.877 for the entropy. The screen is therefore a property of base confidence rather than of one particular score; the margin is convenient because it lower-bounds the gold gap of every base error and Lemma 6 ties it directly to how large a score shift an adapter needs to change the answer, and the fitted curve adds a way to place the threshold from a target level.

Table 12: Base-confidence signals compared at equal coverage. AUC: predicting from low confidence, over all items, that an item is rescued by the pool, with the base model wrong and some adapter right. Retained: share of all rescues that fall among the items not skipped when the most confident fraction of items is skipped, for the three signals margin, maximum option probability and option entropy; the first coverage is the fraction above the frozen $\tau_0(0.05) = 1.68$ on that workload, the second skips half of the items. Ties are split proportionally.

Workload	AUC margin	AUC max-prob	AUC entropy	Skipped	Retained at τ_0 coverage	Retained at 50%
Full MMLU	0.864	0.857	0.842	67.1%	87.1 / 87.0 / 85.9	96.6 / 96.6 / 96.8
MMLU-CF	0.855	0.846	0.824	60.6%	91.9 / 92.6 / 91.3	98.0 / 98.0 / 97.3
1,500-item subset	0.890	0.890	0.877	66.7%	95.2 / 94.2 / 94.2	100.0 / 100.0 / 100.0
Six-pair records	0.813	0.812	0.803	67.3%	76.5 / 76.9 / 77.4	90.6 / 91.8 / 92.2

Table 13: Complete cost–headroom analysis with original anchor labels. Top: frozen threshold, point estimate and 95% bootstrap interval. Bottom: refitting on 200 anchor halves, median [10th, 90th percentile]. Poor: retention below 80% or absolute loss above 1.5 points. Overlap removal excludes 1,124 flagged MMLU items; ARC excludes calibration questions. No missing sensitivity value is imputed.

Workload	K	Passes	Rescues kept (%)	Gain lost
Fresh MMLU	4	2.32	91.7 [89.9, 93.3]	0.63
Fresh MMLU-CF	4	2.58	94.0 [90.4, 97.7]	0.45
Original MMLU	4	2.32	87.1 [85.2, 89.2]	1.07
Original MMLU-CF	4	2.58	91.9 [87.5, 96.2]	0.60
MMLU subset	4	2.33	95.2 [90.9, 99.0]	0.33
Subset + code/math	6	3.00	95.1 [91.0, 99.0]	0.40
MMLU, overlap removed	4	2.29	93.3	0.49
ARC, rescored	4	1.39	86.4 [69.6, 100.0]	0.31

Workload	Passes	Rescues kept (%)	Gain lost	Poor
Fresh MMLU	2.32 [2.11, 2.52]	91.7 [86.0, 95.2]	0.63 [0.36, 1.05]	0.0%
Fresh MMLU-CF	2.58 [2.33, 2.81]	94.0 [88.7, 97.4]	0.45 [0.20, 0.85]	0.0%
Original MMLU	2.32 [2.11, 2.52]	87.1 [81.9, 91.1]	1.07 [0.73, 1.50]	5.0%
Original MMLU-CF	2.58 [2.33, 2.81]	91.9 [85.2, 96.6]	0.60 [0.25, 1.10]	0.0%
MMLU subset	2.33 [2.13, 2.53]	95.2 [92.3, 96.2]	0.33 [0.27, 0.53]	0.0%
Subset + code/math	3.00 [2.69, 3.29]	95.1 [92.7, 96.7]	0.40 [0.27, 0.60]	0.0%
MMLU, overlap removed	–	–	–	–
ARC, rescored	–	–	–	–

G Calibration Sensitivity of the Transferred Screen

The following resampling results use the original anchor labels; Section H repeats the samples after correction. Table 13 separates two sources of variation: target-sample uncertainty of the primary threshold, from an item bootstrap on each workload with the calibration fixed, and calibration variability, from thresholds refitted on resampled calibration sets with the target fixed. For the second, the tail-error curve is refitted with the unchanged grid and fitter on every calibration sample and $\tau_0(0.05)$ is applied, frozen, to each workload; no sample is selected or discarded, and no fit failed. Table 14 varies the calibration sample and the label budget, and Table 15 recomputes the alternative rules of Section E on the same 200 halves. Base option scores have bf16 precision, so margins take discrete values and nearby thresholds give identical operating points; this is why several rows share their percentiles. With 200 or more labels the median share of rescues kept equals that of the primary threshold on all three workloads. On the same halves, SGR keeps a similar median share of rescues at a similar number of passes, the 30% quantile skips more items and keeps correspondingly fewer rescues, and the maximum option probability at the fitted threshold's coverage matches the margin; the entropy is slightly lower.

Table 14: Label budget and resampling of the calibration set for the original pool. For each calibration sample the tail-error curve is refitted and $\tau_0(0.05)$ applied, frozen, to the workload with base plus four adapters. Rescues kept: median [10th, 90th percentile] in %; lost: median points of overall oracle gain above the threshold; poor: share of samples keeping under 80% of rescues or losing more than 1.5 points. No fit failed.

	Full MMLU			MMLU-CF			1,500-item MMLU subset		
Calibration samples	Kept	Lost	Poor	Kept	Lost	Poor	Kept	Lost	Poor
200 random halves of 200 labels	87.1 [81.9, 91.1]	1.07	5.0%	91.9 [85.2, 96.6]	0.60	0.0%	95.2 [92.3, 96.2]	0.33	0.0%
1,000 bootstrap resamples of 400	87.1 [81.9, 91.1]	1.07	3.5%	91.9 [85.2, 96.6]	0.60	0.2%	95.2 [92.3, 96.2]	0.33	0.2%
200 subsamples of 300 labels	87.1 [84.1, 88.6]	1.07	0.5%	91.9 [86.6, 94.0]	0.60	0.0%	95.2 [93.3, 95.2]	0.33	0.0%
200 subsamples of 200 labels	87.1 [81.9, 91.1]	1.07	4.0%	91.9 [85.2, 96.6]	0.60	0.0%	95.2 [92.3, 96.2]	0.33	0.0%
200 subsamples of 100 labels	85.9 [79.4, 92.9]	1.17	12.5%	89.3 [85.2, 98.0]	0.80	3.0%	95.2 [91.2, 96.2]	0.33	3.0%

Table 15: Threshold rules for the original pool, recomputed on the same 200 random calibration halves and applied, frozen, to each workload with base plus four adapters. Passes: median candidate passes per item, where "=" marks the coverage of the fitted threshold on that half; kept: median [10th, 90th percentile] share of rescues kept in %; lost: median points of overall oracle gain above the threshold. SGR returned no certified threshold on 19 of the 200 halves; on those halves the adapters are evaluated on every item with five passes, keeping all rescues, and its row summarizes all 200 halves.

	Full MMLU			MMLU-CF			1,500-item MMLU subset		
Rule	Passes	Kept	Lost	Passes	Kept	Lost	Passes	Kept	Lost
Fitted curve, $\tau_0(0.05)$	2.32	87.1 [81.9, 91.1]	1.07	2.58	91.9 [85.2, 96.6]	0.60	2.33	95.2 [92.3, 96.2]	0.33
SGR at target risk 0.10	2.34	87.4 [81.8, 98.5]	1.04	2.61	91.9 [85.2, 100.0]	0.60	2.35	95.2 [92.3, 100.0]	0.33
30% margin quantile	2.12	82.0 [78.3, 84.6]	1.49	2.33	85.2 [83.9, 87.2]	1.10	2.13	92.3 [89.9, 94.2]	0.53
Max. probability, coverage of τ_0	=	87.0 [81.8, 91.0]	1.08	=	92.6 [83.9, 97.3]	0.55	=	94.2 [93.3, 95.2]	0.40
Entropy, coverage of τ_0	=	85.9 [77.2, 91.0]	1.17	=	91.3 [77.9, 96.6]	0.65	=	94.2 [88.5, 95.2]	0.40

H Post-hoc Correction of the Calibration Labels

Original predictions, corrected labels. Eight science questions in the 400-item anchor encode the gold option as a number, whereas the scorer predicts a letter. We map the gold option by position and compare it with the original saved score-vector argmax; no model is evaluated again. The recorded scores reproduce all 400 original base, science and medical correctness records and exactly match the anchor margins. Of the eight affected questions the original base prediction is correct on six, changing anchor accuracy from $320/400$ to $326/400$. Seven records change at least one candidate's correctness; on the eighth all three candidates remain wrong. A later ARC rescoring answers all eight correctly, but it is a separate evaluation and is not substituted for the original predictions.

Frozen experiment versus corrected fit. The threshold 1.6767057312475637 was frozen before the fresh pool was trained. We preserve all results from that experiment as originally run. Label correction is a separate, post-hoc analysis after the target results were known, with the same tail-error grid, fitter, target $\epsilon = 0.05$ and unchanged target scores. It gives $\delta = 0.1918203$, $\alpha = 0.8755305$ and $\tau_0(0.05) = 1.5356814958$. SGR at risk 0.10 changes from 1.3749999702 to 1.2499999702; the label-free 30% quantile remains 1.3749999404. On MMLU the corrected SGR threshold retains 79.8% rather than 82.1% of rescues and loses 1.67 rather than 1.48 points. The historical fitted-curve and holdout statistics elsewhere in the paper describe the original labels, not the corrected calibration. Neither the refitted threshold nor its observed target performance is a new prospective validation.

Sensitivity to the calibration sample. We reuse all 1,803 original calibration samples: three fixed halves, 200 random halves, 1,000 size-400 bootstraps, and 200 subsamples at each budget of 100, 200 and 300. The same indices are used before and after correction. The supplementary material contains thresholds and all nine targets' retention, passes and lost gain for the fitted, SGR and quantile rules; if SGR returns no threshold, its specified fallback evaluates every adapter. For the fitted rule, the bootstrap threshold median moves from 1.660 to 1.514 and its 10th–90th percentile range from [1.359, 2.052] to [1.241, 1.894].

The corrected fit lowers the threshold, so it evaluates fewer items on every target; the share of rescues kept is unchanged on the two subset targets and 1.2 to 2.7 points lower on the other seven. On the

Table 16: All nine released targets, each cell giving the value under the original frozen threshold and then under the post-hoc corrected fit. Original and fresh denote the pool; union is both pools together. "Poor" is the percentage of 1,000 calibration bootstraps with $< 80\%$ rescues kept or > 1.5 points of oracle gain lost. Passes count base plus evaluated adapters. Target predictions and labels are unchanged.

Target	Adapters	Kept (%)	Passes	Lost (pp)	Poor (%)
MMLU, original	4	87.1/85.9	2.32/2.25	1.07/1.17	3.5/11.1
MMLU-CF, original	4	91.9/89.3	2.58/2.50	0.60/0.80	0.2/0.9
MMLU subset	4	95.2/95.2	2.33/2.25	0.33/0.33	0.2/0.9
MMLU subset	6	95.1/95.1	3.00/2.88	0.40/0.40	0.2/0.9
MMLU, 5-shot	4	79.7/77.4	2.17/2.11	2.07/2.30	87.5/94.4
MMLU, fresh	4	91.7/90.0	2.32/2.25	0.63/0.75	0.2/0.9
MMLU, union	8	86.4/84.8	3.63/3.50	1.45/1.62	43.6/66.2
MMLU-CF, fresh	4	94.0/92.1	2.58/2.50	0.45/0.60	0.0/0.3
MMLU-CF, union	8	93.0/90.5	4.16/4.00	0.70/0.95	1.0/3.8

fresh MMLU and MMLU-CF pools it keeps 90.0% and 92.1% at 2.25 and 2.50 passes, and on the original pool 85.9% and 89.3% at the same passes.

I Freshly Trained Pool

Registration. The pool, its training recipe, the frozen screens and the endpoints of Section 4.5 were registered before training started, as the registration document in the supplementary material records: four adapters with one training seed, 42, the final checkpoint, no early stopping and no replacement after results; screens $\tau_0(0.05) = 1.6767$ from the original anchor, SGR and the anchor 30% quantile, both at $\theta = 1.375$, and maximum probability, entropy and random screening at the fitted threshold's coverage; endpoints per workload as reported in Tables 5 and 17 to 19. The effect sizes used as the yardstick in Section 4.5 were registered as well.

Training and scoring. As Table 17 lists, each adapter is trained on the MMLU validation and dev questions of its subjects with the recipe of the history and philosophy adapters: a system prompt naming the domain, each question repeated 8 times, 2 epochs, batch 4×4, learning rate 2×10^{-4} with warmup and cosine decay, bf16, and LoRA rank 16, $\alpha = 32$ and dropout 0.05 on the same 128 modules. None of the new subjects is among the original history and philosophy training subjects, and the test split is never used for training; the overlap audit of Section L flags 0/25, 2/14, 5/7 and 2/6 MMLU test items for the law, economics, mathematics and CS training questions, counting exact matches before the slash and shared 13-grams in the question or question plus gold option after it. Scoring uses the zero-shot letter scorer of the main experiments. Before scoring, we re-score the base model and the original science adapter on every tenth MMLU item, 1,405 in all, and on 200 MMLU-CF items and compare with the original records: predictions agree on 99.0% and 100% of the items for the base model and on 99.3% and 99.5% for the science adapter. The analysis uses the base rows of the original evaluation; the last two paragraphs of this section examine what the re-scoring differences can affect and repeat the analysis with the base model re-scored on all items.

Results. Table 18 gives the rescue rate by margin quartile and Table 19 all frozen screens. The new pool's rescues overlap only partly with the original pool's: of its 1,055 MMLU rescues, 728 are also rescued by the original pool and 327 are not. Per subject, among the 50 MMLU subjects with at least five rescues, the median share kept by the frozen threshold is 97.1%, and in the pool's own subjects it keeps 229 of 250 law rescues and 136 of 142 mathematics and statistics rescues.

Re-scoring stability and a no-change control. The fresh adapters are scored in new evaluations, while the analysis uses the base rows of the original evaluation, so a difference between evaluations could in principle masquerade as an adapter effect. Each of the five new scoring evaluations, for the four fresh adapters and the code adapter, first re-scores the base model and the original science adapter on every tenth MMLU test item, 1,405 in all, and on 200 MMLU-CF items and compares them with the original rows in Table 20. The five re-scorings give identical agreement and deviation statistics,

Table 17: The freshly trained pool. Each adapter is trained on the unique MMLU validation and dev questions of its subjects, each repeated 8 times, with the recipe of the history and philosophy adapters, seed 42, 2 epochs and the final checkpoint, and scored on all 14,042 MMLU test and 2,000 MMLU-CF items; accuracy in %.

Adapter	MMLU subjects	Questions	Steps	Final loss	MMLU	MMLU-CF
Law	professional law, international law, jurisprudence	209	210	0.028	75.94	64.45
Economics/ business	econometrics, HS macro-/microeconomics, marketing, management, prof. accounting, business ethics	194	194	0.055	76.34	64.75
Mathematics/ statistics	abstract algebra, college/elementary/HS mathematics, HS statistics	140	140	0.048	78.33	64.95
CS/engineering	college/HS computer science, computer security, machine learning, electrical engineering	83	84	0.046	76.58	65.35
Base model	–	–	–	–	78.81	65.55
Oracle, base + 4 adapters	–	–	–	–	86.32	73.10

Table 18: Rescue rate h of the freshly trained pool among base errors, by base-margin quartile over all items, Q1 lowest, with 95% Wilson intervals over base errors, and the AUC of a low margin for predicting rescue among base errors with 95% item-bootstrap intervals.

	Full MMLU			MMLU-CF		
Quartile	Base err. (%)	Errors	h (%)	Base err. (%)	Errors	h (%)
Q1	52.7	1795	47.7 [45.4, 50.0]	63.3	305	39.0 [33.7, 44.6]
Q2	24.8	895	20.8 [18.3, 23.6]	43.6	223	13.0 [9.2, 18.1]
Q3	6.7	237	5.5 [3.2, 9.2]	18.8	95	3.2 [1.1, 8.9]
Q4	1.4	49	0.0 [0.0, 7.3]	13.2	66	0.0 [0.0, 5.5]
AUC	0.72 [0.70, 0.74]			0.77 [0.73, 0.81]		

so re-scoring under the fixed scoring configuration is repeatable; against the original rows, the base model changes its prediction on 1.0% of the MMLU check items and on none of the MMLU-CF check items, and the largest change of any option score is 0.374, with a mean per-item maximum of 0.090. This no-change control sets the scale of the variation between evaluations: on the same check items the new adapters change the base prediction on 9.9–14.2% of MMLU items, and the original adapters, scored in the original evaluation, on 7.3–14.4%. A change of at most 0.374 per score moves any gap between two options by at most 0.75, so by Lemma 6 a deviation of this size cannot change the base answer on items whose margin exceeds 0.75: 83.0% of the MMLU and 79.1% of the MMLU-CF items, including every item above the threshold $\tau_0(0.05) = 1.68$. The value $\epsilon^* = 0.374$ is the largest change observed on the check items, not a bound established for every item: for an item with base margin above $2\epsilon^* = 0.748$ whose largest score change is at most ϵ^*, whether the base model errs is the same under either base scoring, while items with smaller margins can change, as on the MMLU check items, where the re-scored base model answers 1,090 correctly against 1,094 in the original rows; such a change can move an item across the threshold only through a margin change of at most 0.75.

Paired re-scoring of the base model on all items. The registered analysis S5-A of Section P re-scores the base model on all 14,042 MMLU and 2,000 MMLU-CF items in the configuration used for the fresh adapters, after the same check, whose statistics equal those of Table 20, and repeats every fresh-pool endpoint with these rows as the base in Table 21. The re-scored base model keeps its answer on 99.1% of the MMLU and 99.3% of the MMLU-CF items, with accuracy 78.75% against 78.81% and 65.55% in both cases; every changed answer has an original margin of at most 0.25. The median per-item maximum score change is 0.093 and 0.096; the largest is 0.81 and 0.58, above

Table 19: Screening rules on the freshly trained pool, all frozen with thresholds from the 400-item anchor of the original study before its outcomes were examined. Passes are out of 5, and "=" marks the coverage of the fitted threshold; kept: share of the new pool's rescues below the threshold in %; lost: points of overall oracle gain above it; halves: median [10th, 90th percentile] share kept when the rule is recomputed on 200 random calibration halves; for SGR all 200 halves count, with every item evaluated on the 19 halves without a certified threshold.

	Full MMLU, 1,055 rescues				MMLU-CF, 151 rescues			
Rule	Passes	Kept	Lost	Halves	Passes	Kept	Lost	Halves
Fitted curve, $\tau_0(0.05) = 1.68$	2.32	91.7	0.63	91.7 [86.0, 95.2]	2.58	94.0	0.45	94.0 [88.7, 97.4]
SGR at target risk 0.10, $\theta = 1.375$	2.12	86.1	1.05	91.8 [85.8, 99.5]	2.34	88.7	0.85	94.0 [88.7, 99.3]
Anchor 30% margin quantile, $\theta = 1.375$	2.12	86.0	1.05	86.0 [82.7, 88.5]	2.33	88.7	0.85	88.7 [87.3, 90.7]
Max. probability, coverage of τ_0	=	90.6	0.71	90.6 [86.4, 94.6]	=	94.7	0.40	94.7 [88.7, 97.4]
Entropy, coverage of τ_0	=	89.9	0.76	89.9 [81.1, 94.1]	=	92.7	0.55	92.7 [82.8, 98.0]
Random, coverage of τ_0, expected	=	32.9	5.04	–	=	39.5	4.57	–

Table 20: Re-scoring of the base model and the original science adapter on the check items, compared with the rows of the original evaluation; all five re-scorings give identical values. Agreement: share of items with the same prediction; max and mean: largest and average per-item maximum absolute change of an option score. Last column: share of check items on which each adapter's prediction differs from the original base prediction.

Check set	Candidate	Agreement (%)	Max change	Mean change	Adapters vs. base (%)
MMLU, 1,405 items	Base	99.0	0.374	0.090	fresh 9.9–14.2
	Science adapter	99.3	0.364	0.091	original 7.3–14.4
MMLU-CF, 200 items	Base	100.0	0.332	0.092	fresh 7.5–13.0
	Science adapter	99.5	0.297	0.086	original 9.0–12.5

the check-set maximum on 15 and 3 items, among which one answer changes, on an MMLU-CF item with original margin 0. The threshold assignment changes for 92 MMLU items, 46 in each direction, so the share below the threshold stays at 32.9%, and for 24 MMLU-CF items, moving the share from 39.5% to 39.7%. With the re-scored rows the fresh pool has 1,072 and 150 rescues, of which 1,024 and 146 are rescues under the original rows as well; the rescue rate by margin quartile, the AUC and the retention at the frozen threshold stay within the intervals of the original analysis in Tables 5 and 18, and the registered target of at least 85% kept, at most 40% of the items evaluated and at most 1.5 points lost holds on both workloads. The same re-scored rows leave the original pool's endpoints within 0.3 points of retention on MMLU and 0.8 points on MMLU-CF; this comparison is descriptive, since those adapters were scored in the original evaluation.

Table 21: S5-A: fresh-pool endpoints with the base rows of the original evaluation and with the base model re-scored on all items in the configuration of the fresh adapters; base accuracy 78.81% and 78.75% on MMLU, 65.55% under both on MMLU-CF. The rescue sets share 1,024 MMLU and 146 MMLU-CF items; 31 and 5 are rescues only under the original rows, 48 and 4 only under the re-scored rows. h Q1→Q4: rescue rate among base errors in the lowest and highest base-margin quartile. AUC: low margin predicting rescue among base errors. Eval.: share of items below the frozen threshold $\tau_0(0.05) = 1.68$; kept: share of rescues below it, with 95% intervals from 1,000 item-bootstrap resamples; lost: oracle gain above it.

Workload	Base rows	Rescues	h Q1→Q4 (%)	AUC	Eval. (%)	Kept (%)	Lost (pts)
MMLU	original	1,055	47.7→0.0	0.72 [0.70, 0.74]	32.9	91.7 [89.9, 93.3]	0.63
	re-scored	1,072	48.1→0.0	0.73 [0.71, 0.74]	32.9	91.5 [89.8, 93.2]	0.65
MMLU-CF	original	151	39.0→0.0	0.77 [0.73, 0.81]	39.5	94.0 [90.4, 97.7]	0.45
	re-scored	150	37.9→0.0	0.77 [0.73, 0.81]	39.7	94.0 [90.3, 97.4]	0.45

J Additional Results

J.1 Calibration Scaling Trend

Tables 22 and 23 give the per-scale base accuracy, margin distribution and tail-error fit summarized in Section 4.3; only the base model is evaluated at the smaller scales.

Table 22: Single-seed calibration scaling across 4 Qwen3.5 model scales on the science+medical anchor items; for 9B this is the 400-item anchor sample of Section 4.3. Adapter-composition transfer is not evaluated in this table.

Model	Base Acc.	Mean Margin	% with $m < 1.0$	δ
Qwen3.5-0.8B	0.410	1.01	63.0%	0.699
Qwen3.5-2B	0.595	1.52	49.3%	0.385
Qwen3.5-4B	0.625	1.25	56.8%	0.378
Qwen3.5-9B	0.800	3.36	24.5%	0.209

Table 23: Tail-error fits across model scales and families on science+medical MCQ items. The Llama row uses position-corrected labels; its original base/alpha/delta were 0.654/0.436/0.352. Other rows retain their original labels.

Model	Family	Base acc.	α	δ	R^2
Qwen3.5-0.8B	Qwen3.5	0.410	0.809	0.699	0.919
Qwen3.5-2B	Qwen3.5	0.595	0.746	0.385	0.980
Qwen3.5-4B	Qwen3.5	0.625	0.934	0.378	0.990
Qwen3.5-9B	Qwen3.5	0.800	0.854	0.209	0.982
Qwen3-8B	Qwen3	0.743	0.333	0.280	0.980
Llama-3-8B	Llama-3	0.660	0.433	0.344	0.990

J.2 Calibration Model Selection and Sample Efficiency

We test whether the exponential tail-error curve is uniquely preferred over simple alternatives. We fit tail error curves $\mathbb{P}(\text{base wrong} \mid m \geq \tau)$ with four functional forms on the same threshold grid and compute the Bayesian information criterion (BIC) from the binomial tail likelihood, on the anchor sample that defines τ_0, on the seed-1 science+medical trace, and on all 36 pooled pair traces, as Table 24 reports. The threshold observations are nested, so the information criteria should be read as a diagnostic rather than as a formal independent-sample likelihood.

The practical threshold recommendation is therefore more robust than the exact parametric family: within each dataset the exponential, logistic and stretched forms imply $\tau_0(0.05)$ values within 0.032 of each other, with 2.75, 2.75 and 2.77 on the pooled traces, and the exponential form has the lowest BIC on both 400-item samples. In practice, users can fit any monotone tail-error curve and read off the margin where the estimated tail error falls below the target ϵ; the exponential form is a compact closed-form instance.

Residuals of the anchor fit. On the 50-point grid of the anchor fit, the empirical tail error lies above the fitted curve at 27 points and on or below it at 23; the largest positive deviation is 0.019, an empirical 0.143 against a fitted 0.124 at margin 0.61, with 335 items above that margin. The fitted curve is therefore a central estimate of the tail error, which is how Equation (3) uses it; the thresholds it implies are evaluated by the rescues they keep on the target workloads in Sections 4.4 and 4.6, and SGR gives a threshold with a certified bound on the calibration distribution in Section E.

We also estimate how many labels are needed to fit a usable $\tau_0(0.05)$ threshold. Across the 18 pair-seed traces of sample seeds 1–3, we repeatedly subsample labeled calibration sets and compare the fitted threshold to the full-trace threshold.

A label budget is therefore chosen against explicit targets for retention, lost gain and the share of poor operating points, those that keep under 80% of the rescues or lose over 1.5 points of oracle

Table 24: Calibration model selection with the binomial tail likelihood on a 30-point threshold grid over $[0, 0.9 \max m]$ with at least 5 items per tail and 100 for the pooled records. Lower BIC is better. The anchor sample is the one whose 50-point-grid fit gives $\tau_0(0.05) = 1.68$ in the main text; on this coarser grid its exponential fit gives 1.71. The seed-1 trace scores the same seed-1 questions in a separate evaluation. Pooled records repeat items across seeds and pairs, so the information criteria are diagnostics, not independent-sample likelihoods.

Data	Model	BIC	ΔBIC	R^2	$\tau_0(0.05)$
Anchor sample, 400 items	Exponential	2333.65	0.00	0.978	1.71
	Power law	2343.77	10.12	0.939	1.62
	Logistic	2346.15	12.50	0.979	1.68
	Stretched exponential	2346.69	13.03	0.980	1.67
Seed-1 science+medical trace	Exponential	2368.62	0.00	0.982	1.76
	Logistic	2380.30	11.68	0.984	1.73
	Stretched exponential	2380.70	12.08	0.984	1.73
	Power law	2382.44	13.81	0.938	1.66
Pooled pair traces, 14,400 records	Logistic	102881.81	0.00	0.964	2.75
	Stretched exponential	102892.66	10.86	0.964	2.75
	Exponential	102916.74	34.94	0.964	2.77
	Power law	103787.87	906.07	0.914	2.79

Table 25: Small labeled calibration-set sensitivity for $\tau_0(0.05)$, measured over 1800 subsample fits per row. The target-workload margin histogram remains label-free, but fitting (α, δ) requires labels.

Calibration items	Median $\lvert\Delta\tau_0\rvert$	P90 $\lvert\Delta\tau_0\rvert$
50	0.67	1.73
100	0.45	1.20
200	0.27	0.69

gain. Table 14 reports these shares for every label budget with the original labels, and Table 16 the bootstrap shares under both label versions; the per-subsample results after correction are in the supplementary material.

J.3 Seed-1 Six-Pair Quartiles

Table 26 breaks the seed-1 oracle gain of every domain pair down by base-margin quartile.

Table 26: Oracle gain by base-margin quartile for all six domain pairs of Qwen3.5-9B, seed 1, with $n = 400$ per pair. The oracle selects among base and the two pair adapters.

Domain pair	n	Base	Oracle	Gap	Q1 gap	Q4 gap
science+medical	400	0.802	0.840	+3.8pp	+12.9pp	0.0pp
science+history	400	0.882	0.915	+3.25pp	+12.9pp	0.0pp
science+philosophy	400	0.807	0.885	+7.8pp	+22.8pp	0.0pp
medical+history	400	0.760	0.825	+6.5pp	+23.8pp	0.0pp
medical+philosophy	400	0.685	0.788	+10.2pp	+21.2pp	+2.0pp
history+philosophy	400	0.765	0.865	+10.0pp	+25.5pp	0.0pp
Mean		0.784	0.853	+6.9pp	+19.8pp	+0.3pp

Oracle gap and individual adapter accuracy. Table 27 lists the individual adapter accuracies of the six seed-1 pairs. In five pairs the base model is the most accurate single model; in science+philosophy both adapters exceed it, and the oracle gap is +7.75 points, close to the six-pair mean. The oracle gap depends on adapters being right on different items than the base model, which aggregate accuracy does not show.

Table 27: Individual adapter accuracies on the seed-1 six-pair traces with $n = 400$ per pair.

Pair A+B	Base	Adapter A	Adapter B	Oracle	Gap
science+medical	0.8025	0.7875	0.7800	0.8400	+3.75pp
science+history	0.8825	0.8750	0.8500	0.9150	+3.25pp
science+philosophy	0.8075	0.8125	0.8275	0.8850	+7.75pp
medical+history	0.7600	0.7325	0.7475	0.8250	+6.50pp
medical+philosophy	0.6850	0.6575	0.6825	0.7875	+10.25pp
history+philosophy	0.7650	0.7125	0.7650	0.8650	+10.00pp

J.4 SEED-1 SCIENCE+MEDICAL QUARTILES

Table 28 gives the full quartile breakdown of the science+medical pair, including the number of helpable items per quartile.

Table 28: Quartile breakdown for the seed-1 science+medical trace of $n = 400$ records, with Q1 holding the lowest margins. Helpable items are base errors that at least one adapter answers correctly; the oracle never loses an item the base model gets right.

Quartile	Margin range	N	Base acc.	Oracle acc.	Gap	Helpable items
Q1, hardest	$[0, 1.0)$	101	0.495	0.624	+12.9pp	13
Q2	$[1.0, 3.5)$	99	0.747	0.768	+2.0pp	2
Q3	$[3.5, 5.4)$	100	0.980	0.980	0.0pp	0
Q4, easiest	$[5.4, 7.9)$	100	0.990	0.990	0.0pp	0
Overall		400	0.802	0.840	+3.8pp	15

J.5 ALL-PAIR MULTI-SEED ROBUSTNESS

We repeat the all-pair oracle stratification for six independent sample seeds. Across 36 pair-seed evaluations, the overall oracle gap is $+6.97 \pm 3.07$pp; the hardest quartile has $+19.2 \pm 5.5$pp gap, while Q4 has $+0.39 \pm 0.97$pp. Table 29 shows that the seed-1 matrix in Table 26 is not a one-off sample artifact.

Table 29: All-pair multi-seed oracle robustness over 6 sample seeds per pair with $n = 400$ per seed. Values are mean $\pm$ std in percentage points.

Pair	Overall oracle gap	Q1 gap	Q4 gap
science+medical	$+3.8 \pm 1.0$pp	$+14.5 \pm 3.6$pp	$+0.0 \pm 0.0$pp
science+history	$+3.1 \pm 0.5$pp	$+11.8 \pm 1.4$pp	$+0.0 \pm 0.0$pp
science+philosophy	$+7.8 \pm 0.6$pp	$+21.2 \pm 2.2$pp	$+0.0 \pm 0.0$pp
medical+history	$+6.1 \pm 1.1$pp	$+20.0 \pm 5.3$pp	$+0.0 \pm 0.0$pp
medical+philosophy	$+11.0 \pm 0.8$pp	$+22.3 \pm 1.7$pp	$+2.3 \pm 1.1$pp
history+philosophy	$+9.9 \pm 1.0$pp	$+25.1 \pm 3.2$pp	$+0.0 \pm 0.0$pp
Mean	$\mathbf{+7.0 \pm 3.1}$pp	$\mathbf{+19.2 \pm 5.5}$pp	$\mathbf{+0.4 \pm 1.0}$pp

J.6 FOUR-ADAPTER ORACLE RECONSTRUCTION

The evaluations are organized by domain pair. The pair records also allow an $N = 4$ oracle over distinct items: an item of a given domain is evaluated in the three pairs that contain that domain, which together cover all four adapters, and its candidate scores are identical across the seeds that sample it. Table 1 counts each item once, so the estimate does not depend on how often, or in which order, an item was sampled. The base model is more accurate than every single adapter on these items,

Table 30: Item-disjoint router evaluation on the 10,083 item–pair records of the 3,361 six-pair items, with 5-fold cross-validation grouped by item, help-probability and lead thresholds tuned on training items only, and fold assignment 1 unless noted. Improvement over base in points, 95% item-cluster bootstrap intervals.

Policy, no labels at test time	Improvement	95% CI	Gains / damage
RF score-feature router	$+1.02$	$[+0.28, +1.81]$	370 / 267
RF, items and test pair disjoint, 30 folds	$+1.03$	$[+0.36, +1.71]$	330 / 226
RF, no pair or domain indicator features	$+1.13$	$[+0.36, +1.92]$	403 / 289
HistGB score-feature router	$+0.68$	$[+0.17, +1.19]$	177 / 108
Oracle, label-aware ceiling	$+7.16$	$[+6.38, +7.98]$	722 / 0

and the four-adapter oracle adds $+9.28$ points, with $+25.76$ points in the lowest-margin quartile and 0.00 in the highest; 79.5% of rescues lie below $\tau_0(0.05)$.

J.7 Item-Disjoint Router Details

Records and folds. We identify an item by its domain and its exact base option-score vector, a label-free signature that serves as the item identity for all item-grouped splits: all 3,361 resulting groups carry a single gold label, and records of the same group in the same pair have identical candidate scores and correctness under every seed. Across all 14,400 records no signature is associated with two different correctness or score patterns for the same candidate, so no two distinct questions are merged into one group with conflicting outcomes. The opposite error, one question split into two groups by numerically different scores, would let copies of it fall on both sides of a split; only five pairs of distinct signatures agree even up to a constant shift within 10^{-2}, and the provenance audit of Section L maps the 3,361 groups one-to-one to source rows. We therefore keep one record per item and pair, giving 10,083 records, and assign whole items to folds with 5-fold grouped cross-validation stratified by whether any adapter rescues the item. Router thresholds, a help probability in $[0.02, 0.80]$ and a lead over the other adapter in $\{0, 0.005, 0.01, 0.02, 0.05, 0.1\}$, are chosen on out-of-fold predictions of an inner 3-fold grouped cross-validation inside the training items, maximizing inner improvement with ties broken toward less damage. The random forest uses 300 trees of depth 5 with balanced class weights; gradient boosting uses 120 iterations with 8 leaves.

Robustness. Table 31 repeats the evaluation with four further fold assignments, holds out the test domain pair in addition to the test items, training only on the other five pairs' records of the training items, and restores the original weighting of all 14,400 records. The gain stays between $+0.99$ and $+1.22$ points with every 95% interval above zero, and with the test pair also held out the router gains $+1.03$ points.

Table 31: Robustness of the random-forest router. Rows 1–5 repeat the item-grouped 5-fold evaluation with different fold assignments, row 1 being the primary endpoint of Table 30; "items and pair disjoint" additionally holds out the test domain pair over 30 folds, training only on the other five pairs' records of the training items; the last row weights every one of the 14,400 original records instead of one record per item and pair. Improvements are over the base model in points with 95% item-cluster bootstrap intervals.

Evaluation	Improvement	95% CI	Gains / damage	Share of ceiling
Fold assignment 1	$+1.02$	$[+0.28, +1.81]$	370 / 267	14.3%
Fold assignment 2	$+1.22$	$[+0.49, +2.00]$	384 / 261	17.0%
Fold assignment 3	$+1.13$	$[+0.39, +1.94]$	379 / 265	15.8%
Fold assignment 4	$+0.99$	$[+0.22, +1.81]$	396 / 296	13.9%
Fold assignment 5	$+1.16$	$[+0.47, +1.87]$	339 / 222	16.2%
Items and pair disjoint, 30 folds	$+1.03$	$[+0.36, +1.71]$	330 / 226	14.4%
All 14,400 records weighted	$+1.19$	$[+0.41, +1.98]$	497 / 326	17.0%

Follow-up analyses. Table 32 reports further router evaluations under the same grouped protocols. *Features:* without the pair and domain indicators the router gains $+1.13$ points, $[+0.36, +1.92]$. *Label noise:* retraining after removing the 16 ARC items whose numeric option labels the letter scorer never matches gives $+1.31$ $[0.59, 2.04]$ for RF and $+1.33$ $[0.78, 1.87]$ for gradient boosting on 10,035 records. *Learners:* gradient boosting stays positive over all five fold assignments, with mean $+0.95$ and every lower bound $\geq +0.17$, and when the test pair is also held out, with $+0.52$ $[+0.03, +0.99]$.

Table 32: Follow-up analyses of the item-disjoint router. Upper row: the RF router of Table 30 on fold assignment 1 without pair and domain indicator features. Middle: routers retrained after removing ARC items with numeric option labels. Lower: gradient boosting over the five fold assignments and the pair-and-item-disjoint split. Improvement over base in points with 95% item-cluster bootstrap intervals.

Evaluation	Improvement	95% CI	Gains / damage
No pair or domain indicator features	$+1.13$	$[+0.36, +1.92]$	403 / 289
RF, numeric-label ARC items removed, 10,035 records	$+1.31$	$[+0.59, +2.04]$	373 / 242
HistGB, numeric-label ARC items removed, 10,035 records	$+1.33$	$[+0.78, +1.87]$	199 / 66
HistGB, five fold assignments	$+0.95$ (SD 0.19)	range $[+0.68, +1.22]$	–
HistGB, items and pair disjoint	$+0.52$	$[+0.03, +0.99]$	179 / 127

J.8 Practical Decision-Rule Diagnostics

For workload screening, base-margin evaluation uses one candidate pass per item, whereas full pair-oracle measurement evaluates base plus two adapters before any selected follow-up. Table 33 summarizes the resulting decision-rule diagnostics.

Table 33: Decision-rule diagnostics over 36 all-pair seed traces at $\tau_0 = 1.68$. The low-margin region has a much larger local oracle gap and holds most oracle gains. Traces share items, so these are descriptive per-trace statistics; held-out evidence for the threshold is in Tables 4, 11 and 12.

Diagnostic	Value
Oracle gap	$+6.97 \pm 3.07$pp
Low-margin local gap	$+18.11 \pm 4.58$pp
High-margin local gap	$+2.37 \pm 2.33$pp
Oracle gains in low-margin region	$81.8 \pm 15.4\%$

J.9 Measured Scoring Time per Candidate Pass

Cost in the main text is counted in candidate forward passes. To relate passes to time in the post hoc, descriptive analysis S9-A of Section P, we use the six scoring evaluations of Section I: in each, the base model and the science adapter, applied as an unmerged LoRA to all 128 target modules, score the same 1,405 MMLU and 200 MMLU-CF check items one after the other under identical scoring conditions. An adapter pass takes 1.12–1.16 times, on average 1.15 times, as long as a base pass on the MMLU check items and 1.16–1.27 times, on average 1.18 times, on the smaller MMLU-CF check set, and the base model scores 27.6–30.4 MMLU items per second. Full-workload passes, run in different evaluations, agree: 31.2 items per second for the re-scored base model and 24.7–26.8 for the four fresh adapters and the code adapter on all 14,042 MMLU items. Applying the mean MMLU ratio of 1.15 to the candidate-pass counts gives estimated scoring-only costs, in base passes of time per item: 2.51 for the screened original or fresh pool at 2.32 candidate passes, against 5.59 for running all five candidates and 2.15 for the base model plus one adapter; over the eight-adapter union, 4.02 screened and 10.18 unscreened. The training-fold subject lookup of Section 4.7 runs one candidate per item, an adapter on 60%, 48% and 76% of the items for the original, fresh and union pools, for estimated costs of 1.09, 1.07 and 1.11 base passes. These estimates combine measured per-pass times with pass counts; they are not timed executions of the screened subset or of a complete selector pipeline, and they do not include loading or switching adapters in a serving system.

K LLAMA-3-8B REPLICATION

We repeat the MCQ measurements with independently trained Llama-3-8B adapters. Values below use a post-hoc correction of native ARC option labels to their positions in the fixed A–D scoring space; all model scores and predictions are unchanged. Eight seed-1 questions occur in three science-containing pairs, giving 24 affected records. Seven original base predictions become correct. Two numeric-label questions in seed 2 and one in seed 3 are also normalized, without changing their correctness. The supplementary material gives the source option order, every changed record before and after the correction, and the code that derives the corrected records. The prompts, which show the native option labels, are unchanged; the correction concerns the evaluation labels only.

Llama-3-8B MCQ experiment. Across three sample seeds, science+medical base accuracy is 0.660 ± 0.008, oracle accuracy 0.730 ± 0.011, and oracle gap $+7.00 \pm 0.66$pp. The Q1 gap is 22.7 ± 3.1pp and the Q4 gap is zero. The pooled base-error fit has $R^2 = 0.990$; pooled Q1/Q4 oracle gaps are $22.67/0.00$pp. Across the six seed-1 pairs, the mean gap is 12.13 ± 4.31pp and mean Q1/Q4 gaps are $30.61/0.00$pp. The concentration of oracle gain at low margin holds under both label versions.

Table 34: Llama-3-8B MCQ replications, mean $\pm$ sample standard deviation. Original and post-hoc position-corrected labels use identical predictions.

Experiment	Replicates	Base	Oracle	Gap
Original science+medical	3 seeds	0.654 ± 0.017	0.723 ± 0.022	$+6.92 \pm 0.76$pp
Original all-pair	6 pairs	0.623 ± 0.075	0.743 ± 0.059	$+12.00 \pm 4.38$pp
Corrected science+medical	3 seeds	0.660 ± 0.008	0.730 ± 0.011	$+7.00 \pm 0.66$pp
Corrected all-pair	6 pairs	0.631 ± 0.081	0.753 ± 0.064	$+12.13 \pm 4.31$pp

Table 35: Llama-3-8B all-pair MCQ replication after positional label correction, seed 1, $n = 400$ per pair. Oracle complementarity is positive in all domain pairs and concentrated in Q1.

Pair	Base	Oracle	Gap	Q1 gap	Q4 gap
history+philosophy	0.610	0.765	+15.50pp	+32.50pp	+0.00pp
medical+history	0.625	0.733	+10.75pp	+34.00pp	+0.00pp
medical+philosophy	0.505	0.657	+15.25pp	+32.28pp	+0.00pp
science+history	0.757	0.838	+8.00pp	+27.00pp	+0.00pp
science+medical	0.652	0.718	+6.50pp	+19.63pp	+0.00pp
science+philosophy	0.637	0.805	+16.75pp	+38.26pp	+0.00pp

Frozen threshold within Llama-3-8B. The seed-1 six-pair records of Llama-3-8B contain 800 distinct items, 200 per domain, each scored by the base model and all four Llama adapters across the three pairs that contain its domain. We apply the unchanged grid and fitter to 400 records of one pair and evaluate on the 400 distinct items of the other two domains. The original science+medical threshold was 4.576380449; label correction gives 4.525123724 post hoc. The history+philosophy threshold remains 3.183733957. Original and corrected thresholds select exactly the same target items, so costs are unchanged. Table 36 reports corrected labels; the original reverse-direction rescue count was 50, now 51. In both directions the frozen threshold keeps every rescue of the four-adapter oracle; because the Llama base model is less accurate, its tail error reaches 5% only at a larger margin, and the threshold sends more items to the adapters than on Qwen3.5-9B. The rescue rate falls from the lowest to the highest margin quartile in both target sets.

Table 36: Position-corrected, post-hoc threshold transfer on Llama-3-8B with base plus four adapters and 400 target items per direction. τ_0: threshold fitted on the anchor pair, with δ, α equal to 0.340, 0.424 and to 0.325, 0.588. Eval.: share of target items with margin below τ_0; passes: candidate passes per item out of five; kept: share of rescues below τ_0 with a 95% Wilson interval, since the item bootstrap is degenerate when every rescue is kept; lost: oracle gain above τ_0 in points; AUC: low margin predicting rescue among base errors; h: rescue rate among base errors by target-margin quartile in %.

Anchor → target	τ_0	Base acc. (%)	Rescues	Eval. (%)	Passes	Kept (%)	Lost	AUC [95% CI]	h Q1 / Q2 / Q3 / Q4
sci+med → hist+phil	4.53	61.0	91	86.0	4.44	100.0 [96.0, 100]	0.00	0.70 [0.61, 0.78]	73.7 / 64.1 / 25.8 / 0.0
hist+phil → sci+med	3.18	65.2	51	57.0	3.28	100.0 [93.0, 100]	0.00	0.73 [0.65, 0.81]	52.2 / 28.0 / 6.7 / 0.0

L Question Provenance and Training-Text Overlap

Question identity. We reproduced the question selection with the sampling code and seeds of the evaluations and mapped every record to its dataset, split and row in the ARC-Challenge test split, the MedMCQA validation split and the MMLU test split by subject; the mapping is in the supplementary material. All 72 blocks of 200 records, one per seed, pair and domain, match the gold answers of their reconstructed source rows, 14,400 of 14,400; every combination of seed, domain and position carries a single source row and base score vector across the pairs that contain it; and the 3,361 label-free item groups used for grouped splits correspond one-to-one to 3,361 source rows: 802 science, 1,071 medical, 634 history and 854 philosophy. The grouped splits therefore coincide with splits by source row. The anchor scores the seed-1 science+medical questions in a separate evaluation: its domains are in the same order, its margins correlate with those of the seed-1 trace at $r = 0.9993$, and base correctness under the original labels agrees on 395 of 400 items. The recorded option scores exactly reproduce every original anchor correctness bit; Section H uses these scores, not a later evaluation, for label correction.

Overlap with adapter training inputs. Table 37 compares evaluation questions, normalized to lower case with non-alphanumerics removed, with each adapter's training inputs, counting exact question matches and items that share at least one 13-gram with the question or with the question plus its gold option. A question shorter than 13 tokens can be flagged only by an exact match or through its gold option text; the 1,124 flagged MMLU items are the union of both criteria, and 1,123 of them carry a shared 13-gram. The medical row uses all rows of the flashcard dataset, which contain the training examples. The history and philosophy overlaps come from text that MMLU reuses across its own splits: 895 of the philosophy-flagged items are moral-scenarios questions, whose fixed question template exceeds 13 tokens, 56 are formal-logic items, and all history-flagged items are high-school world, European and US history questions, which quote long source passages. Removing all 1,124 flagged MMLU test items, which hold 217 of the 1,163 rescues, leaves 12,918 items with base accuracy 80.3%, oracle gain 7.32 points from 946 rescues, rescue rate among base errors 37.1%, and a rescue AUC among base errors of 0.72; the frozen threshold sends 32.3% of them to the adapters at 2.29 passes and keeps 93.3% of the rescues, losing 0.49 points of oracle gain. For the six-adapter pool of Section 4.6, MathInstruct overlaps MMLU test items almost only in high-school mathematics, 79 of the 85 flagged items; dropping all four mathematics subjects from the 1,500-item sample leaves 1,389 items on which the six-adapter pool has 105 rescues, the frozen threshold keeps 94.3% of them at 2.89 passes, and 15 of the 16 rescues added by code and math lie below it. This overlap analysis is a sensitivity analysis outside the registered audit.

Rescored ARC-Challenge. The primary rescoring, a variant fixed in advance, relabels the options of each of the 1,172 ARC-Challenge test items positionally as letters and scores the item over its complete option set, including the seven items with three or five options. Compared with scoring the same items under the original letter matching, it changes the outcome of exactly the 22 items with numeric labels, for the base model and every adapter; on the 802 questions it shares with the six-pair records, that letter scoring reproduces the original correctness except for 5, 2, 2, 1 and 0 items for the base model and the science, medical, history and philosophy adapters, all with letter labels. On all

Table 37: Overlap of evaluation questions with adapter training inputs: exact normalized-question matches before the slash and items sharing a 13-gram of the question or question plus gold option with the training text after it. Counts in the text are the union of both criteria.

Training inputs, adapter	Rows	ARC test	MedMCQA val.	MMLU test	MMLU-CF
SciQ train, science	11,679	1 / 0	0 / 0	0 / 0	0 / 0
Medical Meadow flashcards, medical	33,955	0 / 0	0 / 0	0 / 0	0 / 0
MMLU val.+dev, history subjects	145	0 / 0	0 / 0	1 / 167	0 / 0
MMLU val.+dev, philosophy subjects	229	0 / 0	0 / 0	1 / 956	0 / 0
CodeAlpaca-20k, code	20,022	0 / 0	0 / 0	0 / 0	0 / 0
MathInstruct, math	262,039	0 / 0	0 / 0	74 / 78	0 / 4

1,172 items the base model reaches 94.8%, the adapters 93.4–94.5%, and the oracle 97.1% with 27 rescues and an oracle gain of $+2.30$ points; the rescue AUC among base errors is 0.77 [0.64, 0.88], and the frozen threshold sends 10.0% of the items to the adapters at 1.40 passes and keeps 24 of the 27 rescues. Excluding the 200 seed-1 science questions of the calibration anchor leaves the 972 items of Table 13; both item lists are in the supplementary material. Substituting the rescored outcomes into the six-pair records moves the 36-trace mean oracle gap from 6.97 to 6.94 points, with 19.2 in the lowest and 0.39 in the highest quartile in both cases.

M Oracle Estimands

Several oracle gaps appear in the paper; Table 38 lists how they differ in the item population, the candidate pool, and the unit of counting.

Table 38: Oracle estimands used in the paper. "Records" count evaluations of an item within a domain pair; "items" count distinct questions.

Quantity	Population	Candidates	Unit	Oracle gap
Anchor fit, Figure 1a	400 seed-1 science+medical questions, separate scoring	base + science, medical	item	–
Seed-1 trace, Table 28	400 records, science+medical, sample seed 1	base + 2 pair adapters	record	$+3.8$
Pair-seed traces, Table 29	36 traces of 400 records; traces share items	base + 2 pair adapters	trace mean	$+6.97$
Router population, Tables 4 and 30	10,083 item–pair records of 3,361 items	base + 2 pair adapters	record	$+7.16$
Four-adapter oracle, Table 1	3,361 distinct items of the six-pair records	base + 4 adapters	item	$+9.28$
Full MMLU, Table 6	14,042 test rows, 57 subjects	base + 4 adapters	item	$+8.28$
MMLU-CF, Table 4	2,000 items	base + 4 adapters	item	$+7.45$
MMLU subset, Table 4	1,500 MMLU test rows from 51 subjects, a subset of full MMLU	base + 4 adapters	item	$+6.93$

N Experimental Details

LoRA training configuration. All domain adapters use rank $r = 16$, $\alpha = 32$ and dropout 0.05, and adapt the query, key, value and output projections of attention and the gate, up and down projections of the MLP. In Qwen3.5-9B these are the attention projections of the 8 full-attention layers, layers 3, 7, …, 31, and the MLP projections of all 32 layers, so each adapter has 128 LoRA pairs; the projections of the 24 linear-attention layers are not adapted. Training data: the 11,679 SciQ training examples for science, 19,765 Medical Meadow medical flashcards for medical, and the 145 and 229 unique MMLU validation and dev questions of the history and philosophy subjects, each repeated 8 times. Training uses supervised fine-tuning with batch size 4, gradient accumulation 4, learning rate 2×10^{-4} with cosine schedule, 2 epochs, bf16 precision. No adapter is trained on a test split: ARC-Challenge and MedMCQA are used only for evaluation, and the history and philosophy adapters see only MMLU validation and dev questions. Text shared between these inputs and test questions is audited in Section L. The code and math adapters used for the six-adapter pool of Section 4.6 follow the same recipe on CodeAlpaca-20k (Chaudhary, 2023) and MathInstruct (Yue et al., 2024).

Evaluation details. Logprob MCQ scoring computes the mean per-token logprob of each answer option A, B, C and D. The margin $m(x)$ is computed as the difference between the highest and second-highest option logprobs after length normalization. All reported results use the same evaluation framework with fixed random seeds for reproducibility. For full MMLU, MMLU-CF and the freshly trained pool, the zero-shot prompt consists of the question, the four lettered options and an answer cue, and an option's score is the log-probability of its single letter token, preceded by a space, at the next position, so the mean per-token logprob equals that token's logprob; the model is evaluated in bf16 with batches of 16 and left padding, and the log-softmax is taken in float32. Rows are keyed by their index in the MMLU test split over all subjects in dataset order, and in the 2,000-item MMLU-CF selection, the first 2,000 rows of the validation split shuffled with seed 1; the six-pair records map to dataset rows as described in Section L.

Reproduction. The supplementary material contains the per-item option scores and correctness of every candidate for the six-pair records, full MMLU, MMLU-CF, and the 1,500-item MMLU subset, together with the code for all router, screening, and rescue-rate analyses, so every number in Tables 1, 2, 4, 5, 7, 8, 11, 12, 14, 15, 18, 19, 30 to 32 and 37 can be recomputed from the released scores, as can Tables 3, 9, 20, 27 and 36. The supplementary code recomputes the key numbers of Sections 4.2, 4.4 and 4.5, namely the full-MMLU accuracies and rescues, h by quartile and AUC on the three MMLU workloads and for the fresh pool, and the passes, rescues kept and gain lost at the primary threshold, and prints them next to the values printed in the paper; all agree to the printed rounding. The supplementary index maps every table to the analysis that produces it, and each registered analysis of Section P to its registration entry, training records and per-item rows. The adapter training follows the protocol above.

O PROOFS

Proof of Equation (2) *and of the margin ceiling in Remark 4.*

Step 1: pointwise identity. Since the oracle includes the base model, $g^+(x) = 1$ exactly when the base model is wrong and $A_+(x) \neq \emptyset$:

$$g^+(x) = \mathbf{1}[\text{base wrong at } x] \cdot \mathbf{1}[A_+(x) \neq \emptyset].$$

Step 2: factorization. Taking conditional expectations given $m(x) \geq \tau$:

$$\begin{aligned} C_>(\tau) &= \mathbb{E}[g^+(x) \mid m(x) \geq \tau] && \text{(Definition 3)} \\ &= \mathbb{P}(\text{base wrong},\ A_+(x) \neq \emptyset \mid m(x) \geq \tau) && \text{(Step 1)} \\ &= e(\tau) \cdot h_>(\tau) && \text{(definition of } h_>), \end{aligned}$$

where the last step requires $e(\tau) > 0$. If $e(\tau) = 0$, the event conditioned on in $h_>(\tau)$ has probability zero; we then set $h_>(\tau) = 0$ by convention, and both sides are zero because the left side is at most $e(\tau)$ by Step 1.

Step 3: ceiling. Since $h_>(\tau) \leq 1$,

$$C_>(\tau) \leq e(\tau) = \mathbb{P}(\text{base wrong} \mid m(x) \geq \tau).$$

The same argument applies within any margin bin. □

Proof of the router improvement decomposition in Proposition 5.

Step 1: improvement. Define the improvement of router r as

$$\mathrm{Imp}(r) = \mathbb{P}(f_{r(x)}(x) \text{ correct}) - \mathbb{P}(f_0(x) \text{ correct}). \tag{8}$$

Step 2: partition. Partition all items into four mutually exclusive events:

1. E_1: base correct, router correct, no change;
2. E_2: base wrong, router correct, a gain from routing;
3. E_3: base correct, router wrong, damage from routing;

4. E_4: base wrong, router wrong, no change.

Since the router is correct on $E_1 \cup E_2$ and the base model on $E_1 \cup E_3$,

$$\mathrm{Imp}(r) = \mathbb{P}(E_2) - \mathbb{P}(E_3).$$

Step 3: gains. A gain occurs only when some adapter can help *and* the router selects one of them, so

$$\mathbb{P}(E_2) = \mathbb{P}(A_+(x) \neq \emptyset) - \mathbb{P}\big(A_+(x) \neq \emptyset,\ r(x) \notin A_+(x)\big).$$

Step 4: combination. Substituting Step 3 into Step 2, and noting that $\mathbb{P}(E_3)$ is the damage term:

$$\begin{aligned}\mathrm{Imp}(r) = \mathbb{P}(A_+ \neq \emptyset) \\ - \underbrace{\mathbb{P}(A_+ \neq \emptyset,\ r \notin A_+)}_{\mathbb{E}[\mathrm{Miss}(r,x)]} \\ - \underbrace{\mathbb{P}(\text{base correct},\ f_{r(x)}\ \text{wrong})}_{\mathbb{E}[\mathrm{Damage}(r,x)]},\end{aligned} \tag{9}$$

which gives the stated decomposition

$$\mathrm{Imp}(r) = \mathrm{OC} - \mathbb{E}[\mathrm{Miss}] - \mathbb{E}[\mathrm{Damage}].$$

□

Proof of the perturbation stability in Lemma 6.

Step 1: flip condition. A prediction flip at item x requires some option $c \neq \hat{y}$ with

$$f_M(x)[c] \geq f_M(x)[\hat{y}]$$

where equality covers a tie resolved against $\hat{y}$, that is,

$$\delta_M(x)[c] - \delta_M(x)[\hat{y}] \geq f_0(x)[\hat{y}] - f_0(x)[c].$$

Since $f_0(x)[\hat{y}] - f_0(x)[c] \geq m(x)$ for every $c \neq \hat{y}$ by Definition 1, this implies

$$\delta_M(x)[c] - \delta_M(x)[\hat{y}] \geq m(x).$$

Step 2: necessary condition. Since

$$\delta_M(x)[c] - \delta_M(x)[\hat{y}] \leq 2\|\delta_M(x)\|_\infty,$$

a necessary condition for a flip is $\|\delta_M(x)\|_\infty \geq m(x)/2$.

Step 3: Markov's inequality. Conditional on base correct and $m \in [\tau, \tau + w]$ with $\tau > 0$ and $w > 0$, write $\cdot$ for this event. Then

$$\begin{aligned}\mathbb{P}(\text{flip} \mid \cdot) &\leq \mathbb{P}(\|\delta_M(x)\|_\infty \geq \tau/2 \mid \cdot) && \text{(Step 2, } m \geq \tau) && (10)\\ &\leq \frac{\mathbb{E}[\|\delta_M(x)\|_\infty^2 \mid \cdot]}{(\tau/2)^2} && \text{(Markov)} && (11)\\ &= \frac{4\,\mathbb{E}[\|\delta_M(x)\|_\infty^2 \mid \cdot]}{\tau^2},\end{aligned}$$

where the Markov step applies to the non-negative random variable $\|\delta_M(x)\|_\infty^2$ and requires a finite conditional second moment. □

P Registered Analyses

Table 39 lists every analysis that was registered, with its endpoints, before its outputs existed, in the order of registration, together with the registered endpoint values; the last column points to where the analysis is used in the paper. Values are in percentage points of accuracy unless marked; intervals are 95% bootstrap intervals clustered as in the registration, by item or, for R2, by subject. The registration document is in the supplementary material.

Table 39: Registered analyses and their registered endpoints. "Improvement": selector accuracy minus base accuracy. Pools: O = original four adapters, F = freshly trained four adapters.

ID	Analysis	Registered endpoint values	In paper
R1	Item-disjoint router, six-pair records	RF improvement $+1.02$ $[+0.28, +1.81]$ on fold assignment 1, mean over five assignments $+1.10$; HistGB $+0.68$ $[+0.17, +1.19]$; logistic $+0.14$ $[-0.84, +1.10]$	Table 30
R1	Label-free policies under the same folds	best fixed adapter $+0.39$ $[-0.43, +1.15]$; margin-threshold routing $+0.06$ $[-0.92, +1.00]$; score averaging $+0.14$ $[-0.66, +0.92]$; plurality vote -0.22 $[-0.69, +0.26]$	–
R1b, R1c	Items and test pair disjoint; original record weighting	RF $+1.03$ $[+0.36, +1.71]$; RF $+1.19$	Table 31
R2	Router across held-out MMLU subjects, full MMLU, 5 candidates	RF -1.16 $[-1.69, -0.63]$; HistGB -0.10 $[-0.28, +0.11]$; logistic -2.89 $[-3.50, -2.28]$	–
R2b, R2c	Same router trained on full MMLU, applied to MMLU-CF / 1,500-item subset	RF -0.50 $[-1.40, +0.35]$ / $+0.13$ $[-0.93, +1.20]$	–
R3	Frozen-parameter calibration hold-out	median frozen R^2 over 200 halves 0.892	Section 4.3
R4	ARC items with numeric option labels	36-trace oracle gap $6.97 \rightarrow 7.00$; router, Table 32	Section 4
R5	Frozen screen on three workloads	Table 4	Section 4.4
M	Validation-tuned TA / TIES / DARE-TIES / DARE-linear merges, 6 pairs $\times$ 3 seeds	all 18 pair-seed evaluations: mean gain over base $+0.81$ / $+0.64$ / $+0.19$ / $+0.49$; fixed-setting TA / sign-agreement / DARE $+0.56$ / $+0.31$ / $+0.54$; per-run records and aggregation in the supplement	–
S2-C1	Calibration sensitivity: halves, bootstrap, label budgets	Tables 5, 14 and 15	Section 4.6
S2-C1	Same, 5-shot full MMLU, pool O	primary threshold: 2.17 passes, 79.7% of rescues kept, 2.07 lost; rescue AUC 0.68	–
S2-C2	Provenance and training-text overlap	Table 37	Section L
S2-C3	ARC-Challenge rescoring	Table 5, Section L	Section 4.6
G1	Freshly trained pool F and union O+F	Tables 5 and 17 to 19; union: 86.4% kept at 3.63 passes	Section 4.5
G2	Code and math adapters on full MMLU and MMLU-CF	code adapter scored on both; math adapter scored on the 1,500-item subset of Section 4.6; six-adapter and replacement pools on the full workloads not formed	Table 9
S3-A	Re-scoring no-change control	Table 20	Section I
S3-A	Re-scoring-robust rescues, with base and adapter gaps $> 2\epsilon^* = 0.748$ and $\epsilon^* = 0.374$ the largest check-item score change, pools O / F on MMLU, MMLU-CF	robust rescues 373 / 31 for O and 286 / 33 for F; h Q1$\rightarrow$Q4 19.3$\rightarrow$2.0, 5.5$\rightarrow$0.0, 18.6$\rightarrow$0.0, 8.3$\rightarrow$0.0%; AUC 0.52 $[0.49, 0.56]$, 0.60 $[0.53, 0.66]$, 0.59 $[0.56, 0.62]$, 0.62 $[0.54, 0.70]$; kept 68.4, 77.4, 78.7, 78.8%	–
S3-B	In-domain net gain, adapter minus base	science on ARC -0.43 $[-1.37, +0.51]$; medical on MedMCQA -1.77 $[-3.73, +0.09]$; history -3.36 $[-5.36, -1.54]$; philosophy $+3.86$ $[+1.90, +5.81]$; law -4.71 $[-6.86, -2.49]$; economics -3.97 $[-5.75, -2.19]$; mathematics -0.94 $[-3.29, +1.41]$; CS -2.15 $[-4.85, +0.18]$; code -0.24 $[-2.18, +1.70]$	–
S3-B	Shift size and per-adapter profiles	Table 9	Section C
S3-B	Magnitude-matched shift control, O / F on MMLU, MMLU-CF	rescues adapters/control 0.90 $[0.86, 0.93]$, 0.63 $[0.55, 0.70]$, 0.67 $[0.64, 0.70]$, 0.49 $[0.42, 0.54]$; changed base errors ending on gold 66.8/58.7, 53.8/53.6, 64.7/63.8, 58.5/59.2%	Section C
S3-C	SGR over all 200 halves with fallback	Tables 15 and 19	Section E
S3-D	Key numbers recomputed from released rows	all printed values reproduced	Section N
S5-A	Same-configuration base rows for the fresh pool: paired re-scoring of all MMLU and MMLU-CF items, completed after the registered date and analysed as registered	answers unchanged 99.1 / 99.3%; threshold crossings 92 / 24; rescue sets in both, only original and only re-scored rows 1,024 / 31 / 48 and 146 / 5 / 4; kept 91.5% $[89.8, 93.2]$ / 94.0% $[90.3, 97.4]$, evaluated 32.9 / 39.7%, lost 0.65 / 0.45; target met on both workloads, Table 21	Section 4.5, Section I
S5-B	Frozen threshold within Llama-3-8B, two directions	kept 91/91 and 50/50 rescues, lost 0.00 / 0.00; evaluated 86.0 / 57.0% of items	Table 36
S5-C	Answer changes vs. rescues below the threshold, five workloads	change share lower by 0.1–9.0 points on every workload; quartile Spearman 0.95–1.00	Table 3
S5-D	Residuals of the anchor fit	27 of 50 grid points above the curve, largest deviation 0.019	Section J.2

Rules of S6-A to S10. The registrations fixed these rules, each evaluated per target: (a) the screen keeps a selector's accuracy if the lower 95% bound of screened minus unscreened accuracy exceeds -0.25 points, for selectors whose unscreened gain over base has a lower bound above zero; (b) a label-free selector, averaging or vote, realizes a gain below the threshold if its screened gain has

Table 39, continued.

ID	Analysis	Registered endpoint values	In paper
S6-A	Fixed selectors, namely averaging, vote, confidence and cross-validated logistic, with and without the frozen screen, six targets	Table 40; rule (a) met by the logistic selector on MMLU with pools O and O+F; rule (c) met on MMLU by the logistic selector on O, F and O+F, averaging on F and vote on F and O+F	Section 4.7
S7-A	S6-A on MMLU with repeated questions kept in one fold and resampled together	Table 41; rule (a) met by the logistic selector on O and O+F; rule (c) met by the logistic selector on O, F and O+F, averaging on F and vote on F and O+F	Section 4.7
S7-B	Training-fold subject lookup, screened lookup, subject-aware logistic selector, MMLU	Table 42; rule (d) met on F; rule (e) met on O, F, O+F; rule (f) met on F	Section 4.7
S8-A	S7-A selectors on the magnitude-matched shift control, MMLU	Table 43; rule (g) met in both arms by averaging, vote and confidence on O, F and O+F and by the logistic selector on O and O+F	Section 4.2
S8-B	Training-selected fixed candidate, MMLU	base model chosen in every fold; rule (h) met on O and O+F, Table 44	Section 4.7
S8-C	Decomposition terms of the S7-A arms	Table 45	Section 4.7
S9-A	Measured scoring time per candidate pass, post hoc and descriptive	adapter/base time ratio 1.12–1.16 on MMLU check items over six evaluations, 1.16–1.27 on MMLU-CF; screened pool 2.51 base passes of time per item for O and F, 4.02 for O+F; lookup 1.07–1.11	Section J.9
S10-A	S7/S8 policies under 20 question-group fold assignments, MMLU	Table 46; rule (i) met on O and O+F; rule (j) met on O, F, O+F	Section 4.7
S10-B	Selector and inexpensive policies trained on common label budgets of 400, 1,000 and 2,000 rows, MMLU	Table 47; rule (k) met on F at all three budgets	Section 4.7
S10-C	Selector and fixed candidate frozen on MMLU, applied to MMLU-CF	Table 48; rule (m) met on O and F	–
S11-A	Second training seed of pool F, full MMLU and MMLU-CF	registered; no endpoint computed	–

a lower bound above zero, on the four Qwen targets with four adapters; (c) the screened selector exceeds the best single adapter in hindsight with a lower bound above zero, a comparison with fixed adapters only; (d) the screened subject-aware selector exceeds the training-fold subject lookup with a lower bound above zero; (e) and (f) the lower bound of screened minus unscreened accuracy exceeds -0.25 points, for the subject-aware selector and for the lookup; (g) the real adapters exceed the magnitude-matched shift control under the same selector and arm, with a lower bound of real minus control accuracy above zero; (h) the screened logistic selector exceeds the training-selected fixed candidate with a lower bound above zero; (i) the screened logistic selector's gain over base has a lower bound above zero in at least 18 of 20 fold assignments; (j) the lower bound of screened minus unscreened logistic accuracy exceeds -0.25 points in at least 18 of 20 fold assignments; (k) at a label budget of n training rows, the screened logistic selector exceeds the subject lookup trained on the same rows with a lower bound above zero; (l) a logistic selector frozen on MMLU and applied to MMLU-CF has a screened gain over base with a lower bound above zero; (m) for that frozen selector, the lower bound of screened minus unscreened accuracy exceeds -0.25 points. Cost is counted in candidate forward passes: every candidate scores the same prompt with one pass per item, and the lookup runs only the candidate it selects. Section J.9 converts passes to measured scoring time. On the Llama items the S6-A folds are grouped by item identity in fixed order rather than shuffled.

Table 40: Registered endpoints of S6-A: fixed selectors with all candidates, unscreened, and only below the frozen threshold, screened. Target column: workload and pool, number of adapters, and screened and unscreened candidate passes per item. CV logistic: the logistic selector trained under cross-validation. Values in points of accuracy with 95% paired item-bootstrap intervals; "vs. best single": screened selector minus the best single adapter in hindsight. Rules (a)–(c) as registered in Section P; "–": no rule met. Llama items: the 800 distinct seed-1 items scored by all four Llama-3-8B adapters, with the threshold fitted on the science and medical anchor, whose items are not held out from the fit. Corrected rows refit the threshold and the same cross-validation selector after label correction; these are post-hoc. Keeping the original threshold gives identical decisions.

Target, adapters, passes	Selector	Unscreened gain	Screened gain	Screened − unscreened	vs. best single	Rules
MMLU, O, 4, 2.32/5	averaging	-0.03 $[-0.42, +0.36]$	$+0.01$ $[-0.38, +0.38]$	$+0.04$ $[-0.04, +0.11]$	$+0.18$ $[-0.24, +0.59]$	–
	vote	$+0.06$ $[-0.24, +0.35]$	$+0.08$ $[-0.22, +0.36]$	$+0.02$ $[-0.03, +0.07]$	$+0.25$ $[-0.06, +0.53]$	–
	confidence	-0.24 $[-0.69, +0.22]$	-0.27 $[-0.71, +0.17]$	-0.03 $[-0.18, +0.12]$	-0.10 $[-0.56, +0.36]$	–
	CV logistic	$+0.39$ $[+0.08, +0.71]$	$+0.36$ $[+0.05, +0.67]$	-0.04 $[-0.09, +0.02]$	$+0.53$ $[+0.22, +0.83]$	a,c
MMLU-CF, O, 4, 2.58/5	averaging	-0.40 $[-1.25, +0.55]$	-0.45 $[-1.30, +0.45]$	-0.05 $[-0.15, +0.00]$	$+0.20$ $[-0.70, +1.20]$	–
	vote	-1.10 $[-1.90, -0.35]$	-1.10 $[-1.85, -0.35]$	$+0.00$ $[-0.15, +0.15]$	-0.45 $[-1.30, +0.40]$	–
	confidence	-1.45 $[-2.45, -0.45]$	-1.45 $[-2.40, -0.45]$	$+0.00$ $[-0.15, +0.15]$	-0.80 $[-1.85, +0.30]$	–
	CV logistic	-0.45 $[-0.95, +0.05]$	-0.45 $[-0.95, +0.05]$	$+0.00$ $[+0.00, +0.00]$	$+0.20$ $[-0.80, +1.20]$	–
MMLU, F, 4, 2.32/5	averaging	-0.07 $[-0.48, +0.34]$	$+0.08$ $[-0.31, +0.49]$	$+0.15$ $[+0.06, +0.24]$	$+0.56$ $[+0.19, +0.95]$	c
	vote	-0.40 $[-0.76, -0.06]$	$+0.03$ $[-0.31, +0.35]$	$+0.43$ $[+0.30, +0.57]$	$+0.51$ $[+0.16, +0.85]$	c
	confidence	-0.26 $[-0.68, +0.16]$	-0.12 $[-0.53, +0.31]$	$+0.14$ $[+0.06, +0.23]$	$+0.36$ $[-0.03, +0.77]$	–
	CV logistic	$+0.12$ $[-0.12, +0.36]$	$+0.11$ $[-0.14, +0.35]$	-0.01 $[-0.04, +0.00]$	$+0.58$ $[+0.24, +0.94]$	c
MMLU-CF, F, 4, 2.58/5	averaging	$+0.10$ $[-0.85, +1.10]$	$+0.05$ $[-0.90, +1.00]$	-0.05 $[-0.15, +0.00]$	$+0.25$ $[-0.80, +1.35]$	–
	vote	$+0.50$ $[-0.35, +1.40]$	$+0.50$ $[-0.35, +1.40]$	$+0.00$ $[-0.15, +0.15]$	$+0.70$ $[-0.35, +1.75]$	–
	confidence	$+0.15$ $[-0.90, +1.15]$	$+0.10$ $[-0.90, +1.15]$	-0.05 $[-0.15, +0.00]$	$+0.30$ $[-0.75, +1.40]$	–
	CV logistic	-0.20 $[-0.90, +0.50]$	-0.20 $[-0.85, +0.50]$	$+0.00$ $[-0.15, +0.15]$	$+0.00$ $[-1.05, +1.05]$	–
MMLU, O+F, 8, 3.63/9	averaging	$+0.19$ $[-0.20, +0.57]$	$+0.21$ $[-0.17, +0.59]$	$+0.03$ $[-0.04, +0.10]$	$+0.38$ $[-0.02, +0.76]$	–
	vote	$+0.25$ $[-0.06, +0.57]$	$+0.26$ $[-0.05, +0.58]$	$+0.01$ $[-0.04, +0.07]$	$+0.43$ $[+0.09, +0.78]$	c
	confidence	$+0.04$ $[-0.42, +0.50]$	$+0.02$ $[-0.43, +0.46]$	-0.02 $[-0.16, +0.11]$	$+0.19$ $[-0.26, +0.63]$	–
	CV logistic	$+0.54$ $[+0.22, +0.84]$	$+0.53$ $[+0.21, +0.83]$	-0.01 $[-0.06, +0.03]$	$+0.70$ $[+0.34, +1.03]$	a,c
Llama, original, 4, 4.19/5	averaging	$+0.50$ $[-1.62, +2.75]$	$+0.50$ $[-1.62, +2.75]$	$+0.00$ $[+0.00, +0.00]$	-0.38 $[-2.62, +1.88]$	–
	vote	-0.12 $[-1.38, +1.12]$	-0.12 $[-1.38, +1.12]$	$+0.00$ $[+0.00, +0.00]$	-1.00 $[-3.00, +0.88]$	–
	confidence	$+0.62$ $[-1.75, +3.12]$	$+0.88$ $[-1.50, +3.38]$	$+0.25$ $[+0.00, +0.62]$	$+0.00$ $[-2.75, +2.62]$	–
	CV logistic	$+0.25$ $[-2.12, +2.75]$	$+0.25$ $[-2.12, +2.75]$	$+0.00$ $[+0.00, +0.00]$	-0.62 $[-2.38, +1.12]$	–
Llama, corrected, 4, 4.19/5	averaging	$+0.38$ $[-1.88, +2.63]$	$+0.38$ $[-1.88, +2.63]$	$+0.00$ $[+0.00, +0.00]$	-0.25 $[-2.62, +2.00]$	–
	vote	-0.38 $[-1.75, +0.88]$	-0.38 $[-1.75, +0.88]$	$+0.00$ $[+0.00, +0.00]$	-1.00 $[-3.00, +0.88]$	–
	confidence	$+0.50$ $[-1.88, +3.00]$	$+0.75$ $[-1.62, +3.12]$	$+0.25$ $[+0.00, +0.62]$	$+0.12$ $[-2.50, +2.75]$	–
	CV logistic	-0.12 $[-2.38, +2.25]$	-0.12 $[-2.38, +2.25]$	$+0.00$ $[+0.00, +0.00]$	-0.75 $[-2.62, +0.88]$	–

Table 41: Registered endpoints of S7-A: the selectors of Table 40 on full MMLU with repeated questions, the 105 groups of 210 rows with identical normalized question-and-option text, kept in the same cross-validation fold and resampled together by a question-cluster bootstrap. Columns as in Table 40.

Target, adapters, passes	Selector	Unscreened gain	Screened gain	Screened − unscreened	vs. best single	Rules
MMLU, O, 4, 2.32/5	averaging	-0.03 $[-0.43, +0.37]$	$+0.01$ $[-0.38, +0.39]$	$+0.04$ $[-0.04, +0.11]$	$+0.18$ $[-0.23, +0.58]$	–
	vote	$+0.06$ $[-0.24, +0.34]$	$+0.08$ $[-0.21, +0.36]$	$+0.02$ $[-0.02, +0.07]$	$+0.25$ $[-0.06, +0.57]$	–
	confidence	-0.24 $[-0.72, +0.24]$	-0.27 $[-0.71, +0.18]$	-0.03 $[-0.19, +0.12]$	-0.10 $[-0.58, +0.36]$	–
	CV logistic	$+0.46$ $[+0.14, +0.78]$	$+0.43$ $[+0.11, +0.74]$	-0.03 $[-0.08, +0.02]$	$+0.60$ $[+0.30, +0.91]$	a,c
MMLU, F, 4, 2.32/5	averaging	-0.07 $[-0.47, +0.33]$	$+0.08$ $[-0.31, +0.46]$	$+0.15$ $[+0.06, +0.23]$	$+0.56$ $[+0.21, +0.94]$	c
	vote	-0.40 $[-0.77, -0.01]$	$+0.03$ $[-0.31, +0.38]$	$+0.43$ $[+0.29, +0.57]$	$+0.51$ $[+0.15, +0.86]$	c
	confidence	-0.26 $[-0.67, +0.16]$	-0.12 $[-0.53, +0.29]$	$+0.14$ $[+0.06, +0.22]$	$+0.36$ $[-0.02, +0.75]$	–
	CV logistic	$+0.23$ $[-0.01, +0.48]$	$+0.21$ $[-0.03, +0.46]$	-0.01 $[-0.04, +0.00]$	$+0.69$ $[+0.31, +1.05]$	c
MMLU, O+F, 8, 3.63/9	averaging	$+0.19$ $[-0.20, +0.57]$	$+0.21$ $[-0.16, +0.59]$	$+0.03$ $[-0.04, +0.09]$	$+0.38$ $[+0.00, +0.78]$	–
	vote	$+0.25$ $[-0.06, +0.58]$	$+0.26$ $[-0.05, +0.59]$	$+0.01$ $[-0.04, +0.07]$	$+0.43$ $[+0.11, +0.79]$	c
	confidence	$+0.04$ $[-0.41, +0.49]$	$+0.02$ $[-0.41, +0.46]$	-0.02 $[-0.15, +0.11]$	$+0.19$ $[-0.26, +0.65]$	–
	CV logistic	$+0.60$ $[+0.28, +0.92]$	$+0.56$ $[+0.24, +0.87]$	-0.04 $[-0.09, +0.01]$	$+0.73$ $[+0.38, +1.10]$	a,c

Table 42: Registered endpoints of S7-B: policies with training-fold subject information on full MMLU, with folds and bootstrap as in Table 41. Lookup: per subject, the candidate with the highest training-fold accuracy; lookup, screened: the lookup's adapter only below the frozen threshold; subject-aware: the logistic selector with each candidate's training-fold subject accuracy as an added feature. Upper block: gain over base in points with 95% interval, and candidate passes per item in parentheses; middle block: paired differences. Rules (d)–(f) as registered in Section P; "–": no rule met.

Policy	MMLU, O, 4 adapters	MMLU, F, 4 adapters	MMLU, O+F, 8 adapters
Lookup	$+0.63$ $[+0.24, +1.02]$ (1.00)	-0.31 $[-0.61, -0.01]$ (1.00)	$+0.61$ $[+0.17, +1.02]$ (1.00)
Lookup, screened	$+0.33$ $[-0.01, +0.66]$ (1.23)	-0.19 $[-0.48, +0.11]$ (1.15)	$+0.26$ $[-0.11, +0.65]$ (1.27)
Subject-aware	$+0.53$ $[+0.23, +0.83]$ (5.00)	$+0.19$ $[-0.04, +0.42]$ (5.00)	$+0.71$ $[+0.39, +1.04]$ (9.00)
Subject-aware, screened	$+0.43$ $[+0.13, +0.72]$ (2.32)	$+0.17$ $[-0.06, +0.41]$ (2.32)	$+0.67$ $[+0.36, +0.99]$ (3.63)
Screened subject-aware − lookup	-0.20 $[-0.56, +0.16]$	$+0.48$ $[+0.16, +0.80]$	$+0.06$ $[-0.33, +0.47]$
Screened − unscreened, subject-aware	-0.11 $[-0.17, -0.04]$	-0.01 $[-0.04, +0.00]$	-0.04 $[-0.10, +0.02]$
Screened − unscreened, lookup	-0.30 $[-0.46, -0.14]$	$+0.11$ $[+0.04, +0.20]$	-0.34 $[-0.50, -0.18]$
Rules met	e	d,e,f	e

Table 43: Registered endpoints of S8-A: the selectors of Table 41 applied to the real adapters and to the magnitude-matched shift control of Section C, which assigns each adapter's per-item score shift to random options and is averaged over 20 draws, with the folds, threshold and question-cluster bootstrap of S7-A. Gains over base in points with 95% interval, and gained and damaged items in parentheses, averaged over draws for the control; last numeric column: real minus control accuracy. Rule (g) as registered in Section P; "–": rule not met.

Target, adapters	Selector	Arm	Real adapters	Shift control	Real − control	Rule
MMLU, O, 4	averaging	unscreened	-0.03 $[-0.43, +0.37]$ (398/402)	-1.17 $[-1.40, -0.94]$ (317/481)	$+1.14$ $[+0.81, +1.49]$	g
		screened	$+0.01$ $[-0.38, +0.39]$ (385/384)	-1.11 $[-1.35, -0.89]$ (315/471)	$+1.12$ $[+0.80, +1.46]$	g
	vote	unscreened	$+0.06$ $[-0.24, +0.34]$ (217/209)	-0.35 $[-0.45, -0.26]$ (108/158)	$+0.41$ $[+0.14, +0.68]$	g
		screened	$+0.08$ $[-0.21, +0.36]$ (213/202)	-0.34 $[-0.44, -0.24]$ (108/156)	$+0.42$ $[+0.15, +0.69]$	g
	confidence	unscreened	-0.24 $[-0.72, +0.24]$ (543/577)	-3.45 $[-3.79, -3.14]$ (455/940)	$+3.21$ $[+2.83, +3.63]$	g
		screened	-0.27 $[-0.71, +0.18]$ (481/519)	-2.67 $[-3.00, -2.37]$ (425/801)	$+2.40$ $[+2.03, +2.80]$	g
	CV logistic	unscreened	$+0.46$ $[+0.14, +0.78]$ (282/218)	-0.06 $[-0.09, -0.02]$ (14/23)	$+0.51$ $[+0.20, +0.83]$	g
		screened	$+0.43$ $[+0.11, +0.74]$ (273/213)	-0.06 $[-0.09, -0.02]$ (14/22)	$+0.48$ $[+0.17, +0.79]$	g
MMLU, F, 4	averaging	unscreened	-0.07 $[-0.47, +0.33]$ (423/433)	-1.67 $[-1.93, -1.42]$ (364/599)	$+1.60$ $[+1.25, +1.97]$	g
		screened	$+0.08$ $[-0.31, +0.46]$ (414/403)	-1.46 $[-1.72, -1.22]$ (358/563)	$+1.54$ $[+1.19, +1.90]$	g
	vote	unscreened	-0.40 $[-0.77, -0.01]$ (312/368)	-1.40 $[-1.55, -1.25]$ (192/388)	$+1.00$ $[+0.65, +1.33]$	g
		screened	$+0.03$ $[-0.31, +0.38]$ (293/289)	-1.00 $[-1.15, -0.86]$ (182/322)	$+1.03$ $[+0.72, +1.34]$	g
	confidence	unscreened	-0.26 $[-0.67, +0.16]$ (454/490)	-2.94 $[-3.26, -2.65]$ (444/856)	$+2.68$ $[+2.32, +3.07]$	g
		screened	-0.12 $[-0.53, +0.29]$ (446/463)	-2.26 $[-2.58, -1.97]$ (430/748)	$+2.14$ $[+1.79, +2.51]$	g
	CV logistic	unscreened	$+0.23$ $[-0.01, +0.48]$ (176/144)	-0.01 $[-0.01, +0.00]$ (2/3)	$+0.23$ $[-0.01, +0.49]$	–
		screened	$+0.21$ $[-0.03, +0.46]$ (174/144)	-0.01 $[-0.01, +0.00]$ (2/3)	$+0.22$ $[-0.02, +0.47]$	–
MMLU, O+F, 8	averaging	unscreened	$+0.19$ $[-0.20, +0.57]$ (397/371)	-1.01 $[-1.22, -0.80]$ (283/424)	$+1.19$ $[+0.87, +1.53]$	g
		screened	$+0.21$ $[-0.16, +0.59]$ (388/358)	-0.98 $[-1.19, -0.78]$ (281/419)	$+1.19$ $[+0.87, +1.53]$	g
	vote	unscreened	$+0.25$ $[-0.06, +0.58]$ (286/251)	-0.55 $[-0.67, -0.44]$ (135/213)	$+0.80$ $[+0.51, +1.11]$	g
		screened	$+0.26$ $[-0.05, +0.59]$ (279/242)	-0.52 $[-0.64, -0.41]$ (134/207)	$+0.78$ $[+0.50, +1.08]$	g
	confidence	unscreened	$+0.04$ $[-0.41, +0.49]$ (537/531)	-2.69 $[-2.99, -2.40]$ (454/831)	$+2.73$ $[+2.36, +3.14]$	g
		screened	$+0.02$ $[-0.41, +0.46]$ (490/487)	-2.14 $[-2.44, -1.86]$ (431/731)	$+2.16$ $[+1.80, +2.55]$	g
	CV logistic	unscreened	$+0.60$ $[+0.28, +0.92]$ (312/228)	-0.06 $[-0.10, -0.03]$ (20/28)	$+0.66$ $[+0.35, +0.99]$	g
		screened	$+0.56$ $[+0.24, +0.87]$ (303/225)	-0.06 $[-0.10, -0.02]$ (20/28)	$+0.62$ $[+0.30, +0.94]$	g

Table 44: Registered endpoints of S8-B: in each fold, the candidate with the highest training-fold accuracy answers every test item of the fold; folds and bootstrap as in Table 41.

	MMLU, O, 4 adapters	MMLU, F, 4 adapters	MMLU, O+F, 8 adapters
Training-selected fixed candidate: gain, passes	$+0.00$, 1, base in every fold	$+0.00$, 1, base in every fold	$+0.00$, 1, base in every fold
Screened logistic − fixed candidate	$+0.43$ $[+0.11, +0.74]$	$+0.21$ $[-0.03, +0.46]$	$+0.56$ $[+0.24, +0.87]$
Rule (h) met	h	–	h

Table 45: S8-C: terms of Proposition 5 for every S7-A arm, in points of accuracy over all items. $\mathrm{Gain}_{\mathrm{in}}$: base errors in the helpable set answered correctly; $\mathrm{Gain}_{\mathrm{out}}$: base errors outside it answered correctly, possible only for averaging; $\mathrm{Imp} = \mathrm{Gain}_{\mathrm{in}} + \mathrm{Gain}_{\mathrm{out}} - \mathrm{Damage}$ and $\mathrm{Gain}_{\mathrm{in}} = \mathrm{OC} - \mathrm{Miss}$ hold exactly in every row.

Target	Selector	Arm	$\mathrm{Gain}_{\mathrm{in}}$	$\mathrm{Gain}_{\mathrm{out}}$	Miss	Damage	Imp
MMLU, O, OC 8.28	averaging	unscreened	2.82	0.01	5.46	2.86	−0.03
		screened	2.73	0.01	5.55	2.73	+0.01
	vote	unscreened	1.55	0.00	6.74	1.49	+0.06
		screened	1.52	0.00	6.77	1.44	+0.08
	confidence	unscreened	3.87	0.00	4.42	4.11	−0.24
		screened	3.43	0.00	4.86	3.70	−0.27
	CV logistic	unscreened	2.01	0.00	6.27	1.55	+0.46
		screened	1.94	0.00	6.34	1.52	+0.43
MMLU, F, OC 7.51	averaging	unscreened	3.00	0.01	4.52	3.08	−0.07
		screened	2.93	0.01	4.58	2.87	+0.08
	vote	unscreened	2.22	0.00	5.29	2.62	−0.40
		screened	2.09	0.00	5.43	2.06	+0.03
	confidence	unscreened	3.23	0.00	4.28	3.49	−0.26
		screened	3.18	0.00	4.34	3.30	−0.12
	CV logistic	unscreened	1.25	0.00	6.26	1.03	+0.23
		screened	1.24	0.00	6.27	1.03	+0.21
MMLU, O+F, OC 10.61	averaging	unscreened	2.83	0.00	7.78	2.64	+0.19
		screened	2.76	0.00	7.85	2.55	+0.21
	vote	unscreened	2.04	0.00	8.57	1.79	+0.25
		screened	1.99	0.00	8.62	1.72	+0.26
	confidence	unscreened	3.82	0.00	6.79	3.78	+0.04
		screened	3.49	0.00	7.12	3.47	+0.02
	CV logistic	unscreened	2.22	0.00	8.39	1.62	+0.60
		screened	2.16	0.00	8.45	1.60	+0.56

Table 46: Registered endpoints of S10-A: the policies of Tables 41, 42 and 44 under 20 question-group fold assignments, of which assignment 0 is the one of S7. Each cell: median [minimum, maximum] over the 20 assignments, in points; in parentheses, the number of assignments whose 95% question-cluster bootstrap lower bound exceeds 0 for gains and differences, or -0.25 for screened minus unscreened. Rules (i) and (j) as registered in Section P; "–": rule not met.

	MMLU, O, 4 adapters	MMLU, F, 4 adapters	MMLU, O+F, 8 adapters
Logistic, unscreened: gain	+0.42 [+0.37, +0.50] (20/20)	+0.22 [+0.16, +0.28] (7/20)	+0.58 [+0.50, +0.65] (20/20)
Logistic, screened: gain	+0.38 [+0.32, +0.46] (20/20)	+0.21 [+0.15, +0.27] (2/20)	+0.55 [+0.46, +0.59] (20/20)
Subject-aware logistic, screened: gain	+0.43 [+0.38, +0.48] (20/20)	+0.20 [+0.16, +0.24] (1/20)	+0.61 [+0.56, +0.67] (20/20)
Subject lookup: gain	+0.57 [+0.28, +0.76] (17/20)	−0.19 [−0.44, −0.04] (0/20)	+0.63 [+0.39, +1.00] (18/20)
Training-selected fixed candidate: gain	+0.00 [−0.26, +0.00] (0/20)	+0.00 [+0.00, +0.00] (0/20)	+0.00 [−0.26, +0.00] (0/20)
Screened − unscreened logistic	−0.04 [−0.05, −0.03] (20/20)	−0.01 [−0.01, −0.01] (20/20)	−0.03 [−0.06, −0.02] (20/20)
Screened logistic − lookup	−0.20 [−0.43, +0.12] (0/20)	+0.36 [+0.28, +0.71] (13/20)	−0.11 [−0.53, +0.19] (0/20)
Screened subject-aware − lookup	−0.14 [−0.30, +0.18] (0/20)	+0.38 [+0.20, +0.63] (18/20)	−0.03 [−0.39, +0.21] (0/20)
Screened logistic − fixed candidate	+0.40 [+0.32, +0.58] (20/20)	+0.21 [+0.15, +0.27] (2/20)	+0.55 [+0.46, +0.85] (20/20)
Rule (i) met	i	–	i
Rule (j) met	j	j	j

Table 47: Registered endpoints of S10-B: all policies trained on the same random subsample of n training rows, drawn as whole question groups, in each fold of S7-A, 20 draws per budget. Top: gains over base in points of the draw-averaged per-item accuracy; bottom: paired differences. 95% question-cluster bootstrap intervals. Passes per item as in Tables 41 and 42. Rule (k) as registered; "–": rule not met.

Target	n	Screened logistic	Screened subject-aware	Subject lookup	Fixed candidate
MMLU, O, 4	400	$+0.25$ $[-0.01, +0.52]$	$+0.27$ $[+0.00, +0.53]$	$+0.36$ $[+0.13, +0.58]$	-0.24 $[-0.44, -0.04]$
	1,000	$+0.32$ $[+0.05, +0.59]$	$+0.36$ $[+0.10, +0.63]$	$+0.39$ $[+0.11, +0.65]$	-0.22 $[-0.40, -0.06]$
	2,000	$+0.35$ $[+0.07, +0.64]$	$+0.39$ $[+0.12, +0.67]$	$+0.50$ $[+0.20, +0.79]$	-0.19 $[-0.36, -0.02]$
MMLU, F, 4	400	$+0.10$ $[-0.13, +0.34]$	$+0.10$ $[-0.13, +0.33]$	-0.57 $[-0.72, -0.42]$	-0.40 $[-0.59, -0.21]$
	1,000	$+0.16$ $[-0.05, +0.39]$	$+0.16$ $[-0.04, +0.38]$	-0.53 $[-0.71, -0.35]$	-0.16 $[-0.29, -0.03]$
	2,000	$+0.23$ $[+0.00, +0.46]$	$+0.18$ $[-0.03, +0.41]$	-0.55 $[-0.75, -0.35]$	-0.19 $[-0.31, -0.08]$
MMLU, O+F, 8	400	$+0.30$ $[+0.03, +0.59]$	$+0.32$ $[+0.05, +0.61]$	$+0.12$ $[-0.12, +0.35]$	-0.30 $[-0.52, -0.08]$
	1,000	$+0.40$ $[+0.13, +0.69]$	$+0.45$ $[+0.18, +0.72]$	$+0.21$ $[-0.08, +0.50]$	-0.24 $[-0.42, -0.06]$
	2,000	$+0.48$ $[+0.20, +0.79]$	$+0.54$ $[+0.26, +0.84]$	$+0.29$ $[-0.03, +0.60]$	-0.26 $[-0.45, -0.09]$

Target	n	Logistic − lookup	Subject-aware − lookup	Logistic − fixed	Rule (k)
MMLU, O, 4	400	-0.11 $[-0.33, +0.13]$	-0.09 $[-0.31, +0.14]$	$+0.49$ $[+0.34, +0.65]$	–
	1,000	-0.07 $[-0.31, +0.18]$	-0.02 $[-0.26, +0.22]$	$+0.54$ $[+0.36, +0.72]$	–
	2,000	-0.15 $[-0.41, +0.12]$	-0.11 $[-0.36, +0.15]$	$+0.55$ $[+0.32, +0.77]$	–
MMLU, F, 4	400	$+0.68$ $[+0.51, +0.85]$	$+0.67$ $[+0.51, +0.84]$	$+0.50$ $[+0.35, +0.65]$	k
	1,000	$+0.70$ $[+0.53, +0.87]$	$+0.69$ $[+0.54, +0.86]$	$+0.32$ $[+0.17, +0.49]$	k
	2,000	$+0.77$ $[+0.59, +0.96]$	$+0.73$ $[+0.55, +0.91]$	$+0.42$ $[+0.24, +0.61]$	k
MMLU, O+F, 8	400	$+0.18$ $[-0.04, +0.43]$	$+0.20$ $[-0.03, +0.45]$	$+0.60$ $[+0.44, +0.78]$	–
	1,000	$+0.19$ $[-0.06, +0.47]$	$+0.24$ $[-0.01, +0.51]$	$+0.64$ $[+0.45, +0.84]$	–
	2,000	$+0.19$ $[-0.09, +0.49]$	$+0.24$ $[-0.02, +0.54]$	$+0.74$ $[+0.51, +0.99]$	–

Table 48: Registered endpoints of S10-C: the logistic selector and the fixed candidate fitted once on all 14,042 MMLU rows of a pool, frozen with the threshold, and applied to the 2,000 MMLU-CF items with the same pool; gains over base in points with 95% item-bootstrap intervals. Rules (l) and (m) as registered; "–": rule not met.

	MMLU-CF, O, 4 adapters	MMLU-CF, F, 4 adapters
Frozen logistic, unscreened, 5 passes	-0.90 $[-1.70, -0.10]$	-0.20 $[-0.75, +0.35]$
Frozen logistic, screened, 2.58 passes	-0.90 $[-1.70, -0.10]$	-0.20 $[-0.75, +0.35]$
Frozen fixed candidate	$+0.00$, base	$+0.00$, base
Screened − unscreened	$+0.00$ $[-0.15, +0.15]$	$+0.00$ $[+0.00, +0.00]$
Rule (l) met	–	–
Rule (m) met	m	m