THE POST-TRAINING TAX: THE MOVING TARGET OF CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models typically achieve higher predictive accuracy after post-training, but a more fundamental question is whether this improvement in accuracy is accompanied by better confidence quality. To investigate this question, we examine the changes induced by post-training from two aspects of confidence: reliability and discrimination. Across 47 model–task cells comprising 16 base/tuned model pairs, 6 post-training settings, and 3 multiple-choice tasks, we find that in the 35-cell alignment cohort, post-training improves average accuracy by 1.1 percentage points while increasing the Brier reliability component by 0.047, with reliability degrading in 32 of 35 cells; discrimination remains largely stable. We refer to this phenomenon as the reliability tax of post-training. To mitigate this reliability loss, we further evaluate several confidence calibration methods and find that simple temperature scaling can substantially reduce calibration error, but its effectiveness is not fixed: when the model, task, and prompt template are held constant while the evaluation pool changes, the median temperature required for harder pools is approximately 5 times that for easier pools, and the excess ECE induced by calibration transfer is strongly correlated with the accuracy gap between pools. Further experiments show that only 20 target labels can substantially improve calibration on a new evaluation pool, reducing held-out ECE by approximately 83.4% relative to the uncalibrated state. Overall, our study reveals the reliability tax introduced by post-training and demonstrates that calibration is strongly dependent on the evaluation pool, while a small amount of target data can support effective recalibration on new evaluation pools.

1 INTRODUCTION

For large language models, producing an answer is often only part of what a downstream system needs. In settings involving abstention, human escalation, and trust in model outputs, downstream systems also require a meaningful confidence signal to assess how much the model should be trusted in its prediction (Guo et al., 2017; Jiang et al., 2021; Geng et al., 2024). However, while post-training can improve predictive performance, it may simultaneously degrade the quality of the model's confidence estimates (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023; Zhu et al., 2023; He et al., 2023; Leng et al., 2025; Xiao et al., 2025; Wang et al., 2025b). This raises a natural question: **when predictive accuracy improves, does the model's confidence become more reliable as well?**

Answering this question requires distinguishing two aspects of confidence: **reliability**, namely whether confidence matches the actual probability of being correct, and **discrimination**, namely whether confidence can distinguish correct predictions from incorrect ones. These two aspects are not equivalent, and therefore predictive accuracy or discrimination alone cannot fully characterize how post-training affects confidence. When confidence exhibits reliability errors, prior work provides a range of calibration methods to mitigate them, including temperature scaling, which learns a temperature parameter from labeled data, and approaches that use the pretrained base model as a label-free reference (Guo et al., 2017; Luo et al., 2025; Tan et al., 2026). These methods provide direct ways to repair confidence estimates, but they also raise a further question: **when used to mitigate the reliability loss induced by post-training, does their calibration effectiveness change across evaluation pools?** Even with the model, task, and prompt template fixed, changes in evalua-

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.

tion difficulty may alter the relationship between confidence and correctness, raising the question of whether calibration fitted on one pool remains effective on another.

We examine post-training changes in confidence reliability and discrimination, and study how calibration behaves across controlled difficulty shifts under fixed models, tasks, and prompt templates. Here, evaluation pools denote controlled difficulty strata constructed within each fixed benchmark, allowing us to isolate instance difficulty as the primary factor associated with changes in calibration. We find that post-training improves predictive accuracy while introducing a systematic loss in reliability, which we refer to as the **reliability tax**. We further find that calibration depends strongly on the evaluation pool: temperatures fitted on one pool can become substantially mismatched on another, revealing a **moving target** in confidence calibration. Finally, we find that a small amount of labeled target-pool data can substantially reduce the resulting calibration error and recover reliable confidence on a new pool.

2 Related Work

2.1 Confidence Calibration and Post-Training

Confidence calibration aims to align a model's predicted confidence with its actual probability of being correct. Classical post-hoc calibration methods include Platt scaling, isotonic regression, and temperature scaling (Guo et al., 2017). Pretrained language models can be reasonably calibrated in-distribution but deteriorate under distribution shifts (Desai & Durrett, 2020; Chen et al., 2023). Reliable confidence supports selective prediction through abstention or human review (Jiang et al., 2021; Geng et al., 2024). Adaptive calibrators account for differences across samples and tasks. Adaptive Temperature Scaling (ATS) adjusts the calibration temperature (Xie et al., 2024); SMART conditions the temperature on the prediction margin (Guo et al., 2026); Density-Aware Calibration (DAC) incorporates sample-density information into calibration (Tomani et al., 2023); and Thermometer learns across tasks to calibrate unseen tasks (Shen et al., 2024). Calibration-aware tuning instead directly modifies model parameters (Kapoor et al., 2024; Tao et al., 2024).

Prior work decomposes proper scoring rules into calibration and refinement components (Murphy, 1973; DeGroot & Fienberg, 1983), with subsequent work introducing within-bin corrections (Stephenson et al., 2007). These decompositions support separate analyses of reliability and discrimination. Instruction tuning, preference optimization, and other alignment procedures can alter confidence even when improving predictive performance (Zhu et al., 2023; He et al., 2023; Leng et al., 2025; Xiao et al., 2025; Wang et al., 2025b). Because these studies assess confidence with aggregate metrics such as ECE or the Brier score, which conflate calibration error with discrimination, they cannot determine whether post-training impairs the ranking of correct versus incorrect answers or merely rescales probabilities; the decompositions above separate the two; see Section 3.1. Huang et al. (2026) find that instruction tuning can increase confidence in low-resource languages without comparable gains in accuracy. Base-model references offer another approach to post-training calibration. DACA calibrates post-trained models by matching their probability distributions to those of the base model (Luo et al., 2025); BaseCal uses the base model to rescore post-trained responses or maps post-trained states into the base model's representation space (Tan et al., 2026), providing label-free or low-label calibration signals.

2.2 Label-Free Uncertainty Estimation

Label-free methods estimate uncertainty from model outputs or generation behavior. Likelihood-based methods use token- or sequence-level probabilities (Malinin & Gales, 2021; Mahaut et al., 2024); another group of methods uses model self-evaluation to estimate the model's confidence in its answers (Kadavath et al., 2022; Lin et al., 2022b; Tian et al., 2023; Xiong et al., 2024; Zhang et al., 2024). In addition, consistency across multiple sampled generations and semantic entropy have been widely used to characterize uncertainty in generated responses (Manakul et al., 2023; Lin et al., 2024; Kuhn et al., 2023; Farquhar et al., 2024; Nikitin et al., 2024). However, uncertainty signals derived from a model's own behavior do not necessarily reflect its true reliability. When a model exhibits high consistency or confidence in an incorrect answer, consistency-based uncertainty estimation can likewise produce overly confident signals (Tan et al., 2025; Simhi et al., 2025). Lyu et al. (2025) estimate model confidence from consistency across sampled generations and find

that instruction tuning can make calibration more challenging. Liu (2026) further studies response homogenization in aligned models and its implications for sampling-based uncertainty estimation.

3 METHOD

3.1 SEPARATING RELIABILITY FROM DISCRIMINATION

To distinguish confidence reliability from its ability to separate correct and incorrect predictions, we analyze model confidence from three complementary perspectives: reliability, resolution, and discrimination. For a multiple-choice dataset $D = \{(x_i, y_i^*)\}_{i=1}^N$, let $\hat{y}_i = \arg\max_a P(a \mid x_i)$ be the selected option, $c_i = \max_a P(a \mid x_i)$ its confidence, and $z_i = \mathbb{1}[\hat{y}_i = y_i^*]$ its correctness. With $M{=}10$ equal-width bins, the Expected Calibration Error (ECE) is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|C_m|}{N} \left|\mathrm{acc}(C_m) - \mathrm{conf}(C_m)\right|. \tag{1}$$

We retain ECE for comparability with prior work, while using the Murphy decomposition to further characterize confidence quality (Murphy, 1973; DeGroot & Fienberg, 1983). The items are partitioned into $K{=}10$ equal-frequency bins $\{B_k\}$ according to confidence rank. Ties are broken by a stable sort that preserves the original input order, and base and tuned checkpoints are binned independently. The decomposition is

$$\underbrace{\frac{1}{N}\sum_i (\bar{c}_{b(i)} - z_i)^2}_{\text{Brier of bin-mean forecasts}} = \underbrace{\sum_k \frac{|B_k|}{N}\left(\mathrm{conf}(B_k) - \mathrm{acc}(B_k)\right)^2}_{\text{REL (lower is better)}} - \underbrace{\sum_k \frac{|B_k|}{N}\left(\mathrm{acc}(B_k) - \bar{z}\right)^2}_{\text{RES (higher is better)}} + \underbrace{\bar{z}(1-\bar{z})}_{\text{UNC}}, \tag{2}$$

where $b(i)$ identifies the bin containing item i, $\bar{c}_k = \mathrm{conf}(B_k)$, and $\bar{z}$ denotes overall accuracy. This identity is exact for forecasts replaced by their bin means, but does not generally hold for the original item-level confidence values. For the latter, the top-label Brier score $\mathrm{BS} = \frac{1}{N}\sum_i (c_i - z_i)^2$ is evaluated against binary correctness rather than the multiclass simplex and decomposes as

$$\mathrm{BS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC} + \sum_k \frac{|B_k|}{N}\left[\mathrm{Var}_k(c) - 2\,\mathrm{Cov}_k(c, z)\right], \tag{3}$$

where the within-bin variance and covariance use denominator $|B_k|$; the correction vanishes when confidence is constant within every bin (Stephenson et al., 2007). In this decomposition, **reliability** measures the mismatch between confidence and empirical accuracy within confidence groups, with lower values indicating better agreement. **Resolution** measures the variation in empirical accuracy across confidence groups, with higher values indicating stronger separation among confidence levels. We additionally report $\mathrm{RES}/\mathrm{UNC}$ when $\mathrm{UNC} > 0$ and use AUROC for predicting correctness from confidence as a complementary measure of discrimination. AUROC and resolution are distinct: equal AUROC does not imply equal resolution, and base and tuned checkpoints may produce different predictions for the same inputs, resulting in different correctness labels.

3.2 INVARIANCE AND SENSITIVITY UNDER CONFIDENCE TRANSFORMATION

Different calibration methods transform confidence in different ways, and transformations need not affect all aspects of confidence equally. To characterize which properties remain invariant and which may change after calibration, we consider strictly monotone transformations of scalar confidence.

Let g be strictly increasing and applied to a scalar confidence, yielding $c_i' = g(c_i)$, while predictions and correctness labels are held fixed. Under rank-based equal-frequency binning with a fixed tie-breaking rule, g preserves the ordering of confidence values and therefore every bin membership. Binned RES and AUROC are consequently exactly invariant. Binary temperature scaling applied to a scalar log-odds score and positive-slope Platt scaling satisfy this condition. Isotonic regression is nondecreasing but can introduce ties, which breaks exact AUROC invariance.

For multiclass temperature scaling, confidence is defined as

$$c_i(T) = \frac{\exp(a_{i,\hat{y}_i}/T)}{\sum_j \exp(a_{ij}/T)}, \qquad T > 0, \tag{4}$$

where a_{ij} denotes the logit for option j of item i. Temperature scaling preserves the argmax prediction, but $c_i(T)$ depends on the complete logit vector. Thus, multiclass temperature scaling is not a uniform monotone transformation of confidence across items: cross-item confidence ranks can change, altering AUROC and rank-based resolution (Guo et al., 2017). Consequently, no single confidence metric characterizes the effect of calibration. The subsequent analyses separately examine **reliability, resolution, discrimination, and calibration repair**.

4 Experiments

We organize the experiments around four questions: whether post-training introduces a reliability tax, whether calibration repairs it, whether the repair transfers across pools, and how much target-pool data recalibration needs. Model, task, and protocol details: Section 4.1 and Appendix A.2.

4.1 Experimental Setup

4.1.1 Models and Tasks

We evaluate 16 architecture-matched base/tuned pairs spanning Qwen2.5 0.5B–32B, Llama-3.1-8B, OLMo-2-1B, and Tulu-3-8B. The post-training settings include instruction tuning, SFT, DPO, final-stage training, Qwen2.5-Math, and R1 distillation (Guo et al., 2025). The main analysis contains 47 model–task cells across MMLU-Pro (Wang et al., 2024), its mathematics subset, and ARC-Challenge; a broader survey also includes TruthfulQA-mc2 (Lin et al., 2022a) and GSM8K (Cobbe et al., 2021). Multiple-choice confidence is the softmax probability over legal answer options under 3-shot plain completion. TruthfulQA and GSM8K use task-specific sequence scoring and are compared only within the same protocol. See Appendices B.1, A.1, and C.1 for the broader model matrix, scoring checks, and self-evaluation comparisons.

4.1.2 Evaluation and Calibration Protocol

Each pool is split evenly into a calibration half D_A and an independent reporting half D_B: $586/586$ items for ARC, $1000/1000$ for MMLU-Pro, and $675/676$ for MMLU-math. Recalibrators are fitted on D_A and evaluated on D_B; temperature scaling minimizes negative log-likelihood. Base/tuned comparisons use paired bootstrap resampling over shared item indices, while adaptive-calibrator comparisons bootstrap model–task summaries. Accuracy changes are reported in percentage points; ECE, reliability, resolution, and AUROC retain their original scales. See Appendices A.1 and D.3 for fitting, resampling, and dependence checks.

4.2 Does Post-Training Impose a Reliability Tax?

4.2.1 The Shape of the Tax

Figure 1 summarizes base-to-tuned changes in accuracy and ECE, where positive ECE changes indicate calibration degradation. Confidence changes are not determined by accuracy changes alone: in some cells post-training improves accuracy and reduces ECE; in others both rise. For example, on MMLU-math, Qwen-7B-Instruct has ECE 0.380 versus 0.061 for Qwen2.5-Math-7B; on ARC, all three OLMo-2-1B post-trained checkpoints exhibit higher ECE than their base models. Table 1 summarizes 47 model–task cells, including 16 base/tuned pairs, six post-training settings, and three main tasks. The main group contains 35 cells covering instruction tuning, SFT, DPO, and RLVR, with a mean accuracy change of $+1.1$ pp; the control

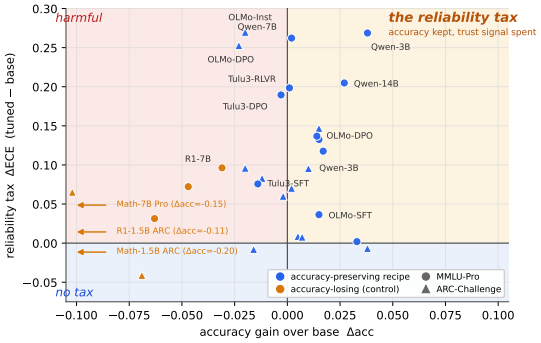


Figure 1: **The reliability-tax landscape.** Changes in accuracy (horizontal) and ECE (vertical). Upper-right points exhibit the tax; circles denote MMLU-Pro and triangles ARC-Challenge.

group contains 12 cells involving mathematical RL and R1 distillation, with a mean accuracy change of -7.9 pp. These drops partly reflect the evaluation format—plain completion suppresses the chain-of-thought scratchpad these reasoning models are trained to use; we treat the cohort as an accuracy-varying contrast, not as evidence about reasoning training itself.

Table 1: **The reliability tax across 47 model–task cells.** Base-to-tuned changes in mean accuracy, reliability, and AUROC. † marks a 95% paired-bootstrap AUROC interval excluding zero; full intervals appear in Table 8.

		ARC		MMLU-Pro		MMLU-math	
Pair	**mean ΔAcc. [pp]**	ΔREL	ΔAUROC	ΔREL	ΔAUROC	ΔREL	ΔAUROC
instruction tuning, 7 pairs, 20 cells: mean ΔREL $= +0.054$, *mean* ΔAUROC $= +0.001$							
Llama-8B	$+3.7$	-0.001	-0.008	$+0.002$	$+0.024$	$+0.005$	$+0.016$
OLMo-1B-Inst	-0.6	$+0.086$	-0.005	—	—	$+0.056$	$+0.030$
Qwen-1.5B	$+0.6$	$+0.017$	$+0.010$	$+0.041$	-0.019	$+0.055$	$+0.020$
Qwen-14B	$+2.5$	$+0.004$	$+0.017$	$+0.070$	-0.011	$+0.087$	$-0.034^{\dagger}$
Qwen-32B	$+1.5$	$+0.002$	-0.007	$+0.038$	$+0.010$	$+0.034$	$+0.019$
Qwen-3B	$+2.2$	$+0.027$	$+0.003$	$+0.114$	-0.019	$+0.158$	$+0.003$
Qwen-7B	$+1.0$	$+0.019$	$+0.004$	$+0.103$	-0.007	$+0.153$	-0.025
SFT, 2 pairs, 6 cells: mean ΔREL $= +0.011$, *mean* ΔAUROC $= +0.013$							
OLMo-1B-SFT	$+1.6$	$+0.024$	$+0.009$	$+0.013$	$+0.020$	-0.005	$+0.037$
Tulu3-SFT	-0.6	$+0.000$	-0.005	$+0.023$	$+0.028^{\dagger}$	$+0.014$	-0.012
DPO, 2 pairs, 6 cells: mean ΔREL $= +0.053$, *mean* ΔAUROC $= +0.011$							
OLMo-1B-DPO	$+0.2$	$+0.076$	$+0.013$	$+0.058$	$+0.049^{\dagger}$	$+0.017$	$+0.047$
Tulu3-DPO	-0.0	$+0.021$	-0.016	$+0.079$	$+0.001$	$+0.066$	-0.030
RLVR, 1 pair, 3 cells: mean ΔREL $= +0.063$, *mean* ΔAUROC $= -0.015$							
Tulu3-RLVR	-0.0	$+0.028$	-0.001	$+0.084$	-0.005	$+0.078$	-0.040
math RL, 2 pairs, 6 cells: mean ΔREL $= +0.002$, *mean* ΔAUROC $= -0.045$							
Math-1.5B	-9.0	$+0.000$	$-0.097^{\dagger}$	$+0.005$	$-0.082^{\dagger}$	$+0.002$	$+0.035$
Math-7B	-8.7	$+0.004$	$-0.065^{\dagger}$	$+0.007$	$-0.034^{\dagger}$	-0.004	-0.024
R1 distillation, 2 pairs, 6 cells: mean ΔREL $= +0.015$, *mean* ΔAUROC $= -0.049$							
R1-1.5B	-7.9	$+0.001$	$-0.047^{\dagger}$	$+0.019$	-0.008	$+0.020$	$-0.070^{\dagger}$
R1-7B	-6.1	-0.003	$-0.067^{\dagger}$	$+0.029$	$-0.045^{\dagger}$	$+0.026$	$-0.054^{\dagger}$

4.2.2 Reliability Degrades More Consistently Than Discrimination

Figure 2 separates changes in confidence reliability and discrimination. In the main group, reliability error increases significantly in 32 of 35 model–task cells, with a mean increase of 0.0470, whereas AUROC-change intervals contain zero in 32 cells.

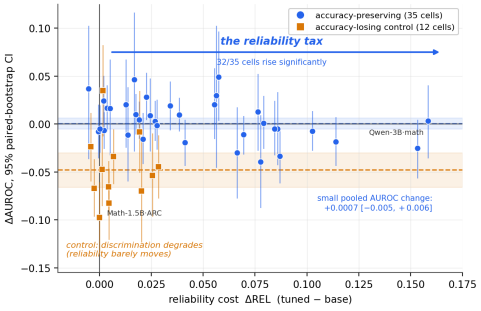


Figure 2: **Reliability and discrimination changes.** Binned reliability versus AUROC changes with 95% paired-bootstrap intervals for 47 cells. Circles denote accuracy-preserving recipes and squares denote controls.

The comparison group shows a smaller reliability increase of $+0.0088$ and a mean AUROC decrease of -0.0465. At the pooled level, the AUROC change in the main group is $+0.0007$ with a 95% CI of $[-0.0050, 0.0064]$, indicating little systematic change in discrimination; see Appendix B.6. As shown in Figure 3, the same pattern emerges along the Tulu-3 pipeline from SFT to DPO and RLVR (Lambert et al., 2025): reliability error increases at each stage on all three tasks, accuracy changes by no more than 2.0 pp relative to the common base model, and all but one AUROC-change interval contain zero. The OLMo-2-1B sequence shows the same ordering; see Appendix B.1. Taken together, the tax is primarily a reliability phenomenon: confidence retains much of its ability to distinguish correct from incorrect predictions, while its mapping to empirical accuracy degrades. The Brier decomposition provides a complementary view. As shown in Table 2, ΔECE correlates with ΔREL at $r = +0.96$, compared with only $+0.08$ for ΔRES and $+0.10$ for ΔAUROC. In the

main group, reliability accounts for 93.4% of the combined mean absolute changes in reliability and resolution, compared with 46.4% in the comparison group.

Table 2: **Brier decomposition by recipe.** Unweighted mean base-to-tuned component changes over 47 cells, using unnormalized resolution. Per-cell results appear in Appendix B.2.

Recipe	Cells	ΔAcc. [pp]	ΔECE	ΔREL	ΔRES	ΔAUROC
Main group: instruction/SFT/DPO/RLVR						
instruction tuning	20	+1.7	+0.1423	+0.0536	-0.0005	+0.0011
SFT	6	+0.5	+0.0388	+0.0114	+0.0015	+0.0129
DPO	6	+0.1	+0.1412	+0.0528	-0.0001	+0.0106
RLVR	3	<0.1	+0.1507	+0.0632	-0.0010	-0.0153
all 35 cells	35	+1.1	+0.1251	+0.0470	-0.0002	+0.0034
Comparison group: math-RL/distillation						
mathematical RL	6	-8.8	+0.0228	+0.0022	-0.0065	-0.0446
R1 distillation	6	-7.0	+0.0547	+0.0154	-0.0105	-0.0485
all 12 cells	12	-7.9	+0.0387	+0.0088	-0.0085	-0.0465

The result is stable across alternative binning choices; detailed per-cell results and binning checks are provided in Appendices B.2 and B.4. The main group's mean reliability increase is 0.0470, while mean resolution changes by only −0.0002. The comparison group instead loses 0.0085 in resolution, alongside its AUROC decline. These component changes distinguish confidence–accuracy mismatch from reduced separation. The reliability share excludes uncertainty and the within-bin correction, so it is not a percentage of total Brier-score change. This pattern suggests that the primary effect of post-training is a shift in the level at which confidence is expressed relative to correctness, rather than a fundamental loss of confidence ranking. Thus, similar discrimination can coexist with substantially degraded reliability, leaving the confidence–accuracy mapping amenable to post-hoc recalibration.

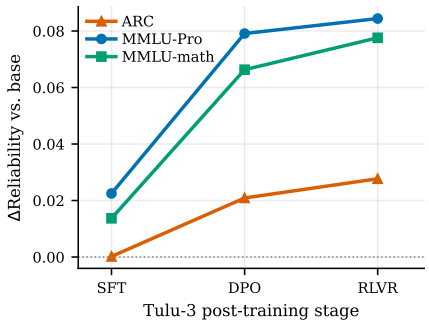


Figure 3: **The tax grows along the pipeline.** Reliability cost at Tulu-3's SFT, DPO and RLVR stages.

4.3 Can Calibration Repair the Reliability Tax?

4.3.1 Supervised Recalibration

Table 3 compares supervised recalibration across 47 model–task cells, with calibrators fitted on D_A and evaluated on an independent D_B. Platt scaling and isotonic regression operate on top-label confidence, matching the top-label metrics reported; full-simplex calibrators such as Dirichlet calibration (Kull et al., 2019) are a natural extension. Counting per-cell minima across all seven configurations in the table, temperature scaling achieves the lowest ECE in 19 cells: 12 for the NLL fit and 7 for the ECE-grid fit. Platt scaling is lowest in 12 cells and isotonic regression in 6. In the remaining ten cells, the base checkpoint is lowest in 6, the tuned checkpoint in 2, and DACA in 2. The median absolute ECE difference between temperature and Platt scaling is 0.007. The mean paired difference is −0.0018, with a 95% interval of $[-0.0051, +0.0012]$. On Qwen-3B MMLU-math, for instance, temperature scaling reduces ECE from 0.376 to 0.026. Calibration can also modify confidence rankings: they change in 38 of 47 cells under temperature scaling while predictions stay fixed. The median $|\Delta\text{AUROC}|$ is 0.0032 and the maximum is 0.0218. See Appendix Figure 8 for the risk–coverage curves.

4.3.2 Base-Model References

Base-model references are effective on standard alignment pipelines: across the 35 main-group cells, DACA reduces mean held-out ECE from the uncalibrated tuned value of 0.1770 to 0.0504. This is a

Table 3: **Held-out recalibration across 47 model–task cells.** Mean ECE on D_B after fitting on D_A; "Best in" counts cell-wise minima. Detailed results appear in Tables 9 and 17.

	by task			by group			
Confidence source	**ARC**	**MMLU-Pro**	**MMLU-math**	**all 47**	**main 35**	**control 12**	**Best in**
Uncalibrated							
base checkpoint	0.0351	0.0661	0.0704	0.0570	0.0542	0.0654	6/47
tuned checkpoint	0.1040	0.1910	0.1813	0.1581	0.1770	0.1029	2/47
Supervised, labels from the fitting pool							
temperature, NLL	**0.0312**	0.0315	**0.0277**	**0.0301**	**0.0299**	**0.0306**	**12**/47
temperature, ECE grid	0.0356	0.0359	0.0299	0.0338	0.0332	0.0355	7/47
Platt, top-label	**0.0312**	**0.0307**	0.0338	0.0319	0.0300	0.0375	**12**/47
isotonic, top-label	0.0433	0.0387	0.0376	0.0399	0.0379	0.0457	6/47
Label-free, no correctness labels							
DACA, base-referenced	0.0612	0.0840	0.0693	0.0712	0.0504	0.1319	2/47

Table 4: **Base-referenced repair on harder pools.** Mean ECE across 15 checkpoints per task. Base-referenced methods use the full unlabeled target pool; per-checkpoint results appear in Table 25.

	ECE on harder pools ↓		
Method / reference	**ARC**	**MMLU-Pro**	**MMLU-math**
Tuned model, uncalibrated	0.172	0.292	0.250
Source-fitted			
Transferred temperature	0.218	0.213	0.148
Base-referenced			
Mean-match to base	0.104	0.166	0.126
Quantile map to base	0.105	0.166	0.126
Reference predictor			
Base checkpoint	0.067	0.171	0.137

71% reduction without correctness labels and slightly improves on the uncalibrated base checkpoint value of 0.0542. Their dependence on the reference's own calibration is the failure mode. On ARC, Llama-8B has a base ECE of 0.042 and a tuned ECE of 0.038; matching the tuned model to the base-model confidence distribution raises ECE to 0.098. The control-group base-reference mean is likewise 0.132, compared with 0.103 without calibration. When the base model is itself well calibrated, base-referenced methods are also competitive on harder pools, as shown in Table 4. See Appendix C.3 for cross-task transfer.

4.4 Is Calibration a Moving Target?

We isolate pool dependence by varying evaluation difficulty while holding the model, task, and prompt template fixed.

4.4.1 Pool-Dependent Calibration

Within each task, we construct controlled-difficulty pools using mean correctness across the other 14 checkpoints; this ranking is never used for calibration fitting. Across 45 model–task cells, the harder pool requires a higher temperature in every case, with a median ratio of 4.98; across 90 pools, $\log T$ correlates with accuracy at $r = -0.710$. Seven harder-pool fits reach $T{=}200$, all on near-chance pools where a near-uniform NLL solution is expected; see Appendix D.1. The result persists after controlling for option count, MMLU-Pro subcategory, and base-model geometry, with median ratios of 5.19–5.22 and 3.13–3.23; see Appendix D.2.

4.4.2 Calibration Transfer under Difficulty Shifts

Transporting an easier-pool temperature to a harder pool can worsen calibration. On ARC, mean ECE rises from 0.172 without calibration to 0.218. It still helps on the more overconfident MMLU pools, but remains well behind target-refitted calibrators on all three tasks; see Table 4. Figure 4

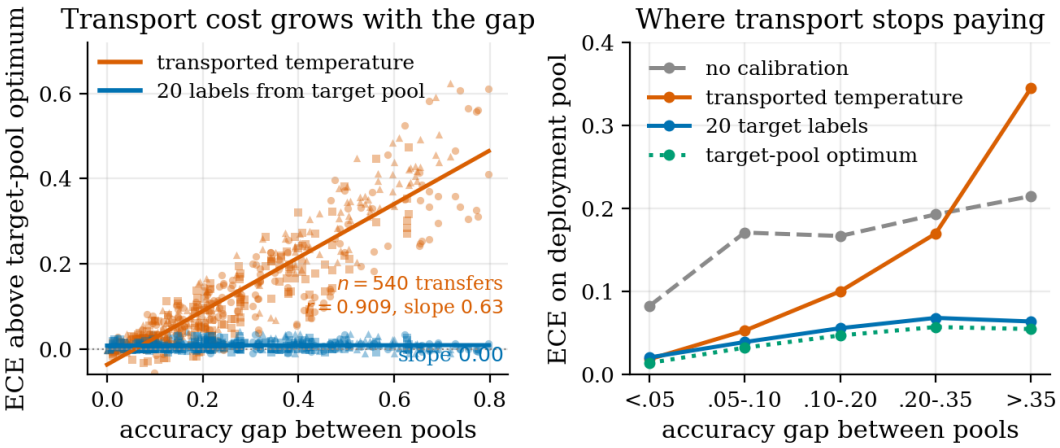


Figure 4: **Calibration transfer under difficulty shifts.** Excess ECE versus the accuracy gap for 540 ordered quartile transfers across 45 cells (left), and ECE by gap bin (right).

quantifies this across all 540 ordered transfers between difficulty quartiles of the 45 cells. The pooled correlation is $r = 0.909$, with slope 0.628 and a cluster-bootstrap 95% interval of $[0.890, 0.929]$. Recalibrating with 20 target-pool labels reduces the slope to approximately zero; see Appendix D.1. This dependence has a simple mechanistic core: temperature scaling preserves predictions and anchors mean confidence near the source-pool accuracy, so a transported fit inherits at least the base-rate gap, $\mathrm{ECE}_{\mathrm{tgt}} \geq |\mathrm{Acc}_{\mathrm{tgt}} - \bar{c}| \approx |\mathrm{Acc}_{\mathrm{tgt}} - \mathrm{Acc}_{\mathrm{src}}|$ (Zou et al., 2023). Transported confidence remains closer to source than target accuracy. Flatter logit distributions on harder pools permit partial self-adjustment, explaining why the slope is below one, but the principal failure remains the confidence level rather than the conditional ranking. Operationally, transported temperature is useful for small and moderate difficulty gaps but becomes worse than no calibration once the accuracy gap exceeds roughly 0.35. Reliability and selective-prediction thresholds are likewise pool-dependent. Appendices D.1, D.1, and D.3 provide the detailed decomposition and threshold analysis.

4.4.3 Beyond Scalar Temperature

To test whether greater calibrator expressiveness improves transfer, we compare scalar temperature with Platt scaling, isotonic regression, SMART, DAC, and ATS. Each method is fitted either on the source pool or on an independent target-pool calibration split with its own label budget, then evaluated on the same held-out harder-pool items across 45 model–task cells; see Table 5.

Under source-fit evaluation, DAC obtains 0.1919, compared with 0.1956 for scalar temperature, and a conservative source-validated multiplier obtains 0.1865. These gains are far smaller than the source–target gap. After target-pool refitting, scalar temperature, Platt, isotonic, SMART, and both DAC variants all reach 0.033–0.040, whereas ATS obtains 0.1181. ATS is the exception rather than a success: its target-fit ECE remains 0.1181, so its transfer gap of $+0.0723$ reflects a weak target fit rather than robust transfer. Its linear head is fitted on hidden states of dimension 1536–5120 from a mean of 421 calibration items, so the adaptation is under-determined; we report it as a protocol limitation rather than as evidence about the method. Fitting objectives also affect transfer: SMART with SoftECE reaches 0.2441, close to the uncalibrated 0.2413, and DAC with squared error reaches 0.2030 and no longer improves over scalar temperature. Expressiveness and alternative objectives do not close the source–target gap. Selective recalibration also retains transfer error: under an unlabeled target-quantile threshold, source-fit ECE rises from 0.2001 at full coverage to 0.2810 at 30% coverage, while target-fit ECE stays within 0.032–0.046; see Appendix D.3.

4.5 How Much Target Data Is Needed?

Figure 5 evaluates target-label budgets from 1 to 200 under strict held-out evaluation. A 20-label fit already beats both the transported temperature and the label-free reference, reaching a mean within-cell median ECE of 0.0394. Increasing the budget to 50 labels lowers the median to 0.0338 and the

Table 5: **Calibration methods under difficulty shift.** Mean ECE across 45 cells on a common held-out harder-pool split. Gap is source-fit minus target-fit ECE; fitting sizes and full protocols appear in Appendix D.3.

	Held-out ECE ↓		Gap
Method	**Source-fit**	**Target-fit**	ΔECE
Scalar temperature	0.1956	0.0325	+0.1631
Platt	0.2311	0.0325	+0.1986
Isotonic	0.2261	0.0350	+0.1911
SMART (adapted)	0.2084	0.0397	+0.1687
DAC (adapted)	0.1919	0.0364	+0.1555
Conservative temperature γT_{src}	0.1865	n/a	n/a
DAC without density (fitted scalar)	0.1998	0.0393	+0.1605
ATS (adapted)	0.1904	0.1181	+0.0723
No calibration	0.2413		n/a

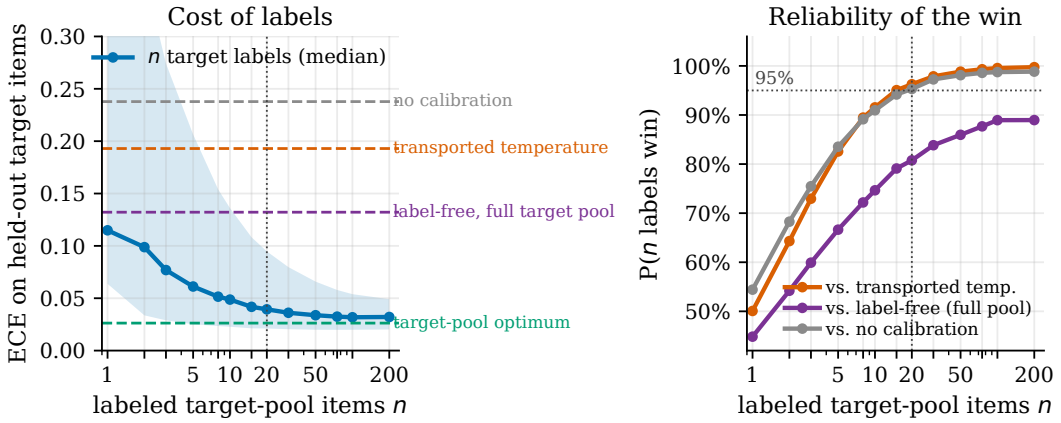


Figure 5: **Target-label budget and calibration quality.** Held-out ECE versus target-label budget with label-free and target-optimum references (left), and the win probability of an n-label fit (right).

P90 ECE from 0.0948 to 0.0659. Win probabilities rise steeply through roughly 50 labels and then begin to saturate. Very small budgets remain unstable; see Appendix D.1 for the sampling protocol and boundary cases. The median and upper tail require different budgets. Beyond 100 labels, median ECE is nearly flat at approximately 0.032, while P90 continues to improve from 0.0539 at 100 labels to 0.0491 at 200. Thus, 20–50 target labels provide a strong typical-case recalibration, whereas about 200 labels are needed to bring P90 below 0.05. Deployments that use probabilities for automated abstention or selective classification should therefore budget for the tail rather than the median. These recommendations assume that fitting labels are sampled from the deployment stream so that the scalar tracks its current difficulty mixture.

5 Discussion

Better predictive performance does not necessarily imply more reliable confidence. Our results show that post-training can improve accuracy while systematically degrading the alignment between confidence and correctness, introducing a reliability tax that can be substantially mitigated by post-hoc calibration. However, calibration is not fixed for a model: its parameters and effectiveness change across evaluation pools, making calibration a property of the model together with its target pool. This finding has a direct implication for calibration transfer. A calibrator fitted on one pool should not be assumed to remain effective on another, even when the model, task, and prompt template are fixed. At the same time, a small amount of labeled target-pool data can substantially reduce the resulting calibration error, providing a practical route for recalibration under pool shifts. Extending this analysis to naturally occurring topic, format, and temporal shifts, as well as free-form generation, remains an important direction for future work.

Reproducibility Statement

Section 4.1 and Appendices A.1 and A.2 specify the experimental setup, dataset splits, confidence definitions, calibration objectives, and resampling procedures. The appendices also provide per-cell results, sensitivity analyses, adaptive-calibrator protocols, and diagnostic measurements supporting the main conclusions.

Ethics Statement

This study evaluates publicly available benchmarks and pretrained model checkpoints and does not involve human subjects or private user data. Reliable confidence estimates may support abstention and human review, but calibration does not guarantee factuality, safety, or fairness. These methods should complement application-specific evaluation and human oversight, especially in high-stakes settings.

AI Use Statement

Generative AI tools were used to assist with language editing, organization, and LaTeX formatting of the manuscript. They were not used to generate experimental data or execute the reported analyses. The authors reviewed all AI-assisted text and take full responsibility for the content, results, and conclusions of the paper.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1343–1367, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.75. URL https://aclanthology.org/2023.acl-long.75/.

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

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1–2):12–22, 1983. doi: 10.2307/2987588.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, June 2024. doi: 10.1038/s41586-024-07421-0.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017. URL https://mlanthology.org/icml/2017/guo2017icml-calibration/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Haolan Guo, Linwei Tao, Haoyang Luo, Minjing Dong, and Chang Xu. Sample margin-aware recalibration of temperature scaling, 2026. URL https://arxiv.org/abs/2506.23492.

Guande He, Jianfei Chen, and Jun Zhu. Preserving pre-trained features helps calibrate fine-tuned language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NI7StoWHJPT.

Jerry Huang, Peng Lu, Qiuhao Zeng, Yusuke Iwasawa, Yutaka Matsuo, Sarath Chandar, Edison Marrese-Taylor, and Irene Li. Investigating the multilingual calibration effects of language model instruction tuning. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–59, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-381-4. doi: 10.18653/v1/2026.eacl-short.1. URL https://aclanthology.org/2026.eacl-short.1/.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe (eds.), *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 1–14, St Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.uncertainlp-1.1. URL https://aclanthology.org/2024.uncertainlp-1.1/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/kuhn2023iclr-semantic/.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8ca01ea920679a0fe3728441494041b9-Paper.pdf.

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

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 16484–16517, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/29fb6e1456b3d8b57ede5c45aa2c6537-Paper-Conference.pdf.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-Box Large Language Models. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/lin2024tmlr-generating/.

Mingyi Liu. The alignment tax: Response homogenization in aligned LLMs and its implications for uncertainty estimation, 2026. URL https://arxiv.org/abs/2603.24124.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained LLM is secretly an unsupervised confidence calibrator. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=I4PJYZvfW5.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In T. Walsh, J. Shah, and Z. Kolter (eds.), *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19260–19268. Association for the Advancement of Artificial Intelligence, 2025. doi: 10.1609/aaai.v39i18.34120. URL https://doi.org/10.1609/aaai.v39i18.34120.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluis Marquez. Factual confidence of LLMs: On reliability and robustness of current estimators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4554–4570, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.250. URL https://aclanthology.org/2024.acl-long.250/.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, June 1973. ISSN 1520-0450. doi: 10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8901–8929. Curran Associates, Inc., 2024. doi: 10.52202/079017-0283. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10c456d2160517581a234dfde15a7505-Paper-Conference.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates, Inc.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. Thermometer: Towards Universal Calibration for Large Language Models. In *International Conference on Machine Learning*, volume 235, pp. 44687–44711, 2024. URL https://mlanthology.org/icml/2024/shen2024icml-thermometer/.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, i'm wrong: LLMs hallucinate with certainty despite knowing the answer. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14665–14688, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.792. URL https://aclanthology.org/2025.findings-emnlp.792/.

D. B. Stephenson, C. Coelho, and I. T. Jolliffe. Two extra components in the brier score decomposition, 2007. URL https://www.ecmwf.int/node/15460.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.238. URL https://aclanthology.org/2025.emnlp-main.238/.

Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun, and Xueqi Cheng. BaseCal: Unsupervised confidence calibration via base model signals. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5172–5187, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.234. URL https://aclanthology.org/2026.acl-long.234/.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. When to trust LLMs: Aligning confidence with response quality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5984–5996, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.357. URL https://aclanthology.org/2024.findings-acl.357/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Christian Tomani, Futa Kai Waseda, Yuesong Shen, and Daniel Cremers. Beyond in-domain scenarios: Robust density-aware calibration. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34344–34368. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/tomani23a.html.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=J5sUOvlLbQ.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Ziming Wang, Zeyu Shi, Haoyi Zhou, Shiqi Gao, Qingyun Sun, and Jianxin Li. Towards objective fine-tuning: How LLMs' prior knowledge causes potential poor calibration? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14830–14853, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.722. URL https://aclanthology.org/2025.acl-long.722/.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. Restoring Calibration for Aligned Large Language Models: A Calibration-Aware Fine-Tuning Approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 68364–68390, 2025. URL https://mlanthology.org/icml/2025/xiao2025icml-restoring/.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1007. URL https://aclanthology.org/2024.emnlp-main.1007/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.173. URL https://aclanthology.org/2024.emnlp-main.173/.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.654. URL https://aclanthology.org/2023.findings-emnlp.654/.

Yuli Zou, Weijian Deng, and Liang Zheng. Adaptive calibrator ensemble: Navigating test set difficulty in out-of-distribution scenarios. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19276–19285, 2023. doi: 10.1109/ICCV51070.2023.01771.

CONTENTS OF APPENDIX

A Evaluation Protocol and Robustness Checks **16**

A.1 Audits of Weights, Prompts, Fitting, and Metrics 16

A.2 Models, Tasks, and Evaluation Protocol 17

B Full Results and Statistical Analyses **17**

B.1 Cross-Task Calibration Survey 17

B.2 Per-Cell Results by Post-Training Recipe 17

B.3 Multiplicity-Adjusted Comparisons Across Cells 18

B.4 Reliability and Resolution under Alternative Aggregation and Binning 18

B.5 Per-Cell Equivalence Margins 24

B.6 Random-Effects Pooling of AUROC Changes 24

C Alternative Confidence Signals and Recalibration Baselines **24**

C.1 Self-Evaluation as a Confidence Signal 25

C.2 Paired Recalibrator Comparisons and Independent-Split Validation 25

C.3 Cross-Task Calibration Transfer 27

D Calibration under Evaluation-Pool Shift **28**

D.1 Pool-Shift Results, Label Budgets, and Label-Free Proxies 28

D.2 Controls for Item Composition and Base-Model Geometry 31

D.3 Expressive Recalibrators under Pool Shift 33

D.4 Confidence-Stratified Analysis 38

E Weight-Space Diagnostics of the Reliability Tax **38**

E.1 Counterfactual Layer Restoration 38

E.2 Held-Out Comparison of Layer Restoration and Temperature Scaling 38

E.3 Validation of the Layer-Sensitivity Ranking 39

E.4 Layer-Scaling Baselines 41

E.5 Scalar and Linear Approximations of Layer Restoration 42

E.6 Summary of Diagnostic Measurements 44

A Evaluation Protocol and Robustness Checks

Seven protocol audits evaluate the model weights, prompts, fitting rules, and metrics used to compare scalar recalibration with weight-space interventions. The final configurations and their sensitivity results are reported below.

A.1 Audits of Weights, Prompts, Fitting, and Metrics

Temperature on the original tuned weights. Temperature scaling is evaluated directly on the original tuned weights. Its ECE is 0.0403 with ECE-grid fitting and 0.0169 with NLL fitting. For comparison, the uncalibrated tuned model and the 21-layer rollback obtain ECE values of 0.0943 and 0.0393, respectively. Positive-temperature scaling preserves the tuned model's predictions and accuracy; Table 37 compares temperature and rollback as separate interventions.

Shared prompting. Temperature and weight-space interventions use the same 3-shot completion builder. On Qwen-1.5B ARC, held-out temperature ECE is 0.035 with ECE-grid fitting and 0.027 with NLL fitting, compared with 0.043 under a 0-shot variant with an additional header line. The selected restoration point obtains ECE 0.025 under the shared 3-shot protocol.

Option-mass verification. The option-restricted softmax conditions confidence on emitting a valid option token. Table 6 reports the mean full-vocabulary probability mass on legal option tokens for the OLMo-2-1B recipe ladder. Mass is uniformly at least 0.98 and slightly higher after post-training. Restricted renormalization therefore inflates base confidence slightly more: on MMLU-Pro, the factor is 1.0198 for the base model and 1.0147 after post-training. Recomputing $\Delta\mathrm{REL}$ with unnormalized full-vocabulary probabilities changes any per-cell value by at most 0.0027 for MMLU-Pro under DPO.

Table 6: **Mean full-vocabulary probability mass on legal option tokens**, OLMo-2-1B recipe ladder, option-restricted scoring protocol.

Task	Base	SFT	DPO	RLVR
ARC	0.9986	0.9997	1.0000	0.9999
MMLU-Pro	0.9806	0.9836	0.9855	0.9854
MMLU-math	0.9950	0.9984	0.9987	0.9983

Truncated prompts. The 27 MMLU-Pro items that exceed the 1,600-token prompt limit account for 1.35% of the pool. Their lengths range from 1,601 to 1,859 tokens, with a median of 1,684 versus 557 pool-wide. The prompt is therefore truncated before the answer position and the scored distribution is not conditional on a completed `Answer:` cue; their mean option mass is below 0.5. The same 27 items are affected at all four checkpoints, so the affected set is shared across base and tuned models. Removing these items changes $\Delta\mathrm{REL}$ by at most 0.0013.

Representation-fit sample size. The representation map is fitted on the last 32 prompt positions, yielding $18{,}752$ training rows. At 1.5B, its off-diagonal energy fraction is 0.053 and its relative distance to $c\,I$ is 0.26. A decision-position-only ablation provides 586 rows in 1536 dimensions and yields corresponding values of 0.88 and 2.74, indicating strong ridge-dominated behavior in that low-sample regime. All reported representation-map results use the 32-position configuration.

Cross-task temperature fitting. All four cross-task directions in Appendix C.3 use the negative log-likelihood objective declared in Appendix A.2. On Qwen-1.5B MMLU-Pro, NLL fitting returns $T{=}2.89$, whereas an ECE grid over $[0.5, 6]$ returns $T{=}2.20$; their transferred ARC ECE values are 0.104 and 0.041, respectively. Under the common NLL objective, the full linear map has the lowest ECE in this direction.

ECE endpoints. The final ECE implementation assigns confidence exactly 1.0 to the last bin and includes every item. A boundary convention that drops confidence exactly 1.0 underestimates ECE

when bin weights are normalized over the full pool; renormalizing only the retained items can shift the estimate in either direction.

The reported baselines and interventions use identical inputs and a shared evaluation function. Positive-temperature scaling of fixed logits preserves predictions and accuracy. NLL is convex in inverse temperature for fixed logits, although its optimum need not minimize held-out ECE.

A.2 Models, Tasks, and Evaluation Protocol

We evaluate architecture-matched base/tuned pairs from Qwen2.5 at 0.5B–32B, Llama-3.1-8B, OLMo-2-1B and Tulu-3-8B after SFT, DPO and the final recipe stage, Qwen2.5-Math, and R1-distilled checkpoints (Guo et al., 2025). The decomposition and recalibration analyses use 47 model–task cells from sixteen model pairs with per-item records from both checkpoints; the self-evaluation comparison of Appendix C.1 covers all of them. The broader survey covers 13 tuned checkpoints across five tasks; Table 7 in Appendix B.1 reports representative rows.

Tasks are MMLU-Pro (Wang et al., 2024), MMLU-math, ARC-Challenge, TruthfulQA-mc2 (Lin et al., 2022a), and GSM8K (Cobbe et al., 2021). MMLU-math comprises the 1,351 items of MMLU-Pro's mathematics category; the 2,000-item MMLU-Pro evaluation sample shares 230 of them, while ARC-Challenge is disjoint from both. The answer-letter multiple-choice experiments use 3-shot plain completion, with confidence computed by softmax restricted to legal option logits. Option-softmax confidence has higher AUROC than Kadavath-style self-evaluation (Kadavath et al., 2022) in all 47 cells, averaging 0.724 versus 0.588; see Appendix C.1. Within each comparison, checkpoints, interventions, and baselines use the same inputs. TruthfulQA and GSM8K use task-specific sequence scoring, so their ECE comparisons are within task and protocol. All analyses in Section 4.2.2 onward use the answer-letter multiple-choice setting.

Unless stated otherwise, each pool is stored in its seed-0 shuffled sampling order and split evenly into a fitting half D_A and a reporting half D_B. The split sizes are $586/586$ for ARC, $1000/1000$ for MMLU-Pro, and $675/676$ for MMLU-math. Recalibrators are fitted on D_A and reported on D_B. Temperature is fitted by minimizing negative log-likelihood, a convex objective in $\beta{=}1/T$ for $\beta \in (0, \infty)$. Appendix A.1 reports the protocol audits. For base-to-tuned comparisons, paired bootstrap uses 2000–4000 resamples with seed 2026, resampling item indices jointly for base and tuned. The separate adaptive-calibrator suite bootstraps paired cell summaries instead; see Appendix D.3. The 230 items shared between the two MMLU-derived pools make their cell summaries correlated beyond what cell-level clustering removes.

Accuracy is reported as a percentage unless explicitly labeled as a fraction, and accuracy changes in percentage points, denoted pp. ECE, reliability, resolution, and AUROC retain their original scales.

B Full Results and Statistical Analyses

The complete numbers behind Section 4.2.1 and Section 4.2.2, together with multiplicity control, an exploratory equivalence-margin analysis, and random-effects pooling.

B.1 Cross-Task Calibration Survey

Table 7 gives the cross-task calibration overview referenced in Section 4.2.1. It lists representative rows of the 13×5 survey, including TruthfulQA and GSM8K, whose scoring rules differ from the main multiple-choice analysis. Figure 6 shows the OLMo recipe ladder.

B.2 Per-Cell Results by Post-Training Recipe

Table 8 lists all 47 model–task cells with confidence intervals behind the per-pair matrix of Table 1. Table 9 gives the per-cell values summarized by Table 3. Table 10 supplies the 20 instruction-tuning cells aggregated in Table 2, retaining accuracy and risk–coverage changes alongside reliability, resolution, and AUROC.

Table 7: **Representative rows from the calibration matrix** under a unified prompting protocol. ECE values are comparable within a task column, not across tasks with different scoring rules.

Model	**MMLU-Pro**, 2000		**MMLU-math**, 1351		**ARC**, 1172		**TQA-mc2**, 817		**GSM8K**, 1319	
	ECE↓	Acc. [%]↑	ECE↓	Acc. [%]↑	ECE↓	Acc. [%]↑	ECE↓	Acc. [%]↑	ECE↓	Acc. [%]↑
Qwen-1.5B-Inst	0.194	29.4	0.219	21.5	0.115	76.4	0.427	54.1	0.656	7.4
Qwen-7B-Inst	0.298	46.3	0.380	35.8	0.079	89.9	0.414	55.1	0.481	41.5
Math-1.5B	0.081	21.4	0.078	20.4	0.036	56.6	0.426	40.0	0.616	19.5
Math-7B	0.089	32.2	0.061	32.0	0.068	79.5	0.435	45.7	0.668	11.8
R1-1.5B	0.154	16.6	0.162	12.7	0.056	45.3	0.415	59.7	0.656	7.6
R1-7B	0.183	28.2	0.169	23.3	0.033	72.5	0.412	63.4	0.595	16.1
OLMo-SFT	0.116	15.6	0.062	11.6	0.144	50.8	0.433	30.4	0.468	30.6
OLMo-DPO	0.221	15.5	0.155	11.4	0.262	46.9	0.418	43.8	0.515	32.3
OLMo-Inst	0.257	16.6	0.239	10.7	0.275	47.3	0.421	45.9	0.427	45.5

Additional 0.5B/3B/14B/Llama rows are omitted; all 13 TQA ECE values lie in 0.41–0.44.

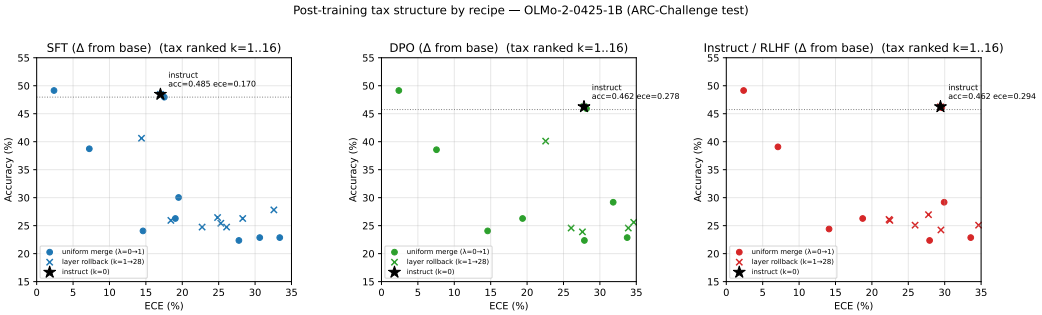


Figure 6: OLMo-2-1B on ARC across post-training stages. ECE rises from SFT to DPO to Instruct, while accuracy does not increase monotonically. The matched base-to-tuned changes for each checkpoint are reported in Table 8.

B.3 Multiplicity-Adjusted Comparisons Across Cells

Section 4.2.2 reports three main-group cells whose 95% intervals for ΔAUROC exclude zero. Because the same test is applied to all 35 main-group cells, three uncorrected rejections at $\alpha = 0.05$ compares with 1.75 expected under the null. Table 11 gives two-sided bootstrap p-values with Bonferroni and Benjamini–Hochberg adjustment, and two alternative discrimination measures for each cell.

No cell is significant after either adjustment. Qwen-14B MMLU-math loses 0.0336 AUROC, with Bonferroni $p = 0.490$ and BH $q = 0.408$. OLMo-1B-DPO MMLU-Pro gains 0.0492, with $p = 1.000$ and $q = 0.408$, and Tulu3-SFT MMLU-Pro gains 0.0281, with $p = 0.840$ and $q = 0.408$. The signs also vary across tasks: Qwen-14B and OLMo-1B-DPO gain 0.0168 and 0.0128 on ARC. These point estimates do not support a consistent checkpoint-level discrimination change.

B.4 Reliability and Resolution under Alternative Aggregation and Binning

Empirical binning protocol. The reported Murphy components use K=10 rank-based equal-frequency bins, distinct from ECE's ten equal-width bins. In `calib_common.murphy`, `np.argsort(conf, kind="stable")` sorts confidence in ascending order, retaining the original input-array order for ties; `np.array_split(order, k)` then divides the ordered indices into ten consecutive groups whose sizes differ by at most one. Within each bin, REL uses squared mean confidence–accuracy mismatch and RES uses squared deviation of bin accuracy from overall accuracy, both weighted by the bin's item fraction. Base and tuned checkpoints are processed in separate calls in `exp_all_cells_v2.py`: they share neither a partition nor bin boundaries, even when evaluated on the same items. Equal-confidence clusters can therefore be split across rank bins.

Table 8: **All 47 model–task cells by post-training recipe.** The upper block contains instruction/SFT/DPO/RLVR and the lower block math-RL/distillation. The interval column reports the 95% interval for ΔAUROC.

Pair	Task	Recipe	ΔAcc. [pp]	ΔREL	ΔAUROC	95% CI
Instruction/SFT/DPO/RLVR, main group, 35 cells						
Llama-8B	ARC	instruct	+3.8	-0.0006	-0.0079	[-0.035,+0.020]
Llama-8B	MMLU-Pro	instruct	+3.3	+0.0019	+0.0243	[-0.000,+0.050]
Llama-8B	MMLU-math	instruct	+4.0	+0.0050	+0.0162	[-0.032,+0.067]
OLMo-1B-Inst	ARC	instruct	-2.0	+0.0859	-0.0051	[-0.042,+0.033]
OLMo-1B-Inst	MMLU-math	instruct	+0.9	+0.0564	+0.0299	[-0.045,+0.102]
Qwen-1.5B	ARC	instruct	+0.2	+0.0174	+0.0099	[-0.010,+0.031]
Qwen-1.5B	MMLU-Pro	instruct	+1.7	+0.0411	-0.0194	[-0.043,+0.004]
Qwen-1.5B	MMLU-math	instruct	-0.1	+0.0554	+0.0204	[-0.015,+0.058]
Qwen-14B	ARC	instruct	+0.5	+0.0037	+0.0168	[-0.007,+0.040]
Qwen-14B	MMLU-Pro	instruct	+2.7	+0.0695	-0.0108	[-0.031,+0.008]
Qwen-14B	MMLU-math	instruct	+4.2	+0.0869	-0.0336	[-0.061,-0.005]
Qwen-32B	ARC	instruct	+0.7	+0.0022	-0.0068	[-0.025,+0.012]
Qwen-32B	MMLU-Pro	instruct	+1.5	+0.0384	+0.0095	[-0.007,+0.027]
Qwen-32B	MMLU-math	instruct	+2.4	+0.0341	+0.0190	[-0.006,+0.044]
Qwen-3B	ARC	instruct	+1.0	+0.0267	+0.0029	[-0.017,+0.024]
Qwen-3B	MMLU-Pro	instruct	+3.8	+0.1139	-0.0186	[-0.043,+0.007]
Qwen-3B	MMLU-math	instruct	+1.8	+0.1584	+0.0034	[-0.035,+0.040]
Qwen-7B	ARC	instruct	-0.2	+0.0192	+0.0043	[-0.015,+0.023]
Qwen-7B	MMLU-Pro	instruct	+0.2	+0.1027	-0.0073	[-0.027,+0.013]
Qwen-7B	MMLU-math	instruct	+2.9	+0.1533	-0.0253	[-0.056,+0.004]
OLMo-1B-SFT	ARC	sft	+1.5	+0.0244	+0.0089	[-0.028,+0.047]
OLMo-1B-SFT	MMLU-Pro	sft	+1.5	+0.0128	+0.0203	[-0.024,+0.067]
OLMo-1B-SFT	MMLU-math	sft	+1.8	-0.0054	+0.0369	[-0.036,+0.102]
Tulu3-SFT	ARC	sft	-1.6	+0.0002	-0.0052	[-0.029,+0.019]
Tulu3-SFT	MMLU-Pro	sft	-1.4	+0.0225	+0.0281	[+0.005,+0.053]
Tulu3-SFT	MMLU-math	sft	+1.1	+0.0137	-0.0115	[-0.059,+0.038]
OLMo-1B-DPO	ARC	dpo	-2.3	+0.0763	+0.0128	[-0.027,+0.052]
OLMo-1B-DPO	MMLU-Pro	dpo	+1.4	+0.0575	+0.0492	[+0.004,+0.096]
OLMo-1B-DPO	MMLU-math	dpo	+1.6	+0.0168	+0.0465	[-0.028,+0.116]
Tulu3-DPO	ARC	dpo	-1.2	+0.0209	-0.0156	[-0.041,+0.011]
Tulu3-DPO	MMLU-Pro	dpo	-0.3	+0.0791	+0.0008	[-0.025,+0.029]
Tulu3-DPO	MMLU-math	dpo	+1.5	+0.0663	-0.0299	[-0.077,+0.017]
Tulu3-RLVR	ARC	rlvr	-2.0	+0.0277	-0.0013	[-0.026,+0.025]
Tulu3-RLVR	MMLU-Pro	rlvr	+0.1	+0.0844	-0.0051	[-0.032,+0.024]
Tulu3-RLVR	MMLU-math	rlvr	+1.9	+0.0776	-0.0395	[-0.087,+0.009]
Math-RL/distillation, comparison group, 12 cells						
Math-1.5B	ARC	math-rl	-19.6	+0.0000	-0.0974	[-0.133,-0.061]
Math-1.5B	MMLU-Pro	math-rl	-6.3	+0.0046	-0.0823	[-0.120,-0.047]
Math-1.5B	MMLU-math	math-rl	-1.2	+0.0016	+0.0349	[-0.012,+0.082]
Math-7B	ARC	math-rl	-10.2	+0.0044	-0.0654	[-0.091,-0.039]
Math-7B	MMLU-Pro	math-rl	-14.8	+0.0067	-0.0340	[-0.062,-0.006]
Math-7B	MMLU-math	math-rl	-1.0	-0.0042	-0.0235	[-0.059,+0.011]
R1-1.5B	ARC	distill	-11.3	+0.0012	-0.0474	[-0.085,-0.010]
R1-1.5B	MMLU-Pro	distill	-4.7	+0.0193	-0.0082	[-0.050,+0.034]
R1-1.5B	MMLU-math	distill	-7.7	+0.0202	-0.0700	[-0.123,-0.014]
R1-7B	ARC	distill	-6.9	-0.0025	-0.0671	[-0.096,-0.037]
R1-7B	MMLU-Pro	distill	-3.1	+0.0285	-0.0445	[-0.077,-0.012]
R1-7B	MMLU-math	distill	-8.4	+0.0255	-0.0536	[-0.096,-0.010]

Table 9: **Seven confidence sources on D_B across 47 cells.** Recalibrators are fitted on D_A and preserve tuned predictions; the base checkpoint is a separate predictor. DACA is label-free. Bold marks the lowest row-wise ECE.

		uncalibrated		**supervised, fitted on D_A**				**unsup.**
Pair	**Recipe**	base	tuned	temp, NLL	temp, ECE	Platt	isotonic	DACA
ARC								
Llama-8B	instruct	0.0443	**0.0376**	0.0437	0.0457	0.0416	0.0543	0.0978
OLMo-1B-Inst	instruct	**0.0258**	0.2955	0.0331	0.0418	0.0385	0.0392	0.0382
Qwen-1.5B	instruct	0.0455	0.1157	**0.0269**	0.0350	0.0286	0.0459	0.0291
Qwen-14B	instruct	0.0199	0.0284	0.0200	0.0256	0.0199	0.0214	**0.0193**
Qwen-32B	instruct	**0.0154**	0.0231	0.0187	0.0258	0.0169	0.0196	0.0190
Qwen-3B	instruct	0.0275	0.1230	0.0336	0.0417	**0.0222**	0.0242	0.0344
Qwen-7B	instruct	0.0180	0.0780	0.0164	0.0283	**0.0136**	0.0302	0.0167
OLMo-1B-SFT	sft	**0.0258**	0.1726	0.0359	0.0265	0.0609	0.0750	0.0269
Tulu3-SFT	sft	0.0443	0.0363	0.0245	0.0363	**0.0143**	0.0383	0.0399
OLMo-1B-DPO	dpo	**0.0258**	0.2785	0.0318	0.0307	0.0385	0.0372	0.0377
Tulu3-DPO	dpo	0.0443	0.1272	0.0331	**0.0193**	0.0275	0.0382	0.0665
Tulu3-RLVR	rlvr	0.0443	0.1402	0.0330	0.0399	**0.0166**	0.0280	0.0494
Math-1.5B	math-rl	0.0455	**0.0340**	0.0449	0.0448	0.0475	0.0621	0.2177
Math-7B	math-rl	**0.0180**	0.0832	0.0328	0.0347	0.0418	0.0641	0.1019
R1-1.5B	distill	0.0340	0.0483	0.0304	0.0388	**0.0255**	0.0529	0.1375
R1-7B	distill	0.0832	0.0420	**0.0406**	0.0540	0.0459	0.0624	0.0472
MMLU-Pro								
Llama-8B	instruct	0.0723	0.0742	0.0560	0.0528	0.0519	0.0548	**0.0443**
Qwen-1.5B	instruct	0.0636	0.1812	0.0409	0.0289	**0.0240**	0.0293	0.0381
Qwen-14B	instruct	**0.0287**	0.2335	0.0395	0.0485	0.0353	0.0571	0.0339
Qwen-32B	instruct	0.0446	0.1771	0.0329	0.0323	**0.0154**	0.0357	0.0434
Qwen-3B	instruct	0.0588	0.3275	0.0476	0.0488	0.0489	**0.0448**	0.0479
Qwen-7B	instruct	0.0441	0.3063	**0.0211**	0.0501	0.0320	0.0425	0.0517
OLMo-1B-SFT	sft	0.0838	0.1202	**0.0029**	0.0106	0.0115	0.0210	0.0762
Tulu3-SFT	sft	0.0723	0.1480	0.0271	0.0294	**0.0197**	0.0286	0.0823
OLMo-1B-DPO	dpo	0.0838	0.2206	0.0231	0.0230	**0.0176**	0.0243	0.0755
Tulu3-DPO	dpo	0.0723	0.2618	0.0393	0.0433	**0.0294**	0.0392	0.0801
Tulu3-RLVR	rlvr	0.0723	0.2708	0.0356	0.0428	**0.0326**	0.0431	0.0773
Math-1.5B	math-rl	0.0636	0.0950	**0.0251**	0.0365	0.0312	0.0328	0.1379
Math-7B	math-rl	0.0441	0.0927	0.0413	0.0414	0.0403	**0.0283**	0.1826
R1-1.5B	distill	0.0950	0.1672	**0.0066**	0.0200	0.0318	0.0458	0.1458
R1-7B	distill	0.0927	0.1890	0.0330	**0.0308**	0.0387	0.0539	0.1423
MMLU-math								
Llama-8B	instruct	0.0429	0.0661	0.0124	**0.0123**	0.0277	0.0441	0.0133
OLMo-1B-Inst	instruct	0.1094	0.2257	**0.0071**	0.0356	0.0323	0.0453	0.0749
Qwen-1.5B	instruct	0.0457	0.2101	0.0392	**0.0227**	0.0404	0.0398	0.0351
Qwen-14B	instruct	0.0482	0.2643	**0.0374**	0.0397	0.0395	0.0629	0.0397
Qwen-32B	instruct	0.0624	0.1941	**0.0304**	0.0476	0.0538	0.0629	0.1014
Qwen-3B	instruct	0.0398	0.3762	0.0261	**0.0199**	0.0298	0.0345	0.0225
Qwen-7B	instruct	0.1221	0.4146	0.0683	0.0668	**0.0607**	0.0716	0.0975
OLMo-1B-SFT	sft	0.1094	0.0486	0.0215	**0.0083**	0.0291	0.0281	0.0677
Tulu3-SFT	sft	0.0429	0.0783	0.0282	0.0205	0.0196	**0.0195**	0.0283
OLMo-1B-DPO	dpo	0.1094	0.1477	**0.0123**	0.0250	0.0231	0.0197	0.0725
Tulu3-DPO	dpo	0.0429	0.1881	0.0186	0.0249	0.0128	**0.0127**	0.0392
Tulu3-RLVR	rlvr	0.0429	0.2033	0.0298	0.0304	0.0247	**0.0135**	0.0462
Math-1.5B	math-rl	0.0457	0.0944	**0.0261**	0.0465	0.0373	0.0345	0.0672
Math-7B	math-rl	0.1221	0.0460	0.0357	**0.0259**	0.0517	0.0528	0.0848
R1-1.5B	distill	0.0944	0.1632	**0.0047**	0.0193	0.0264	0.0297	0.1577
R1-7B	distill	0.0460	0.1798	0.0456	0.0336	0.0314	**0.0294**	0.1608

Table 10: **Murphy decomposition for seven instruction-tuned pairs.** Positive $\Delta\mathrm{REL}$ is worse; $\Delta(\mathrm{RES}/\mathrm{UNC})$ and $\Delta\mathrm{AUROC}$ measure discrimination. AURC is lower-is-better. Other recipes appear in Table 8.

Pair	Task	ΔREL	Δ(RES/UNC)	ΔAUROC	ΔAURC	ΔAcc. [pp]
Llama-8B	ARC	-0.0006	-0.0380	-0.0079	-0.0120	+3.8
OLMo-1B-Inst	ARC	+0.0859	+0.0050	-0.0051	+0.0219	-2.0
Qwen-1.5B	ARC	+0.0174	+0.0176	+0.0099	-0.0032	+0.2
Qwen-14B	ARC	+0.0037	+0.0081	+0.0168	-0.0023	+0.5
Qwen-32B	ARC	+0.0022	-0.0286	-0.0068	-0.0006	+0.7
Qwen-3B	ARC	+0.0267	+0.0080	+0.0029	-0.0033	+1.0
Qwen-7B	ARC	+0.0192	+0.0366	+0.0043	+0.0002	-0.2
Llama-8B	MMLU-Pro	+0.0019	+0.0156	+0.0243	-0.0513	+3.3
Qwen-1.5B	MMLU-Pro	+0.0411	-0.0187	-0.0194	-0.0088	+1.7
Qwen-14B	MMLU-Pro	+0.0695	-0.0199	-0.0108	-0.0157	+2.7
Qwen-32B	MMLU-Pro	+0.0384	+0.0135	+0.0095	-0.0132	+1.5
Qwen-3B	MMLU-Pro	+0.1139	-0.0202	-0.0186	-0.0251	+3.8
Qwen-7B	MMLU-Pro	+0.1027	-0.0070	-0.0073	+0.0037	+0.2
Llama-8B	MMLU-math	+0.0050	+0.0214	+0.0162	-0.0639	+4.0
OLMo-1B-Inst	MMLU-math	+0.0564	+0.0014	+0.0299	-0.0257	+0.9
Qwen-1.5B	MMLU-math	+0.0554	+0.0038	+0.0204	-0.0110	-0.1
Qwen-14B	MMLU-math	+0.0869	-0.0398	-0.0336	-0.0251	+4.2
Qwen-32B	MMLU-math	+0.0341	+0.0208	+0.0190	-0.0298	+2.4
Qwen-3B	MMLU-math	+0.1584	+0.0085	+0.0034	-0.0183	+1.8
Qwen-7B	MMLU-math	+0.1533	-0.0293	-0.0253	-0.0201	+2.9

Aggregation rule. For a group of cells R, define the reliability share as

$$\mathrm{REL\ share}(R) = \frac{\frac{1}{|R|}\sum_{r\in R}|\Delta\mathrm{REL}_r|}{\frac{1}{|R|}\sum_{r\in R}|\Delta\mathrm{REL}_r| + \frac{1}{|R|}\sum_{r\in R}|\Delta\mathrm{RES}_r|}. \tag{5}$$

Absolute changes are averaged within the group before taking the ratio. This differs both from averaging per-cell ratios and from taking a ratio of absolute group-mean changes. Here $\Delta\mathrm{RES}_r$ is the difference in unnormalized resolution, not the change in RES/UNC. The statistic compares the magnitudes of two Brier components only; it excludes uncertainty and the within-bin correction and is not a share of ECE or total Brier-score change.

Sensitivity to binning rule and bin count. Table 13 recomputes the decomposition for all 47 cells under ten configurations, using the same definitions as the main analysis. The main group contains 35 cells and the control group 12. The main-group reliability share stays within 93.4–94.1% for equal-frequency bins, but falls to 86.4–91.5% for equal-width bins. Thus the predominance of reliability changes persists, while the precise 93.4% attribution depends on the partition. Equal-width binning also yields more negative mean ΔRES in the main group. The control-group share ranges from 45.8% to 47.9% across the ten configurations. These are descriptive sensitivity results, not confidence intervals.

Unbinned score references. For the same cells, mean changes in item-level top-label Brier score and NLL are $+0.0513$ and $+0.3095$ in the main group, versus $+0.0148$ and $+0.2046$ in the control group. These unbinned scores measure total score degradation, complementing the component-wise attribution above.

Confidence ties and tie-aware re-binning. Confidence ties are common on ARC for stronger checkpoints: 66.5% of Qwen-14B's items share a confidence value with another item, as do 63.3% for Qwen-7B and 56.2% for Qwen-32B. Qwen-14B's largest tie cluster contains 563 of 1172 items. Across the 68 inspected files, the mean tied-item fraction is 5.1% and the median tie-cluster size is two; this file-level summary is separate from the 47 paired cells. Stable sorting allocates tied items according to input order. Splitting a large cluster can create bins with identical confidence but

Table 11: **Multiplicity adjustment over the 35 main-group cells.** ΔAUROC is the plug-in difference, matching Table 8; two-sided p-values come from 4000 paired-bootstrap resamples, with Bonferroni and Benjamini–Hochberg adjustment, alongside two alternative discrimination measures. No cell is significant after either adjustment, and three cells are uncorrected rejections.

Pair	**Task**	ΔAUROC	p	p_{Bonf}	q_{BH}	ΔAURC	$\Delta\rho_s$
Llama-8B	ARC	-0.0079	0.590	1.000	0.816	-0.0120	-0.0475
OLMo-1B-Inst	ARC	-0.0051	0.823	1.000	0.900	$+0.0219$	-0.0092
Qwen-1.5B	ARC	$+0.0099$	0.330	1.000	0.641	-0.0032	$+0.0134$
Qwen-14B	ARC	$+0.0168$	0.151	1.000	0.528	-0.0023	$+0.0195$
Qwen-32B	ARC	-0.0068	0.474	1.000	0.742	-0.0006	-0.0262
Qwen-3B	ARC	$+0.0029$	0.777	1.000	0.878	-0.0033	-0.0074
Qwen-7B	ARC	$+0.0043$	0.645	1.000	0.816	$+0.0002$	$+0.0200$
OLMo-1B-SFT	ARC	$+0.0089$	0.614	1.000	0.816	-0.0170	$+0.0154$
Tulu3-SFT	ARC	-0.0052	0.676	1.000	0.816	$+0.0126$	$+0.0066$
OLMo-1B-DPO	ARC	$+0.0128$	0.508	1.000	0.742	$+0.0158$	$+0.0216$
Tulu3-DPO	ARC	-0.0156	0.251	1.000	0.634	$+0.0121$	-0.0116
Tulu3-RLVR	ARC	-0.0013	0.921	1.000	0.948	$+0.0120$	$+0.0150$
Llama-8B	MMLU-Pro	$+0.0243$	0.059	1.000	0.521	-0.0513	$+0.0494$
Qwen-1.5B	MMLU-Pro	-0.0194	0.101	1.000	0.528	-0.0088	-0.0245
Qwen-14B	MMLU-Pro	-0.0108	0.285	1.000	0.637	-0.0157	-0.0202
Qwen-32B	MMLU-Pro	$+0.0095$	0.309	1.000	0.637	-0.0132	$+0.0135$
Qwen-3B	MMLU-Pro	-0.0186	0.149	1.000	0.528	-0.0251	-0.0210
Qwen-7B	MMLU-Pro	-0.0073	0.479	1.000	0.742	$+0.0037$	-0.0124
OLMo-1B-SFT	MMLU-Pro	$+0.0203$	0.377	1.000	0.694	-0.0332	$+0.0314$
Tulu3-SFT	MMLU-Pro	$+0.0281$	0.024	0.840	0.408	$+0.0058$	$+0.0407$
OLMo-1B-DPO	MMLU-Pro	$+0.0492$	0.035	1.000	0.408	-0.0425	$+0.0671$
Tulu3-DPO	MMLU-Pro	$+0.0008$	0.950	1.000	0.950	$+0.0080$	$+0.0006$
Tulu3-RLVR	MMLU-Pro	-0.0051	0.714	1.000	0.833	$+0.0083$	-0.0082
Llama-8B	MMLU-math	$+0.0162$	0.509	1.000	0.742	-0.0639	$+0.0377$
OLMo-1B-Inst	MMLU-math	$+0.0299$	0.416	1.000	0.729	-0.0257	$+0.0322$
Qwen-1.5B	MMLU-math	$+0.0204$	0.254	1.000	0.634	-0.0110	$+0.0286$
Qwen-14B	MMLU-math	-0.0336	0.014	0.490	0.408	-0.0251	-0.0529
Qwen-32B	MMLU-math	$+0.0190$	0.144	1.000	0.528	-0.0298	$+0.0322$
Qwen-3B	MMLU-math	$+0.0034$	0.854	1.000	0.906	-0.0183	$+0.0110$
Qwen-7B	MMLU-math	-0.0253	0.109	1.000	0.528	-0.0201	-0.0332
OLMo-1B-SFT	MMLU-math	$+0.0369$	0.307	1.000	0.637	-0.0397	$+0.0412$
Tulu3-SFT	MMLU-math	-0.0115	0.659	1.000	0.816	-0.0119	-0.0117
OLMo-1B-DPO	MMLU-math	$+0.0465$	0.215	1.000	0.634	-0.0404	$+0.0515$
Tulu3-DPO	MMLU-math	-0.0299	0.253	1.000	0.634	-0.0022	-0.0360
Tulu3-RLVR	MMLU-math	-0.0395	0.118	1.000	0.528	-0.0009	-0.0484

Table 12: **Reliability share under the verified aggregation rule.** Changes are unweighted, signed cell means; REL share instead uses mean absolute cell changes, as in Equation 5. Resolution is unnormalized.

Group	Cells	Mean ΔREL	Mean ΔRES	REL share
Main group, all tasks	35	+0.0470	-0.0002	93.4%
ARC	12	+0.0253	+0.0011	87.7%
MMLU-Pro	11	+0.0567	-0.0012	94.2%
MMLU-math	12	+0.0599	-0.0005	95.3%
Control group, all tasks	12	+0.0088	-0.0085	46.4%

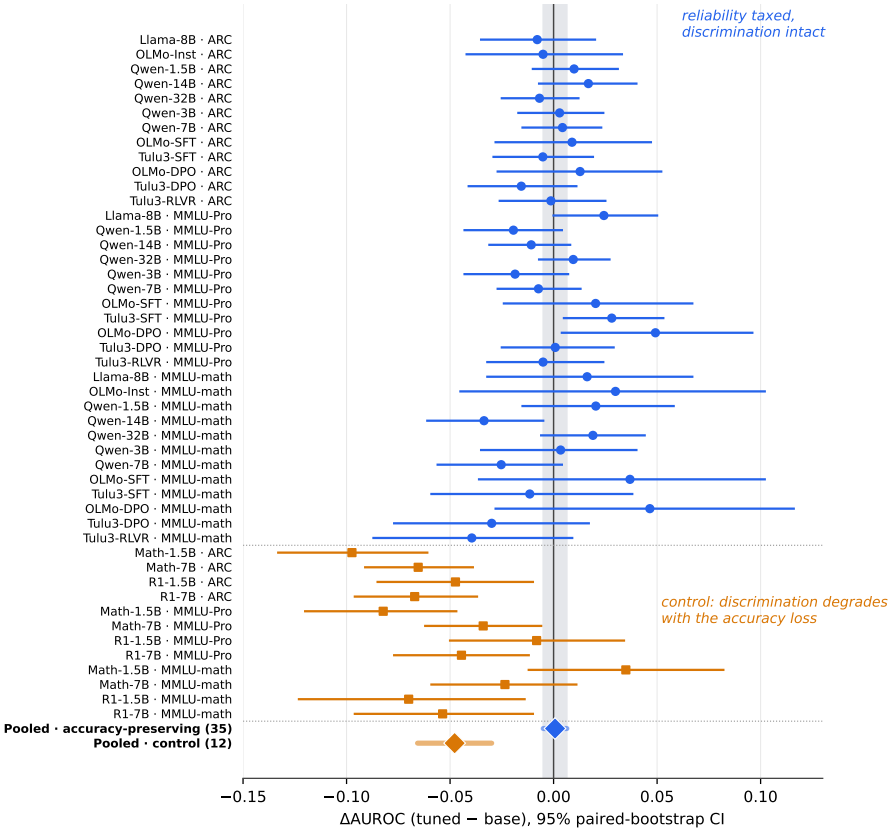


Figure 7: AUROC changes with per-cell 95% paired-bootstrap intervals and random-effects pooled estimates shown as diamonds. The upper block contains instruction/SFT/DPO/RLVR and the lower block math-RL/distillation. Shading marks the main group's pooled 95% interval $[-0.0050, +0.0064]$ under a working independence assumption across cells. Individual intervals show cell-level variation. The two pooled intervals do not overlap.

Table 13: **Binning sensitivity of the Murphy-component comparison.** All 47 cells, retaining the original group definitions. Signed component changes are equal-cell means; reliability shares use mean absolute cell changes as in Equation 5, not ratios of the signed means. The shaded row is the reported protocol.

		Main group (35 cells)			Control (12 cells)
Binning	K	ΔREL	ΔRES	REL share	REL share
Equal-frequency	5	$+0.0460$	-0.0004	94.1%	47.9%
Equal-frequency	10	$+0.0470$	-0.0002	93.4%	46.4%
Equal-frequency	15	$+0.0471$	-0.0003	93.6%	46.4%
Equal-frequency	20	$+0.0474$	-0.0003	93.4%	45.8%
Equal-frequency	30	$+0.0473$	-0.0005	93.5%	46.8%
Equal-width	5	$+0.0435$	-0.0056	86.4%	46.0%
Equal-width	10	$+0.0447$	-0.0036	89.9%	46.3%
Equal-width	15	$+0.0454$	-0.0025	90.7%	47.2%
Equal-width	20	$+0.0456$	-0.0023	90.6%	45.8%
Equal-width	30	$+0.0463$	-0.0014	91.5%	47.0%

differing accuracies, so order-dependent variation, including sampling noise, can contribute to both REL and RES.

We therefore repeat the calculation with greedy approximately equal-frequency binning that never splits an equal-confidence cluster; see Table 14. Main-group mean $\Delta\mathrm{REL}$ changes from $+0.0470$ to $+0.0465$, a reduction of about 1.1%; the control-group mean remains $+0.0088$ at the reported precision. There is one sign flip in the main group, with 32 of 35 cells still positive, and no sign flips in the control group, with 10 of 12 positive. The sole flip is Tulu3-SFT on ARC, from $+0.0002$ to -0.0005; its original $\Delta\mathrm{REL}$ bootstrap interval $[-0.0046, +0.0042]$ already contains zero. These counts track point-estimate signs under the new binning. The largest absolute per-cell shifts are -0.0023 for Qwen-3B on MMLU-math in the main group and $+0.0011$ for Math-1.5B on MMLU-math in the control group.

Table 14: **Sensitivity to keeping confidence ties intact.** Equal-cell mean $\Delta\mathrm{REL}$ under the reported rank-binning protocol and greedy tie-aware re-binning. The difference is tie-aware minus reported. Sign flips and positive-cell counts concern point estimates, not significance. The control-group mean difference rounds to zero.

Group	Cells	Reported	Tie-aware	Difference	Sign flips	Still positive
Main	35	$+0.0470$	$+0.0465$	-0.0005	1/35	32/35
Control	12	$+0.0088$	$+0.0088$	$+0.0000$	0/12	10/12

The invariance argument in Section 3.2 is unchanged: a strictly increasing map of scalar confidence preserves both ranks and ties, and stable sorting preserves the order within ties when the input order is fixed. The rank-bin memberships, and hence binned RES, are then identical. This statement holds for a fixed prediction/correctness vector; it does not imply shared bins between base and tuned checkpoints or rank invariance under multiclass temperature scaling.

B.5 Per-Cell Equivalence Margins

This post-hoc sensitivity analysis quantifies the per-cell equivalence margins supported by the data underlying Section 4.2.2.

Smallest supportable equivalence margin. For each main-group cell we take the smallest symmetric margin containing its 90% interval, $\delta^* = \max(|\mathrm{lo}|, |\mathrm{hi}|)$. A symmetric equivalence margin must exceed this boundary for the interval to lie strictly inside it. Across the 35 main-group cells the median is 0.0399 and the maximum 0.1043: 23 of 35 lie within ± 0.05 and 34 of 35 within ± 0.10, while none reaches ± 0.02. The three widest are the OLMo MMLU-math cells. This per-cell analysis uses unadjusted intervals: the data support relatively broad equivalence margins, with none as narrow as two AUROC points.

B.6 Random-Effects Pooling of AUROC Changes

Table 15 gives DerSimonian–Laird random-effects estimates for the current 35 main-group and 12 control cells, together with task and recipe strata. These weighted pooled estimates differ from the unweighted cell means in Table 2. All intervals use a working independence assumption across cells; dependence from shared checkpoints and evaluation items is not modeled.

All three main-group task intervals contain zero. Main-group task strata have 11–12 cells, control-task strata four, and RLVR three, limiting the precision of between-cell variance estimates. ARC and RLVR estimates reach the variance estimator's nonnegative boundary. These small-stratum summaries use the working independence assumption above; the intervals can understate uncertainty in the estimated between-cell variance.

C Alternative Confidence Signals and Recalibration Baselines

We evaluate alternatives to option-softmax confidence and fitted temperature on the same items and predictions used in Section 4.3.

Table 15: **Random-effects AUROC changes for the current cell population.** DerSimonian–Laird pooling under a working independence assumption. k counts cells, not independent checkpoints; I^2 describes estimated heterogeneity, with zero estimates not proving homogeneity. Task and recipe blocks regroup the same cells and are not additional independent evidence.

Group / stratum	k	Pooled ΔAUROC	95% CI	I^2
Main group	35	+0.0007	$[-0.0050, +0.0064]$	28.7%
Control group	12	-0.0478	$[-0.0657, -0.0299]$	66.8%
Main group by task				
ARC	12	+0.0010	$[-0.0060, +0.0080]$	0.0%
MMLU-Pro	11	+0.0031	$[-0.0080, +0.0142]$	53.3%
MMLU-math	12	-0.0025	$[-0.0186, +0.0136]$	43.3%
Control group by task				
ARC	4	-0.0690	$[-0.0866, -0.0514]$	19.6%
MMLU-Pro	4	-0.0429	$[-0.0701, -0.0158]$	60.2%
MMLU-math	4	-0.0271	$[-0.0687, +0.0145]$	72.0%
Main group by recipe				
Instruction tuning	20	-0.0004	$[-0.0073, +0.0064]$	32.2%
SFT	6	+0.0108	$[-0.0034, +0.0251]$	4.1%
DPO	6	+0.0047	$[-0.0169, +0.0264]$	46.6%
RLVR	3	-0.0079	$[-0.0254, +0.0096]$	0.0%

C.1 Self-Evaluation as a Confidence Signal

The multiple-choice experiments take confidence to be the softmax over legal option logits. An alternative, used as a baseline in much of the confidence-estimation literature, is to ask the model to judge its own answer and read the probability of "Yes" (Kadavath et al., 2022). We evaluate that alternative on the same items and the same predictions: the model answers under the ordinary protocol, and a second forward pass appends the chosen option and asks whether it is correct.

Table 16 reports both across all 47 cells and all six recipes. Self-evaluation confidence has lower discrimination in every cell. Its AUROC is 0.464–0.736, mean 0.588, against 0.534–0.924, mean 0.724, for the option softmax. Its calibration error is also worse in the aggregate, sometimes by a large margin. The confidence bias is recipe-dependent. Instruction-tuned checkpoints tend to answer "No" almost regardless of the question: Qwen-3B gives $P(\text{Yes})$ between 0.011 and 0.057, which drives its ARC ECE to 0.774 while leaving AUROC at 0.542. Math-RL and distilled checkpoints fail the opposite way, answering "Yes" almost regardless: R1-1.5B gives 0.906–0.944 at accuracies of 0.13–0.45, for ECE up to 0.809. Llama-8B has mean $P(\text{Yes}) \approx 0.5$ but still does not match the performance of its option probabilities.

The confidence source also changes the recipe-level calibration ordering. Under option softmax, the math-RL and distilled pairs have lower mean ECE than the main group, 0.098 against 0.174. Under $P(\text{True})$, the ordering reverses to 0.419 against 0.269. Option softmax provides the stronger discrimination signal across all six recipes, with a mean AUROC advantage of about 0.14.

C.2 Paired Recalibrator Comparisons and Independent-Split Validation

Paired comparisons across all 47 cells. Table 17 supplements the means and lowest-ECE counts in Table 3. For each comparator, we subtract its held-out ECE from NLL-fitted temperature's ECE in the same model–task cell, then average the signed differences equally over cells. These comparisons use the D_A fitting and D_B reporting protocol of Table 3. We use a practical margin ϵ=0.005: a temperature win has $d_r < -\epsilon$, a tie has $|d_r| \leq \epsilon$, and a loss has $d_r > \epsilon$, where d_r is the cell's paired ECE difference. These are descriptive point-estimate categories, not tests of statistical equivalence or significance within individual cells. They differ from "Best in," which counts the lowest ECE across all listed methods without this practical margin.

Temperature's mean difference from Platt is -0.0018, with a reported 95% interval $[-0.0051, +0.0012]$ containing zero. The 18/15/14 win/tie/loss split shows heterogeneous direc-

Table 16: **Option-softmax confidence versus Kadavath-style self-evaluation**, on identical items and identical predictions. $P(\text{True})$ has lower AUROC in all 47 cells. Within each task block the main group, instruct/SFT/DPO, precedes the control group, math-RL/distill.

			ECE↓		**AUROC↑**	
Pair	**Recipe**	**Acc. [%]**	softmax	$P(\text{True})$	softmax	$P(\text{True})$
ARC-Challenge						
Llama-8B	instruct	83.9	0.0400	0.2973	0.8370	0.6749
OLMo-1B	instruct	47.3	0.2751	0.2925	0.6766	0.5754
Qwen-1.5B	instruct	76.4	0.1153	0.3753	0.8351	0.5545
Qwen-14B	instruct	94.5	0.0360	0.5384	0.9135	0.7065
Qwen-32B	instruct	96.0	0.0288	0.5721	0.9237	0.7357
Qwen-3B	instruct	82.7	0.1154	0.7738	0.8782	0.5419
Qwen-7B	instruct	89.8	0.0806	0.7534	0.9135	0.6285
OLMo-SFT	sft	50.8	0.1444	0.3009	0.6905	0.4965
Tulu3-SFT	sft	78.5	0.0343	0.1954	0.8398	0.7014
OLMo-DPO	dpo	46.9	0.2618	0.3877	0.6944	0.5489
Tulu3-DPO	dpo	78.9	0.1214	0.0535	0.8294	0.7063
Tulu3-RLVR	rlvr	78.2	0.1376	0.0378	0.8437	0.7061
Math-1.5B	math-rl	56.6	0.0364	0.1719	0.7278	0.5513
Math-7B	math-rl	79.8	0.0672	0.1122	0.8439	0.6746
R1-1.5B	distill	45.3	0.0564	0.4904	0.6804	0.5438
R1-7B	distill	72.9	0.0302	0.0368	0.7767	0.6699
MMLU-Pro						
Llama-8B	instruct	36.1	0.0801	0.1669	0.7622	0.6636
Qwen-1.5B	instruct	29.3	0.1943	0.0901	0.6995	0.5373
Qwen-14B	instruct	54.1	0.2352	0.3041	0.7917	0.6222
Qwen-32B	instruct	59.3	0.1724	0.3713	0.8114	0.6462
Qwen-3B	instruct	36.2	0.3073	0.3456	0.7103	0.5388
Qwen-7B	instruct	46.6	0.2964	0.3122	0.7720	0.5860
OLMo-SFT	sft	15.6	0.1159	0.2058	0.6332	0.5052
Tulu3-SFT	sft	31.4	0.1460	0.2254	0.7660	0.7133
OLMo-DPO	dpo	15.5	0.2207	0.1006	0.6621	0.5538
Tulu3-DPO	dpo	32.5	0.2522	0.2729	0.7388	0.6948
Tulu3-RLVR	rlvr	32.8	0.2602	0.3095	0.7328	0.6855
Math-1.5B	math-rl	21.3	0.0814	0.4761	0.6366	0.5251
Math-7B	math-rl	31.7	0.0932	0.4146	0.7453	0.5441
R1-1.5B	distill	16.6	0.1538	0.7412	0.6284	0.4640
R1-7B	distill	28.6	0.1786	0.3581	0.7007	0.5646
MMLU-math						
Llama-8B	instruct	22.7	0.0604	0.2262	0.6576	0.5920
OLMo-1B	instruct	10.7	0.2391	0.0672	0.5342	0.4827
Qwen-1.5B	instruct	21.5	0.2190	0.0755	0.6282	0.4889
Qwen-14B	instruct	46.1	0.2633	0.2957	0.7564	0.6024
Qwen-32B	instruct	52.6	0.1865	0.2906	0.7873	0.6284
Qwen-3B	instruct	27.2	0.3683	0.2670	0.6776	0.4779
Qwen-7B	instruct	35.7	0.3815	0.2621	0.7363	0.5981
OLMo-SFT	sft	11.5	0.0623	0.2155	0.5412	0.4952
Tulu3-SFT	sft	19.8	0.0872	0.0998	0.6299	0.6014
OLMo-DPO	dpo	11.4	0.1547	0.0811	0.5508	0.4845
Tulu3-DPO	dpo	20.2	0.1941	0.1038	0.6115	0.6032
Tulu3-RLVR	rlvr	20.6	0.2109	0.1370	0.6019	0.5878
Math-1.5B	math-rl	20.4	0.0778	0.4213	0.6427	0.5401
Math-7B	math-rl	31.8	0.0656	0.4353	0.7381	0.4979
R1-1.5B	distill	12.7	0.1623	0.8091	0.5728	0.4875
R1-7B	distill	23.3	0.1685	0.5557	0.6845	0.6243

tions despite the slightly lower mean ECE. The aggregate differences favor temperature over isotonic regression, ECE-grid temperature, DACA, and the base checkpoint, with the win/tie/loss counts showing how each advantage is distributed across cells. The base checkpoint is a separate predictor and can change answers, unlike the recalibrators of fixed tuned predictions.

Table 17: **Paired held-out ECE differences.** Values are NLL-temperature minus the comparator, so negative values favor temperature. Win/tie/loss uses a practical margin of 0.005; the lower block separates DACA by recipe group.

Comparator / group	Cells	Mean difference	95% interval	Win/tie/loss
Platt	47	−0.0018	[−0.0051, +0.0012]	18/15/14
Isotonic	47	−0.0098	[−0.0140, −0.0057]	29/10/8
Temperature (ECE grid)	47	−0.0037	[−0.0067, −0.0008]	22/16/9
DACA	47	−0.0411	[−0.0555, −0.0281]	32/12/3
Base checkpoint	47	−0.0269	[−0.0364, −0.0183]	33/8/6
DACA: main group	35	−0.0205	[−0.0291, −0.0125]	20/12/3
DACA: control group	12	−0.1014	[−0.1278, −0.0733]	12/0/0

The DACA comparison is particularly heterogeneous across recipes: the mean difference is −0.0205 in the 35-cell main group and −0.1014 in the 12-cell control group. The latter has 12/0/0 wins/ties/losses, whereas the main group has 20/12/3. Base-referenced calibration is competitive in the separate harder-pool analysis, ARC mean-matching ECE is 0.104 versus 0.218 for transported temperature, with base-checkpoint ECE 0.067; see Table 4. Reference quality, fitting-pool shift, and whether predictions are held fixed all matter for interpreting these comparisons.

Independent-split comparison. We compare temperature, Platt scaling, and isotonic regression on an independently shuffled Qwen-1.5B ARC pool, fitting on the localization split and evaluating on the decision split. Table 18 shows that Platt scaling reaches lower ECE than restoration on the same split. All output-space methods preserve every prediction.

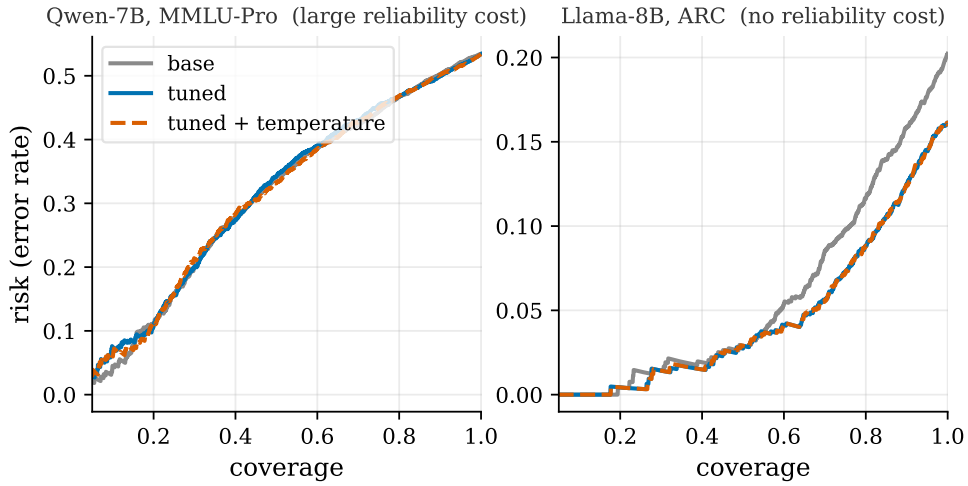


Figure 8: Risk–coverage curves for two cells. The dashed temperature curve overlaps the tuned-model curve in both panels. Left: a large reliability cost with base and tuned nearly indistinguishable in ranking. Right: no reliability cost, but a visible ranking difference.

C.3 Cross-Task Calibration Transfer

We compare cross-task transfer for three recalibrators of increasing capacity, fitting each on one task and reporting on the other in Table 19. This complements the within-task difficulty shifts in Section 4.4 and evaluates the transfer motivation for base-referenced methods (Tan et al., 2026).

Table 18: **Global recalibrators on an independent Qwen-1.5B ARC split.** Output-space methods leave $\arg\max$ untouched by construction; layer restoration changes 5.4% of predictions.

Calibrator, fitted on D_A	ECE↓ on D_B	Accuracy [%]↑	Predictions kept
Tuned model	0.111	76.6	—
Temperature $\theta = 2.0$	<u>0.025</u>	76.6	100%
Platt scaling	**0.021**	76.6	100%
Isotonic regression	0.036	76.6	100%
Layer restoration $k = 22$	0.043	76.3	94.6%

Temperature has the smallest transferred ECE in three of the four directions. For Qwen-1.5B transferring from MMLU-Pro to ARC, the full linear map reaches 0.078 against 0.104 for temperature. The two tasks' accuracies differ by 0.47 for this checkpoint, with accuracies of 0.29 and 0.76, consistent with sensitivity to the fitting pool, but this comparison does not isolate accuracy as a cause. At 7B, where the fitted temperatures on the two tasks are close, 3.51 and 3.37, temperature transfers essentially for free in both directions and is best in both. This is consistent with the pool dependence studied in Section 4.4; these four directions do not establish that accuracy alone determines transferability.

Table 19 uses NLL-fitted temperature throughout. Appendix A.1 reports the corresponding ECE-grid sensitivity result for Qwen-1.5B MMLU-Pro to ARC.

These representation maps are fitted on the last 32 prompt positions of multiple-choice items and evaluated on answer-letter probabilities, whereas BaseCal fits on free-form response tokens (Tan et al., 2026). The comparison evaluates these three methods under our multiple-choice protocol and their respective fitting objectives.

Table 19: **Cross-task transfer.** Target-task ECE after source-task fitting; the in-task column is target-fitted. For representation maps, d is the hidden-state dimension, giving d or d^2 parameters.

Checkpoint	Fit → Report	Method	Params	Transferred↓	In-task
Qwen-1.5B	ARC → MMLU-Pro	temperature	1	**0.034**	0.045
		diagonal	d	0.160	0.163
		linear	d^2	0.232	0.114
	MMLU-Pro → ARC	temperature	1	0.104	0.026
		diagonal	d	0.097	0.098
		linear	d^2	**0.078**	0.071
Qwen-7B	ARC → MMLU-Pro	temperature	1	**0.021**	0.018
		diagonal	d	0.231	0.232
		linear	d^2	0.179	0.132
	MMLU-Pro → ARC	temperature	1	**0.013**	0.015
		diagonal	d	0.071	0.070
		linear	d^2	0.040	0.031

D Calibration under Evaluation-Pool Shift

Per-checkpoint results for Section 4.4, target-label budget curves, composition controls, and comparisons of scalar, adaptive, and label-free repair.

D.1 Pool-Shift Results, Label Budgets, and Label-Free Proxies

Table 25 gives the per-checkpoint numbers underlying the pool characteristics in Table 24 and the corresponding full-pool baseline entries in Table 4. Accuracy means and temperature medians are calculated across all fifteen listed checkpoints within each task. Difficulty is the mean correctness of the other fourteen checkpoints on the same item, so the split never consults the evaluated model. T_E and T_H are fitted by negative log-likelihood on the easier and harder halves; the reported ECE

columns are all evaluated on the harder half. The n=20 column is the median over 400 random draws of twenty labeled items from that half, evaluated on the entire harder half including those fitting items. The full-pool temperature reference is fitted and scored on the harder pool; it is not the held-out target-fit comparison of Appendix D.3. Its mean in-sample ECE across fifteen checkpoints per task is 0.031 on ARC, 0.027 on MMLU-Pro, and 0.021 on MMLU-math.

Dependence checks for quartile transfers. The 540 ordered transfers in Figure 4 comprise twelve directions within each of 45 model–task cells, spanning fifteen checkpoints and three tasks. The pooled estimates are $r = 0.909$ and slope 0.628. Table 20 reports bootstrap intervals using model–task cells as clusters, retaining the twelve directions together rather than resampling individual transfers. The reported interval is $[0.890, 0.929]$ for r and $[0.593, 0.664]$ for the slope.

Leaving out one checkpoint at a time gives $r \in [0.906, 0.913]$ across fifteen analyses; this is a sensitivity range, not a confidence interval. Omitting each task in turn gives correlations from 0.881 to 0.931 and slopes from 0.574 to 0.661. Fitting the relation separately within each cell gives positive slopes in all 45 cells, with median 0.612, showing that the aggregate pattern also holds within individual cells.

Cell clustering models dependence among each cell's twelve transfers. Cross-cell dependence from shared checkpoints and items remains outside this bootstrap model; checkpoint and task omissions assess sensitivity to those groupings. The adaptive-method intervals use the separate procedure in Appendix D.3.

Table 20: **Sensitivity of the quartile-transfer relation.** Excess transferred ECE is regressed on the accuracy gap. Brackets denote bootstrap intervals or leave-one-checkpoint-out ranges; task omissions are point estimates.

Analysis	r	**Slope**
Pooled estimate, 540 transfers	0.909	0.628
Model–task-cell cluster bootstrap	$[0.890, 0.929]$	$[0.593, 0.664]$
Leave one checkpoint out	$[0.906, 0.913]$	—
Without ARC	0.881	0.574
Without MMLU-Pro	0.931	0.661
Without MMLU-math	0.915	0.642
Within-cell fits, median slope (45/45 positive)	—	0.612

Reliability and discrimination by difficulty pool. For the main recipe group, mean reliability cost is 0.0851 on harder halves and 0.0019 on easier halves. The corresponding AUROC changes are 0.0051 and 0.0058, with 95% intervals of $[-0.0058, 0.0160]$ and $[-0.0072, 0.0188]$, respectively. Thus, the large difference in reliability costs accompanies small estimated AUROC changes in both pools.

Held-out and full-pool label-budget evaluation. A full-pool diagnostic evaluates on the whole harder pool, including fitting items. Its mean within-cell median ECE is 0.0377 at n=20 and 0.0321 at n=50, with mean within-cell 90th percentiles 0.0948 and 0.0627. Under this full-pool evaluation, twenty-label cell summaries beat transferred temperature in 44 of 45 cells and base-referenced methods in 38 of 45 under that full-pool evaluation.

The main-text results use strict held-out evaluation: after each permutation, the first n items are used for fitting and the remaining items for reporting. There are 400 draws per cell and budget. For ECE $e_{r,b}^{(n)}$ in cell r and draw b, the reported summaries are

$$\frac{1}{45}\sum_{r=1}^{45} \mathrm{median}_{b=1}^{400}\, e_{r,b}^{(n)}, \qquad \frac{1}{45}\sum_{r=1}^{45} Q_{0.9}\left(\{e_{r,b}^{(n)}\}_{b=1}^{400}\right). \tag{6}$$

Thus, the mean within-cell median is 0.0394 at twenty labels and 0.0338 at fifty; the corresponding mean within-cell 90th percentiles are 0.0948 and 0.0659. Neither summary pools all draws before

taking a quantile. Table 21 gives all thirteen evaluated budgets underlying Figure 5; Table 22 records its reference lines, and Table 23 gives task-specific summaries. P10 and P90 follow the same within-cell-then-across-cell aggregation as the median. Win rates average the within-cell proportions of draws beating each baseline, with equal weight for all 45 cells.

Table 21: **Full target-label budget results under strict held-out evaluation.** Each fit excludes its n labeled items from evaluation. ECE quantiles summarize 400 draws within each cell, then average equally across 45 cells; win rates use the same cell weighting. P90/median reports the ratio of the aggregate summaries.

	Held-out ECE				Win rate [%]		
n	Median↓	P10	P90↓	P90/median	vs. transferred	vs. label-free	vs. uncalibrated
1	0.1149	0.0638	0.5530	4.81	50.1	44.8	54.4
2	0.0988	0.0336	0.3593	3.64	64.3	54.2	68.3
3	0.0769	0.0289	0.2742	3.56	72.9	59.9	75.5
5	0.0612	0.0251	0.2064	3.37	82.6	66.6	83.5
8	0.0515	0.0234	0.1541	2.99	89.5	72.2	89.1
10	0.0487	0.0227	0.1361	2.79	91.5	74.7	91.0
15	0.0418	0.0213	0.1083	2.59	95.1	79.1	94.2
20	0.0394	0.0211	0.0948	2.40	96.3	80.8	95.3
30	0.0361	0.0204	0.0797	2.21	97.9	83.8	97.2
50	0.0338	0.0199	0.0659	1.95	98.8	86.0	98.1
75	0.0324	0.0198	0.0579	1.79	99.3	87.7	98.6
100	0.0318	0.0199	0.0539	1.69	99.6	88.9	98.7
200	0.0321	0.0202	0.0491	1.53	99.7	89.0	98.8

Typical error and upper-tail error improve differently. The mean within-cell median is lowest at n=100 among the evaluated budgets, at 0.0318, and rises slightly to 0.0321 at n=200. Over the same step, mean P90 falls from 0.0539 to 0.0491. The reported P90/median ratio decreases from 4.81 at one label to 1.53 at 200 labels. Additional labels continue to reduce upper-tail error after typical ECE levels off.

Table 22: **Reference lines for the label-budget curves.** ECE is averaged equally across 45 cells. Transferred temperature is fitted on the easier pool. The label-free baseline uses the full unlabeled target pool. The oracle uses all target labels for fitting and is an in-sample reference, not a held-out small-budget result.

Reference	ECE
Uncalibrated	0.2378
Transferred temperature	0.1930
Label-free baseline	0.1322
Target-pool oracle	0.0263

Several harder-half fits reach the temperature search bound of 200, censoring the reported T_H values and their harder-to-easier ratios. Sample-to-sample variation is assessed separately by the held-out label-budget curves in Figure 5. The descriptive variance summary for the bounded $\log T$ fits attributes 19% to between-checkpoint variation and 58% to difficulty. These percentages, the correlation with accuracy, and the temperature ratios are conditional on the search range; robustness to the upper bound has not been established here.

Entropy and agreement as label-free proxies. We evaluate entropy-conditioned temperature and agreement-based accuracy estimation as two ways to recalibrate across difficulty shifts without target labels.

The first conditions the temperature on an observable per-item difficulty proxy, $T_i = \exp(a + b\,h_i)$ with h_i the normalized entropy of the option distribution, fitting (a, b) on the easier half; b=0 recovers ordinary temperature scaling. It transfers *worse* than a single temperature, mean ECE 0.212 against 0.197 on MMLU-Pro, better in 2 of 11 checkpoints, and the fitted b is negative in nine of eleven cases, the opposite of the direction the between-pool relation implies.

Baseline competitiveness and task heterogeneity. The win rate against transferred temperature reaches 91.5% at ten labels, 96.3% at twenty, and 99.7% at 200. Against label-free calibration it reaches 89.0% at 200, leveling off over the largest evaluated budgets. Base-referenced matching remains competitive where the base checkpoint is well calibrated. Task-specific residual errors also differ: at 200 labels, mean within-cell median ECE is 0.0428 on ARC, 0.0297 on MMLU-Pro, and 0.0238 on MMLU-math. ARC rises from 0.0418 to 0.0428 between 100 and 200 labels. The task-specific curves retain residual-error differences that are averaged away in the pooled curve.

Table 23: **Target-label budget results by task.** Mean within-cell median held-out ECE across 15 checkpoints per task, with 400 draws per cell and budget under the strict protocol.

n	ARC	MMLU-Pro	MMLU-math
5	0.1037	0.0472	0.0327
10	0.0757	0.0410	0.0295
20	0.0568	0.0358	0.0257
30	0.0505	0.0337	0.0242
50	0.0456	0.0318	0.0240
100	0.0418	0.0301	0.0236
200	0.0428	0.0297	0.0238

Table 24: **Characteristics of the constructed difficulty pools.** Mean accuracy and median fitted temperature across 15 checkpoints per task. Temperatures are fitted by NLL on the respective halves; some harder-half fits reach the search bound T=200.

Pool statistic	ARC	MMLU-Pro	MMLU-math
Mean accuracy [%], easier half	91.8	54.3	40.4
Mean accuracy [%], harder half	60.3	15.1	13.6
Median T, easier half	0.85	1.68	1.96
Median T, harder half	2.49	8.22	18.03

The second estimates the target pool's accuracy from base–tuned answer agreement and sets the temperature so that mean confidence matches that estimate. The ratio of accuracy to agreement is 0.82 on easier pools and 0.53 on harder ones, and predicting the harder half from a ratio fitted on the easier half gives mean absolute error 0.244. Agreement overestimates accuracy exactly where accuracy is low, because base and tuned checkpoints share errors and agree on the same wrong option.

The proxies share a transfer problem: the relation between each observable signal and accuracy changes across pools. Their harder-pool estimates are systematically overconfident. A labeled target sample directly measures the local confidence–accuracy relation used for recalibration.

D.2 Controls for Item Composition and Base-Model Geometry

We vary the construction of the easier/harder split in Section 4.4 to control legal option count, MMLU-Pro subcategory, and base-checkpoint output geometry separately. The option-count and geometry checks cover 45 model–task cells, whereas subcategory stratification covers the 15 MMLU-Pro cells and its 14 categories. The baseline rows retain each check's reported precision. Here E and H denote the resulting easier and harder pools, and $R = \mathrm{median}_{\mathrm{cells}}(T_H/T_E)$ is the median cellwise ratio, not the ratio of the two displayed temperature summaries. Transferred temperature is fitted on E and evaluated on H. The oracle is the full-target-pool fitted reference, not a held-out small-budget result. Accuracy columns in this subsection are fractions on $[0, 1]$.

Legal option count. Table 26 reports `exp_optmatch.py`: restricting each cell to its modal legal option count, or constructing the split within option-count strata. The median ratio is 5.19 or 5.06, respectively, versus 4.98 on all items, and $T_H > T_E$ in all 45 cells under both controls. Matching option count leaves the temperature contrast intact.

MMLU-Pro subcategory. Table 27 reports `exp_review_checks2.py`, constructing the split within each of MMLU-Pro's 14 categories. The median ratio falls from 8.00 to 5.22, while the

Table 25: **Full-pool per-checkpoint pool-shift diagnostic.** Harder-pool ECE for uncalibrated predictions, transferred temperature, and target-pool fits. Difficulty is the mean correctness of the other fourteen checkpoints, so the split never consults the evaluated model. T_H values at the search bound of 200 are censored, not estimated beyond the imposed bound; the n=20 column is a median over 400 draws, evaluated on the full harder pool *including fitting items*. The full-pool reference is fitted and evaluated on the harder pool.

	Accuracy [%]		**Fitted T**		**ECE on harder half↓**			
Checkpoint	easier	harder	T_E	T_H	none	transferred	n=20	full-pool fit
ARC								
Qwen-1.5B	96.3	59.3	0.91	2.49	0.219	0.233	0.049	0.037
Qwen-3B	98.9	69.2	1.08	3.24	0.211	0.205	0.055	0.037
Qwen-7B	99.2	82.0	1.48	3.78	0.143	0.126	0.043	0.019
Qwen-14B	99.6	90.4	1.31	2.88	0.065	0.057	0.041	0.019
Llama-8B	98.3	71.9	0.51	1.47	0.103	0.196	0.060	0.032
OLMo-SFT	82.4	32.2	0.85	4.25	0.267	0.305	0.058	0.016
OLMo-DPO	77.1	28.1	1.37	8.68	0.402	0.331	0.033	0.034
Math-1.5B	77.9	35.8	0.65	2.30	0.177	0.279	0.050	0.021
Math-7B	96.5	65.7	0.39	0.96	0.048	0.206	0.056	0.039
R1-1.5B	66.0	33.0	0.68	2.31	0.140	0.220	0.066	0.031
R1-7B	91.7	56.6	0.61	1.60	0.133	0.242	0.060	0.034
Qwen-32B	99.8	92.9	1.38	2.32	0.051	0.043	0.044	0.024
Tulu3-SFT	98.3	61.8	0.29	1.49	0.130	0.300	0.061	0.037
Tulu3-DPO	98.1	62.9	0.71	2.75	0.235	0.272	0.059	0.041
Tulu3-RLVR	96.6	62.6	0.94	3.00	0.250	0.256	0.064	0.042
MMLU-Pro								
Qwen-1.5B	50.6	13.0	1.51	18.63	0.282	0.187	0.033	0.023
Qwen-3B	62.6	16.4	1.89	7.57	0.428	0.245	0.044	0.036
Qwen-7B	79.4	23.6	1.70	5.59	0.453	0.319	0.045	0.026
Qwen-14B	79.0	30.4	1.74	4.45	0.387	0.245	0.053	0.028
Llama-8B	66.0	12.6	0.72	4.15	0.215	0.287	0.031	0.030
OLMo-SFT	24.9	8.2	1.83	200.00	0.162	0.095	0.030	0.030
OLMo-DPO	23.0	9.5	2.80	200.00	0.241	0.088	0.015	0.015
Math-1.5B	34.8	10.7	1.35	200.00	0.157	0.114	0.007	0.007
Math-7B	56.0	13.1	0.81	5.67	0.217	0.273	0.027	0.009
R1-1.5B	23.1	11.4	1.97	200.00	0.191	0.085	0.008	0.008
R1-7B	47.0	14.4	1.31	5.85	0.263	0.191	0.037	0.011
Qwen-32B	82.4	37.2	1.37	3.29	0.294	0.218	0.059	0.039
Tulu3-SFT	61.1	7.8	0.96	8.22	0.287	0.295	0.052	0.054
Tulu3-DPO	63.0	8.6	1.68	13.47	0.400	0.283	0.049	0.047
Tulu3-RLVR	62.3	9.6	1.79	15.18	0.402	0.270	0.036	0.035
MMLU-math								
Qwen-1.5B	34.4	11.3	2.10	20.54	0.277	0.126	0.011	0.002
Qwen-3B	46.9	12.0	2.49	11.67	0.462	0.203	0.020	0.018
Qwen-7B	61.8	16.5	2.43	6.55	0.515	0.267	0.054	0.053
Qwen-14B	70.1	29.3	1.96	3.93	0.374	0.192	0.048	0.032
Llama-8B	39.2	9.7	0.88	6.51	0.146	0.167	0.023	0.022
OLMo-SFT	14.2	9.4	2.21	200.00	0.076	0.037	0.008	0.008
OLMo-DPO	13.1	10.0	5.65	200.00	0.158	0.024	0.002	0.002
Math-1.5B	32.9	10.4	1.49	103.34	0.145	0.094	0.002	0.003
Math-7B	54.2	14.4	0.77	2.66	0.176	0.245	0.039	0.026
R1-1.5B	16.6	9.5	3.41	200.00	0.187	0.052	0.007	0.007
R1-7B	38.9	10.8	1.26	7.92	0.262	0.203	0.017	0.019
Qwen-32B	76.3	36.7	1.53	2.85	0.270	0.160	0.054	0.038
Tulu3-SFT	34.2	8.3	1.06	18.03	0.162	0.153	0.025	0.025
Tulu3-DPO	36.6	7.2	1.94	32.84	0.265	0.154	0.034	0.035
Tulu3-RLVR	36.5	8.2	2.18	26.83	0.275	0.140	0.026	0.027

Table 26: **Legal-option-count controls, 45 cells.** k_E and k_H summarize legal option counts in each pool; R is the median cellwise temperature ratio. ECE is reported on the harder pool. The count column reports cells with $T_H > T_E$; r correlates $\log T$ with accuracy across the resulting pools.

						Harder-pool ECE↓				
Split / restriction	k_E	k_H	T_E	T_H	R	**Uncal.**	**Transfer**	**Oracle**	$T_H > T_E$	r
All items	7.61	7.85	1.37	5.85	4.98	0.2378	0.1930	0.0263	45/45	−0.710
Modal k only	8.00	8.00	1.37	6.66	5.19	0.2377	0.1903	0.0259	45/45	−0.702
Within-k split	7.82	7.81	1.37	6.40	5.06	0.2405	0.1940	0.0258	45/45	−0.704

harder-pool temperature remains higher in all 15 cells. The temperature contrast is smaller within subcategories.

Table 27: **Subcategory-stratified splits on MMLU-Pro, 15 cells.** Accuracy is a fraction. R is the median cellwise T_H/T_E ratio; transferred and oracle ECE are evaluated on the resulting harder pool.

Split	acc_E	acc_H	T_E	T_H	R	**Transfer ECE**	**Oracle ECE**	$T_H > T_E$
All items	0.544	0.151	1.68	8.22	8.00	0.2131	0.0267	15/15
Within 14 categories	0.523	0.166	1.66	6.33	5.22	0.2095	0.0255	15/15

Base-checkpoint output geometry. Table 28 reports `exp_geommatch.py`, splitting within deciles of the base checkpoint's margin or entropy. The stratification variables come from the base checkpoint; the fitted temperatures use the tuned logits. The median ratio falls from 4.98 to 3.23 with margin stratification and 3.13 with entropy stratification, while $T_H > T_E$ remains true in all 45 cells. The base-margin summaries become closer, from $1.99/0.99$ without stratification to $1.46/1.39$ and $1.47/1.38$, and the easier–harder accuracy difference also narrows. Both geometry and subcategory stratification therefore attenuate the temperature contrast while preserving its direction. These separate controls leave question format and distractor quality uncontrolled. Oracle ECE rises from 0.0263 to 0.0351 and 0.0377 on the reconstituted harder pools, which have different composition and accuracy from the original harder pools.

Table 28: **Base-geometry-stratified splits, 45 cells.** The margin column reports base-checkpoint margin summaries for E/H, including for the entropy-stratified split. Accuracy is a fraction; R is the median cellwise temperature ratio. ECE is evaluated on each resulting harder pool. r correlates $\log T$ with accuracy across pools.

Split	**Base margin** E/H	acc_E	acc_H	T_E	T_H	R	**Transfer ECE**	**Oracle ECE**	$T_H > T_E$	r
All items	1.99 / 0.99	0.622	0.297	1.37	5.85	4.98	0.1930	0.0263	45/45	−0.710
Base-margin deciles	1.46 / 1.39	0.575	0.333	1.52	4.21	3.23	0.1740	0.0351	45/45	−0.632
Base-entropy deciles	1.47 / 1.38	0.568	0.337	1.49	3.94	3.13	0.1771	0.0377	45/45	−0.620

D.3 Expressive Recalibrators under Pool Shift

Common target evaluation. This additional experiment comprises 45 easier-to-harder transfers, distinct from the 540 ordered quartile transfers in Section 4.4. Let E be the easier source pool and split the harder pool into calibration items H_{cal} and held-out reporting items H_{test}. For each method m, fit on E or H_{cal} and evaluate both fits on H_{test}, defining

$$\Delta^{(m)}_{\mathrm{transfer}} = \mathrm{ECE}_{H_{\mathrm{test}}}(f^{(m)}_E) - \mathrm{ECE}_{H_{\mathrm{test}}}(f^{(m)}_{H_{\mathrm{cal}}}). \tag{7}$$

Mean pool sizes across cells are $|E|$=666, $|H|$=842, and $|H_{\mathrm{cal}}|$=$|H_{\mathrm{test}}|$=421. Mean accuracy is 0.6215 on E, 0.2963 on the full harder pool H, and 0.2964 on H_{test}. The source and target protocols use different fitting budgets; target-fit ECE is measured on held-out items. The twenty-label experiment below provides a common-budget comparison across methods.

The easier/harder assignment is deterministic: the other-checkpoint mean correctness score is thresholded with a strict greater-than-median comparison, with the complementary set forming the harder pool. Ties are not randomly split, so the two pools need not be equal in size. The harder pool is randomly divided into H_{cal} and H_{test} using seed 11, without stratification. Internal learned-module validation uses seed 0; conservative source validation uses seed 3; the hundred twenty-label draws use seed 7.

Paired intervals and dependence. For Tables 29, 30, 31, and 32, each reported 95% interval is a percentile interval from 4000 bootstrap resamples generated with seed 2026. The bootstrap operates on a length-45 array of cell-level paired differences. Each resample draws 45 cells with replacement and takes their mean. Items and draw repetitions are not the resampling units. The intervals treat cells as independent. Shared models across tasks and shared evaluation items across checkpoints introduce cross-cell dependence outside this bootstrap model.

Implementations and scope. The comparison includes scalar temperature, Platt and isotonic calibration, and multiple-choice adaptations of SMART (Guo et al., 2026), ATS (Xie et al., 2024), and DAC (Tomani et al., 2023). The implementations below specify the multiple-choice features, parameterizations, and fitting procedures. The primary comparison uses NLL for SMART and DAC; additional fits use SoftECE and squared error, respectively; see Table 33. Confidence-threshold selective recalibration is evaluated separately; see Table 34.

Common learned-module fitting. The shared fitting routine uses full-batch Adam with learning rate 3×10^{-2} for 300 epochs, reduced to 200 in the twenty-label experiment. Regularization is implemented through Adam's `weight_decay`, not a separately added loss term. Features are standardized before fitting; the fitted means and standard deviations are saved and applied at evaluation. Weight decay is selected using an internal 80/20 calibration split, reserving at least four items for validation, and the module is then refitted on all calibration items. Validation labels count toward the twenty-label budget.

SMART adaptation. The input is the standardized gap $g_i = a_{i,\max} - a_{i,2\text{nd}}$ between the largest and second-largest option logits. A 49-parameter network, $\text{Linear}(1, 16) \to \tanh \to \text{Linear}(16, 1)$, outputs a temperature through $\text{softplus}(\cdot) + 10^{-2}$. Training minimizes standard cross-entropy (NLL). The weight-decay grid is $\{10^{-1}, 10^{-2}, 10^{-3}\}$.

ATS adaptation. For the standardized final hidden state x_i at the answer decision position, the head predicts

$$\log T_i = \log T_0 + \text{clamp}(w^\top x_i, -1.5, 1.5), \tag{8}$$

with w initialized to zero, recovering the global temperature T_0. The training target q_i is one-hot on the true option for a correct prediction and uniform over all k_i legal options for an incorrect prediction, with zero mass on illegal options. Here $k_i = \sum_j \mathbb{1}[a_{ij} > -10^8]$, because illegal option logits are masked to -10^9. The first k_i columns contain the complete legal option set in option order; this is not logit-ranked top-k truncation. ARC items almost all have four options. MMLU-Pro has 3–10 options, including 1667 of 2000 items with ten and 104 with four; 1237 of 1351 MMLU-math items have ten. Thus, legal option count can vary within a task and is not matched across the difficulty pools in this adaptive suite. Appendix D.2 reports separate option-count, subcategory, and base-geometry controls for the scalar-temperature comparison. The loss is $-\sum_j q_{ij} \log \text{softmax}(a_i/T_i)_j$. The weight-decay grid is $\{3 \times 10^{-1}, 10^{-1}, 3 \times 10^{-2}, 10^{-2}, 3 \times 10^{-3}\}$.

DAC adaptation. Three density features use the final, three-quarter-depth, and half-depth layers. At each layer, hidden features are scaled by the reference bank's per-coordinate standard deviations. The mean kNN distance is divided by the median of the bank's own corresponding distances, giving $s_\ell(x)$ on a scale near one. The temperature is

$$T_i = \underbrace{\text{softplus}(w_0) + 10^{-2}}_{\tau_0} + \sum_{\ell=1}^{3} \underbrace{\text{softplus}(\rho_\ell)}_{\alpha_\ell} s_\ell(x_i), \tag{9}$$

with each ρ_ℓ initialized to -3. Its objective is NLL, and the weight-decay grid is $\{10^{-2}, 10^{-3}\}$. The reference bank is E for source fitting, H_{cal} for target fitting, and the sampled twenty items for the small-budget experiment. The bank thus comes from calibration data. The neighbor count is 20 for source and target fitting and 5 in the twenty-label experiment. The density coefficients are learned jointly with the scalar offset.

Table 29: **Source versus target fitting on common held-out target items.** Mean ECE across 45 cells. Gaps and reported 95% intervals refer to Equation 7; intervals bootstrap cells under a working independence assumption. SMART and DAC are NLL adaptations. Controls and the ATS adaptation are separated from the primary comparison; dashes denote quantities not reported.

Method	Parameterization	Source-fit	Target-fit	Gap	95% interval
Temperature	1	0.1956	0.0325	+0.1631	$[+0.1413, +0.1850]$
Platt	2	0.2311	0.0325	+0.1986	$[+0.1696, +0.2270]$
Isotonic	Nonparametric	0.2261	0.0350	+0.1911	$[+0.1624, +0.2195]$
SMART	Gap MLP, 49	0.2084	0.0397	+0.1687	$[+0.1464, +0.1922]$
DAC (adapted, learned)	$\tau_0 + \sum_\ell \alpha_\ell s_\ell$	0.1919	0.0364	+0.1555	$[+0.1358, +0.1753]$
DAC without density	w_0 only	0.1998	0.0393	+0.1605	$[+0.1403, +0.1807]$
Conservative	γT_{src}	0.1865	—	—	—
ATS (adapted)	Linear head	0.1904	0.1181	+0.0723	$[+0.0345, +0.1109]$
Uncalibrated	—	0.2413	—	—	—

Transfer error versus method choice. For temperature, Platt, isotonic, SMART, and learned DAC, mean transfer gaps range from 0.1555 to 0.1986, whereas target-fit ECE ranges from 0.0325 to 0.0397; see Table 29. Switching to target-pool fitting improves each of these methods more than switching among their source-fitted versions. Table 30 reports direct paired comparisons with temperature. Learned DAC has lower source-fit ECE in 32 of 45 cells, with a mean difference of -0.0037. The conservative control has a larger mean improvement but wins in only 12 cells; the mean and win count describe different aspects of its cross-cell behavior.

Table 30: **Paired source-fit comparisons with temperature.** Both methods are evaluated on the same H_{test} in each of 45 cells. Negative mean ECE differences favor the listed method. The 95% intervals bootstrap cells under a working independence assumption. SMART and DAC use NLL adaptations.

Method	Better than temperature	Mean difference	95% interval
DAC (adapted, learned)	32/45	-0.0037	$[-0.0061, -0.0012]$
Conservative	12/45	-0.0091	$[-0.0144, -0.0044]$
SMART	8/45	$+0.0128$	$[+0.0085, +0.0170]$
Isotonic	9/45	$+0.0305$	$[+0.0212, +0.0403]$
Platt	4/45	$+0.0355$	$[+0.0259, +0.0454]$
ATS (adapted)	25/45	-0.0052	$[-0.0278, +0.0183]$

Density ablation and a conservative control. Relative to the scalar-only DAC ablation, the full learned density-aware variant reduces source-fit ECE by 0.0079 on average (paired difference -0.0079, reported 95% interval $[-0.0102, -0.0056]$), and improves 38 of 45 cells. Multiplying the source temperature by a constant c gives numerically lower aggregate ECE than the full density-aware variant; see Table 31. Selected values of c have median 1.00 and range 1.00–1.50. Both learned density features and a source-selected scalar adjustment improve transfer over their respective scalar baselines.

The no-density ablation refits the trainable w_0 using the same fitting routine, weight-decay grid, and internal validation as the full density-aware model. This scalar-only fit retains the DAC optimization pipeline, whereas the global-temperature baseline optimizes NLL independently. The conservative multiplier γ is selected on a 20% source-internal validation subset by ECE, from $\{1, 1.25, 1.5, 2, 3, 4, 6\}$, and γT_{src} is then applied to H_{test}. This selection uses no target-pool data.

Table 31: **Density ablation and conservative temperature adjustment.** Mean source-fit ECE on common held-out harder-pool items across 45 cells.

Variant	Source-fit ECE
Full learned density-aware adaptation	0.1919
Without density, with a fitted scalar	0.1998
Conservative temperature γT_{src}	0.1865

Twenty target labels. Table 32 reports the separate small-budget comparison, including validation labels in the twenty-item budget. There are 100 paired draws per cell with RNG seed 7; the label-budget curves use 400 draws per cell. Each draw takes one permutation-based twenty-item subset of H_{cal} and uses it for all four methods. Every method is evaluated on the same fixed H_{test}. Temperature has lower ECE than learned DAC, SMART, and Platt. Thus, learned DAC's modest advantage under source fitting reverses at this target-label budget.

For method m in cell r, let $e_{r,b}^{(m)}$ denote ECE in draw b. We first compute a separate within-cell median for each method, then take the difference:

$$M_r^{(m)} = \mathrm{median}_{b=1}^{100}\, e_{r,b}^{(m)}, \qquad d_r^{(m)} = M_r^{(m)} - M_r^{(\mathrm{temp})}. \tag{10}$$

The table reports $45^{-1}\sum_r M_r^{(m)}$ and $45^{-1}\sum_r d_r^{(m)}$; intervals bootstrap the 45 values $d_r^{(m)}$. The paired statistic is a difference of within-cell medians. Taking the median after forming draw-wise differences would define a different estimator.

Table 32: **Recalibration with twenty target labels.** ECE averages 45 within-cell medians over 100 shared draws. Paired differences are relative to temperature; positive values favor temperature. Intervals bootstrap cells.

Method	ECE	Paired difference	95% interval	Better cells
Temperature	0.0401	—	—	—
DAC (adapted, learned)	0.0560	+0.0159	[+0.0116, +0.0202]	6/45
SMART	0.0657	+0.0256	[+0.0192, +0.0320]	6/45
Platt	0.0905	+0.0504	[+0.0456, +0.0554]	0/45

Original-objective controls. We refit the SMART and DAC adaptations with SoftECE and squared error, respectively, using the same split and common H_{test} across the 45 cells. Table 33 separates this objective change from the NLL results. Both original-objective fits have higher transferred ECE: SMART increases by 0.0357 and DAC by 0.0111. DAC's mean paired difference from scalar temperature changes from −0.0037 under NLL to +0.0074 under squared error, and its better-cell count falls from 32/45 to 21/45. SMART with SoftECE is slightly worse than no calibration in the aggregate: 0.2441 versus 0.2413. Within these multiple-choice implementations, NLL yields lower transferred ECE for both methods.

Table 33: **Original-objective refits on the common harder reporting pool.** ECE is averaged equally across 45 cells. Δ_{NLL} compares each original-objective fit with its NLL adaptation; Δ_T compares it with scalar temperature. Positive differences favor the comparator. Intervals are the reported paired intervals; better-cell counts refer to temperature. All rows use full coverage.

Method	Objective	ECE	NLL ECE	Δ_{NLL}	Δ_T	Paired interval	Better cells
SMART	SoftECE	0.2441	0.2084	+0.0357	+0.0485	[+0.0349, +0.0624]	6/45
DAC	Squared error	0.2030	0.1919	+0.0111	+0.0074	[+0.0009, +0.0151]	21/45
Temperature	NLL	0.1956	—	—	—	—	—
Uncalibrated	—	0.2413	—	—	—	—	—

Selective recalibration and threshold transfer. We evaluate a confidence-threshold selector with temperature chosen by ECE on the retained fitting items, across the same 45 model–task cells.

Table 34 reports two selection protocols on H_{test}. In the absolute-threshold protocol, a threshold set on the source fitting pool is transferred without adjustment. Its nominal coverage need not hold on the target: 90% becomes 73.1%, and 30% becomes 10.6%. In the quantile-controlled diagnostic, thresholds are set using the unlabeled confidence distribution on H_{test} to enforce the requested coverage. This transductive diagnostic separates coverage control from absolute-threshold transfer using unlabeled reporting-pool data.

Source-fit and target-fit ECE are reported after selection from H_{test}; the absolute protocol's actual-coverage column describes the transferred source selector. Retained-set overlap between source-fit and target-fit pipelines is not recorded in these summaries. At full coverage the selective pipeline gives 0.2001 and 0.0327, distinct from the NLL scalar baseline in Table 29. Table 34 keeps this ECE-based fitting and selection protocol separate from the full-coverage NLL comparison.

Across the reported coverage levels, both source-fit ECE and the source-minus-target fitting gap increase monotonically as coverage falls. With quantile control the gap rises from 0.1674 to 0.2351; with absolute threshold transfer it rises to 0.3223, alongside the loss of coverage. Target-fit ECE remains between 0.0323 and 0.0459. The widening gap and falling actual coverage identify two pool-dependent quantities in this pipeline: the recalibration temperature and the selection threshold.

Table 34: **Selective recalibration under difficulty shift.** Equal-cell means over 45 cells; the transfer gap is source-fit minus target-fit ECE. The upper panel transfers absolute thresholds, while the lower uses target-confidence quantiles. Intervals are 95%.

Nominal coverage	Actual coverage	Source-fit ECE	Target-fit ECE	Transfer gap	95% interval
Absolute source-threshold transfer					
1.0	1.000	0.2001	0.0327	+0.1674	[+0.1441, +0.1912]
0.9	0.731	0.2174	0.0323	+0.1851	[+0.1548, +0.2173]
0.8	0.565	0.2336	0.0331	+0.2005	[+0.1629, +0.2392]
0.7	0.433	0.2532	0.0331	+0.2201	[+0.1774, +0.2652]
0.5	0.235	0.2939	0.0356	+0.2583	[+0.2039, +0.3145]
0.3	0.106	0.3669	0.0446	+0.3223	[+0.2448, +0.3972]
Target-quantile control (diagnostic)					
1.0	—	0.2001	0.0327	+0.1674	[+0.1441, +0.1912]
0.9	—	0.2108	0.0323	+0.1786	[+0.1542, +0.2036]
0.8	—	0.2217	0.0325	+0.1892	[+0.1620, +0.2167]
0.7	—	0.2340	0.0328	+0.2012	[+0.1722, +0.2323]
0.5	—	0.2493	0.0353	+0.2140	[+0.1784, +0.2511]
0.3	—	0.2810	0.0459	+0.2351	[+0.1929, +0.2775]

Table 35 provides a descriptive view of the target confidence mismatch, grouping H_{test} by source-fit confidence. The highest-confidence quintile has the highest accuracy but also the largest mean over-confidence gap, $+0.3328$ versus $+0.1032$ in the lowest quintile. The confidence selector retains the region with the largest average signed bias, consistent with its higher ECE at lower coverage. These quintiles describe group-level mismatch. Multiclass temperature scaling can change the ranking used for selection; see Section 3.2.

Table 35: **Overconfidence by target confidence quintile.** Items in H_{test} are grouped by source-fit confidence within each cell, then statistics are averaged equally over 45 cells. The gap is mean confidence minus accuracy, not ECE.

Confidence quintile	Mean confidence	Accuracy	Overconfidence gap
Highest, 1	0.7104	0.3776	+0.3328
2	0.5644	0.3383	+0.2261
3	0.4825	0.3011	+0.1814
4	0.4105	0.2614	+0.1492
Lowest, 5	0.3103	0.2071	+0.1032

Accuracy gaps and the ATS adaptation. Table 36 relates the method-specific transfer gap in Equation 7 to the source–target accuracy gap across the 45 easier-to-harder transfers. These correlations and slopes use a different transfer set from Figure 4. ATS has high ECE under both source and target fitting, at 0.1904 and 0.1181. Its small transfer gap and low correlation therefore coexist with poor target calibration.

Table 36: **Transfer loss versus source–target accuracy gap.** Correlation r and slope across 45 easier-to-harder transfers, with ECE and accuracy expressed on the $[0, 1]$ scale. The response is each method's own source-fit minus target-fit ECE, not excess over a shared temperature oracle.

Method	r	Slope
DAC (adapted, learned)	+0.893	0.430
DAC without density	+0.920	0.449
Temperature	+0.918	0.489
SMART	+0.923	0.517
Isotonic	+0.933	0.632
Platt	+0.938	0.646
ATS (adapted)	+0.160	0.148

D.4 Confidence-Stratified Analysis

Stratifying by the evaluated model's own confidence gives a descriptive view of where excess confidence occurs, complementing the other-model difficulty split in Section 4.4. On MMLU-Pro, signed bias is $+0.24$ to $+0.30$ in the high-confidence group, against $+0.04$ to $+0.17$ in the low-confidence group. On ARC, the high-confidence stratum is nearly calibrated, with bias ≈ 0, while the low-confidence stratum carries a bias of $+0.21$ for Qwen-7B. The location of the reliability tax thus varies across tasks.

E Weight-Space Diagnostics of the Reliability Tax

These supporting analyses ask how well scalar approximations capture effects observed under weight restoration and representation mapping. Temperature reproduces 40–53% of the per-item restoration effect on ARC and 50–69% on MMLU-Pro. Scalar representation maps recover -2.4 to 4.3% of the calibration gap, versus 52–62% for full linear maps. Restoration and representation mapping therefore change outputs beyond scalar rescaling. Appendix E.5 presents the scalar comparisons, Appendix E.3 validates the restoration instrument, and Appendix E.6 provides the full diagnostic measurements.

E.1 Counterfactual Layer Restoration

Let $T^{(\ell \leftarrow B)}$ denote the tuned model with the seven attention/MLP matrices $\{W_Q, W_K, W_V, W_O, W_{\text{gate}}, W_{\text{up}}, W_{\text{down}}\}$ of block ℓ restored to base values. Normalization, embeddings, the output head, and other parameters remain tuned. On a localization split, define $s_\ell = \text{ECE}(T) - \text{ECE}(T^{(\ell \leftarrow B)})$. This is a marginal sensitivity, not an additive layer contribution. Sorting layers by s_ℓ on a localization split and restoring them cumulatively in that order defines the intervention path used below.

E.2 Held-Out Comparison of Layer Restoration and Temperature Scaling

Table 37 separates fitting on D_A, restoration-depth selection on D_B, and reporting on untouched D_C. In these three comparisons, temperature achieves lower reporting-set ECE than rollback while preserving predictions. Rollback changes 5–6% of predictions and improves accuracy in two cells, but reduces it by 1.55 percentage points on Qwen-1.5B ARC. Appendix A.1 gives the corresponding protocol audits.

Table 37: **Temperature versus restoration on an untouched reporting split.** Layers are ranked and temperatures fitted on D_A; restoration depth k is selected on D_B; all results are reported on untouched D_C. Temperature uses NLL except for Qwen-1.5B on ARC, which uses ECE-grid fitting. Temperature preserves predictions. Bold marks the lowest ECE in each cell.

Model	Task	Method	ECE↓	Accuracy [%]
Qwen-7B	ARC	Tuned model	0.0943	87.98
		Rollback, k=21	0.0393	88.49
		Temperature (NLL)	**0.0169**	87.98
Qwen-1.5B	ARC	Tuned model	0.1209	75.45
		Rollback, k=25	0.0550	73.90
		Temperature (ECE grid)	**0.0335**	75.45
Qwen-7B	MMLU-Pro	Tuned model	0.2993	46.00
		Rollback, k=21	0.0564	47.20
		Temperature (NLL)	**0.0339**	46.00

E.3 Validation of the Layer-Sensitivity Ranking

Appendix E.1 defines an ordering of layers by single-layer calibration sensitivity s_ℓ as a measurement instrument. This appendix evaluates the ordering against shuffled-label controls, measures reproducibility, and tests update directionality. All measurements are on ARC unless noted.

The ordering separates from shuffled-label controls. Figure 9 shows the profile itself: recovery is nonuniform across depth, peaking in a middle band but nonzero throughout. Recomputing every s_ℓ after shuffling each layer's correctness labels, 100 shuffles, gives a rank correlation with the real ordering of -0.02, 95% interval $[-0.40, +0.33]$. Bootstrap intervals identify $\{17, \ldots, 20\}$ as the only cluster excluding zero. At matched restoration depth the sensitivity-informed order reaches lower ECE than random orders, paired bootstrap $P \geq 0.92$ at $k \in \{16, 20, 22, 24\}$, and lower than the reverse order, $P > 0.99$ for $12 \leq k \leq 24$. Figure 10 gives the per-layer bootstrap intervals and Figure 11 the ranked-versus-random-versus-reverse comparison.

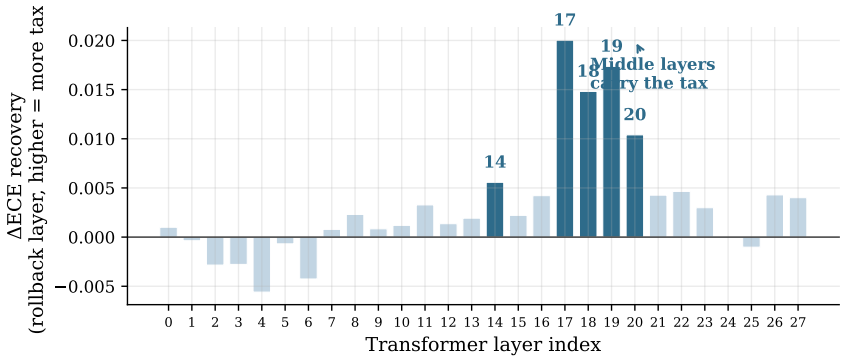


Figure 9: Single-layer ECE recovery is nonuniform across depth, peaking in a middle band but nonzero throughout.

It is reproducible, and transfers partially. Two disjoint Qwen-1.5B pools agree at Spearman 0.78 and share four of five top layers. Qwen-1.5B and Qwen-7B agree at 0.536. For Qwen-7B, ARC and MMLU-Pro profiles agree at 0.468, with a bootstrap 95% interval of $[0.124, 0.710]$ and $p = 0.012$, but share only one of five top layers. Cross-task transfer is therefore partial. Estimating it also needs data: five disjoint 234-item splits give mean off-diagonal Spearman 0.10 against 0.78 at 586 items.

Recovery spans most layers. The minimum-ECE restoration uses 22 of 28 layers at 1.5B and 27 of 28 at 7B. Sensitivity varies with depth, while cumulative recovery draws on most of the

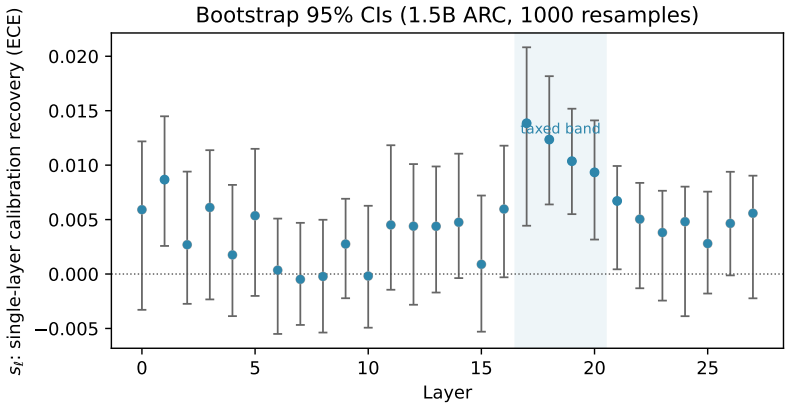


Figure 10: Bootstrap 95% intervals for single-layer calibration recovery on Qwen-1.5B ARC, using 1000 resamples. Only the band $\{17, \ldots, 20\}$ excludes zero.

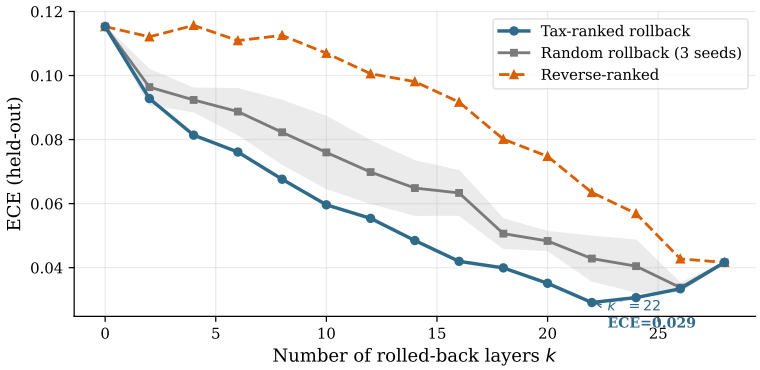


Figure 11: Tax-ranked, random, and reverse-ranked restoration at matched depth. The sensitivity-informed order reaches lower ECE over the informative intermediate range.

stack. Appendix E.4 compares this measured ordering with fixed-form depth priors and temperature scaling.

Update direction predicts calibration sensitivity. Update magnitude alone is a weak predictor of sensitivity: its correlation with s_ℓ is $+0.36$ for Qwen-1.5B, with a 95% interval of $[-0.05, +0.71]$, and -0.16 for Llama-8B, and a magnitude-only regression has five-fold layer-held-out R^2 of -0.05, rising to 0.79 once a permutation-response feature is added. The feature measures a finite ECE difference under the permutation control.

Equal-norm controls are more direct. Interpolating $W_\ell(\alpha) = B_\ell + \alpha \Delta W_\ell$ along the true direction raises ECE monotonically toward the tuned value, whereas equal-norm random element-wise permutations of ΔW_ℓ leave ECE nearly flat at every α in both Qwen and Llama, with the real curve outside the 95% envelope of 50 permutations at full strength, as reported in Figure 12. The contrast is specific to sensitive layers: on band layer L20 the real direction separates from a 30-permutation envelope, 0.1153 against $[0.1051, 0.1096]$, while on control layer L6 it lies inside, 0.1153 against $[0.1134, 0.1204]$.

Joint restoration is subadditive. With $I(S) = s_S - \sum_{\ell \in S} s_\ell$, restoring $\{17, 18, 19, 20\}$ on Qwen-1.5B recovers 0.0387 against a single-layer sum of 0.0459, so $I(S) = -0.0072$. This is nonadditivity of the measured ECE response, combining weight-intervention effects with contributions from the metric's bounds and bin changes.

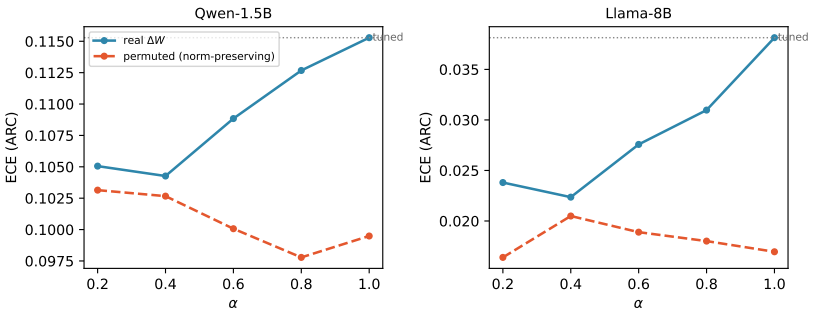


Figure 12: Directional perturbation of a taxed layer. The real direction carries the tax monotonically; an equal-norm permutation of ΔW_ℓ is nearly flat at every α, in Qwen-1.5B layer 17 and Llama-8B layer 11.

Calibration and behavioral sensitivity have different profiles. Single-layer restorations change roughly 1% of predictions everywhere while calibration recovery varies by an order of magnitude. The rank correlations with prediction disagreement and KL are $+0.31$ and -0.20 for Qwen-1.5B, $+0.25$ and $+0.35$ for Llama-8B, and -0.48 and -0.89 for OLMo-1B, as reported in Figure 13. A per-layer correctness probe follows a different depth profile: its AUC grows monotonically from 0.58 to 0.79, while s_ℓ peaks in the middle band, with correlation $+0.40$.

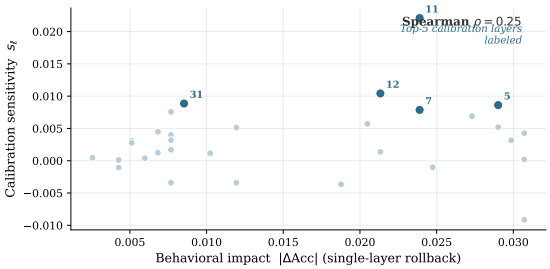


Figure 13: Calibration versus behavioral sensitivity of single-layer restoration. Behavioral impact changes roughly 1% of predictions everywhere; calibration recovery spans an order of magnitude.

Representation scale is otherwise untouched. Single-layer restoration changes the final hidden-state norm by under 1% at 1.5B and the answer-logit standard deviation by 0.98–1.03; at 7B the logit scale stays within 0.95–1.01. The one exception is 7B layer 0, discussed in Appendix E.5. Along the cumulative paths the equivalent temperature climbs monotonically from 1.14 to 1.81 at 1.5B and 1.83 to 3.46 at 7B, the latter matching the independently fitted optimum of 3.45.

E.4 Layer-Scaling Baselines

We compare the sensitivity ordering, estimated from several hundred labeled items, with the prescribed depth profile of LiNeS (Wang et al., 2025a), which scales the update linearly: $W_\ell = B_\ell + \lambda_\ell \Delta W_\ell$ with $\lambda_\ell = \alpha + \beta\,(\ell - 1)/(L-1)$, keeping shallow layers near base and deep layers near tuned. Table 38 compares it against uniform interpolation and against the calibration-informed order, all selected on the same pool, at two scales and on the two tasks whose reliability costs differ by a factor of four.

The ordering is the same in both panels. LiNeS is *worse* than plain uniform interpolation, with ECE 0.046 versus 0.044 on Qwen-1.5B ARC and 0.085 versus 0.072 on Qwen-7B MMLU-Pro. It also trails ranked restoration at 0.028 and 0.050; reversing its direction is worse still. Its depth prior was designed to prevent forgetting, and the calibration tax is not distributed in that shape, which is consistent with the band structure of Figure 9: concentrated in the middle rather than monotone in depth. On the 7B panel the layer ordering is the one estimated on ARC; the transferred ordering outperforms every depth-prior variant.

On Qwen-7B MMLU-Pro, the base model alone reaches 0.0500, within 0.0004 of the best ranked restoration. Each listed partial weight-space interpolation has higher ECE than this base endpoint.

The measured ordering improves on the prescribed depth profiles in both panels. Temperature gives the lowest ECE overall, 0.026 and 0.030, with a single fitted parameter.

Table 38: **Layer-scaling baselines at two scales.** All methods are selected on the same pool and subject to the retained-accuracy constraint. Reported values are measured on the selection set. Depth-profile parameter selection uses ECE and accuracy labels. On the 7B panel the layer ordering is transferred from ARC.

		Qwen-1.5B, ARC		Qwen-7B, MMLU-Pro	
Method	**Labels needed**	ECE↓	Acc. [%]	ECE↓	Acc. [%]
Tuned model	—	0.1153	76.4	0.2851	47.6
Base model, λ=0	none	0.0416	76.5	0.0500	47.6
Uniform interpolation	selection	0.0440	76.4	0.0720	48.0
LiNeS, best α, β	selection	0.0464	76.2	0.0854	48.2
Reverse LiNeS	selection	0.0617	76.9	0.0954	47.3
Tax-ranked restoration	ranking and selection	0.0275	76.6	0.0504	47.6
Temperature, NLL	fitting	**0.0263**	76.4	**0.0295**	47.5

E.5 Scalar and Linear Approximations of Layer Restoration

We compare scalar approximations with the effects of restoration and representation mapping.

Temperature captures part of the restoration effect. For each restoration depth we fit the temperature minimizing a distributional discrepancy to the restored outputs and report the fraction of the per-item confidence change it reproduces. On ARC the cumulative path peaks at 40–53% near k=12–16 and declines, so the maxima are interior rather than at the full stack, as reported in Figure 14. Single layers are about 22% reproduced inside the taxed band of Qwen-1.5B and 2% outside it, 30% and 5% at 7B, and s_ℓ correlates with this fraction at $+0.576$ and $+0.482$: the temperature-like component sits where the tax is.

On MMLU-Pro, where Qwen-7B's tax is four times larger, the same path is 50–69% reproduced against 29–48% on ARC for that checkpoint, so where the reliability cost is large the intervention is substantially, though not entirely, temperature-like. What the scalar misses is about one third of the per-item effect on MMLU-Pro and three fifths on ARC. Consistent with this, the layer that moves representation scale most is 7B layer 0. Its restoration shifts the final hidden-state norm by 17%, compared with under 1% for other layers. It is unremarkable on ARC but the single most sensitive layer on MMLU-Pro, by 57% over the runner-up. On this checkpoint, larger scalar-reproduced effects thus coincide with a stronger response at the scale-sensitive first layer. Separating the roles of first-layer scaling and overall tax magnitude requires varying them independently.

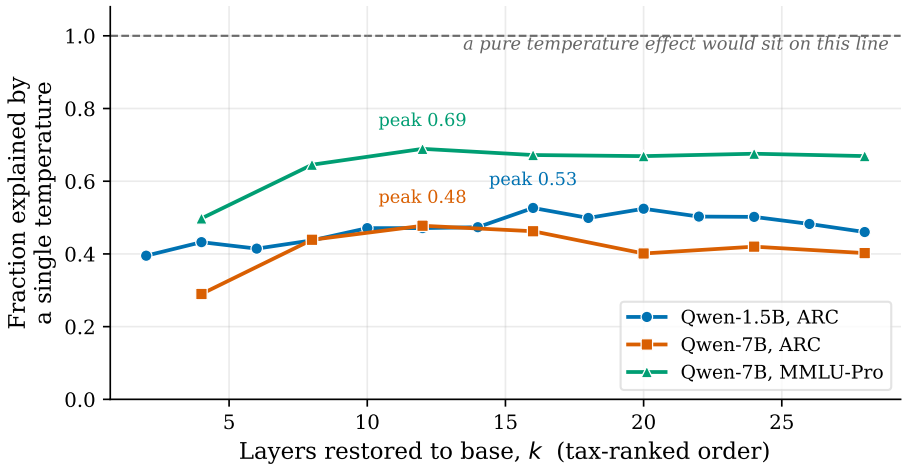


Figure 14: Fraction of the per-item effect of cumulative layer restoration that a single temperature reproduces, against restoration depth. A purely scalar mechanism would sit on the dashed line. The fraction is higher on MMLU-Pro, where the tax is four times larger, than on ARC for the same checkpoint, but does not reach one in either case.

Representation maps need more than a scalar. BaseCal reports that a single linear projection from tuned to base final-layer states recovers substantial calibration (Tan et al., 2026); we ask how much of that map is needed. Fitting maps of increasing capacity and evaluating through the base output layer, as reported in Figure 15, scalar maps recover 4.3% and -2.4% of the calibration gap for Qwen-1.5B and Qwen-7B on ARC, diagonal maps 18.3% and 13.2%, and full linear maps 52.1% and 62.3%. On MMLU-Pro, where Qwen-7B's gap is four times larger, the ordering is unchanged: 2.7%, 24.0%, 61.0%. The fitted scalars are 0.974–0.986, near one, while the full map recovers three times what the diagonal one does. Each capacity level is fitted independently under a least-squares objective, comparing scalar rescaling, per-channel scaling, and cross-channel mixing. The full map lowers MMLU-Pro accuracy by 2.3 points against 0.2 on ARC. The fitted full map has off-diagonal energy fraction 0.053 at 1.5B and 0.270 at 7B on ARC, 0.254 on MMLU-Pro, with relative distance to cI of 0.26, 0.93 and 0.80.

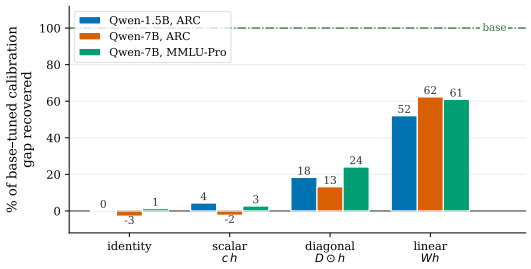


Figure 15: Capacity needed to map tuned final-layer states back to base states, evaluated through the base output layer, at two scales and on the task with the largest tax. A scalar recovers at most 4.3% of the calibration gap, a per-channel diagonal map 13–24%, and a full linear map 52–62%.

E.6 Summary of Diagnostic Measurements

Table 39 collects the magnitude, direction, and interaction measurements for the diagnostic probes. All ARC pools; interpolation $W(\alpha) = B + \alpha\Delta W$ on the indicated layer with all other layers held tuned.

Table 39: **Mechanism experiments, full numbers.** Top: magnitude and direction controls. Middle: measured nonadditivity and submodule decomposition. Bottom: scale and temperature-approximation diagnostics, then representation-map comparisons.

Test	Setting	Value
Magnitude	Qwen-1.5B $\rho(s_\ell, \|\Delta W_\ell\|_F)$	$+0.36$
	Qwen-1.5B $\rho(s_\ell, \|\Delta W_\ell\|_F/\|W_\ell\|_F)$	$+0.44$
	Llama-8B $\rho(s_\ell, \|\Delta W_\ell\|_F)$	-0.16
Direction	Qwen-1.5B L17, real: ECE at $\alpha = 0.2 \to 1.0$	$0.105 \to 0.115$
	Qwen-1.5B L17, permuted: ECE at $\alpha = 0.2 \to 1.0$	$0.103 \to 0.100$
	Llama-8B L11, real: ECE at $\alpha = 0.2 \to 1.0$	$0.024 \to 0.038$
	Llama-8B L11, permuted: ECE at $\alpha = 0.2 \to 1.0$	$0.016 \to 0.017$
50-perm envelope	Qwen-1.5B L17 at α=1.0	mean 0.102, CI $[0.099, 0.106]$
	, real 0.115, outside	
	Llama-8B L11 at α=1.0	mean 0.020, CI $[0.013, 0.027]$
	, real 0.038, outside	
High-/low-tax	L20, band real vs. 30-perm CI at α=1.0	0.1153 vs $[0.1051, 0.1096]$
	L6, low-tax real vs. 30-perm CI at α=1.0	0.1153 vs $[0.1134, 0.1204]$, inside
Regression	$s_\ell \sim \log r_\ell$: R^2	0.16, 5-fold CV: -0.05
	+ permutation-response feature: R^2	0.82, 5-fold CV: 0.79
	$\rho(s_\ell, \Delta\text{acc}_\ell)$ [95% CI]	-0.63 $[-0.84, -0.29]$
Interaction	joint $\{17, 18\}$ vs. sum of singles	0.0235 vs 0.0262, -0.0027
	joint $\{17, 18, 19, 20\}$ vs. sum of singles	0.0387 vs 0.0459, -0.0072
	low-tax control $\{3, 9, 12, 15\}$: $I(S)$	-0.0052
Submodule	total single-layer recovery, MLP only	0.106
	total single-layer recovery, attention only	0.061
Scale invariance	$\|h\|$ ratio under single-layer restoration	0.99–1.00
	answer-logit s.d. ratio	0.98–1.03
Temperature equivalence	band $\{17, \ldots, 20\}$, mean fraction explained	0.219
	remaining 24 layers, mean fraction explained	0.019
	$\rho(s_\ell,$ fraction explained$)$	$+0.576$
	full stack, Qwen-1.5B, peak / final	0.526 / 0.460
	full stack, Qwen-7B, peak / final	0.478 / 0.402
Representation map	recovered gap: scalar / diagonal / linear, 1.5B	4.3% / 18.3% / 52.1%
	recovered gap: scalar / diagonal / linear, 7B	-2.4% / 13.2% / 62.3%
	fitted scalar c, 1.5B / 7B	0.974 / 0.986
	off-diagonal energy fraction of W, 1.5B / 7B	0.053 / 0.270
	relative distance $\|W - cI\|/\|W\|$, 1.5B / 7B	0.26 / 0.93