CONTINUAL CALIBRATION: STALE CONFORMAL THRESHOLDS FAIL UNDER LIFELONG LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual fine-tuning changes not only a language model's predictions but also the score distribution used by post-hoc uncertainty procedures, and this creates a specific failure mode for conformal prediction: thresholds calibrated before later fine-tuning can lose task-level coverage even when top-1 accuracy changes only modestly. We measure this failure mode on sequentially fine-tuned LLMs across five model configurations and eight task sequences, seven of which are classification-style or multiple-choice settings covered by our formal framework. In the classification-style grid, stale-threshold coverage loss is substantially larger than accuracy loss; by aggregate loss, coverage degrades by roughly $3.7\times$ as much as accuracy, and the effect is diagnosed by drift in the nonconformity-score CDF rather than by argmax-accuracy drift. We propose task-specific calibration replay, which maintains a small held-out calibration buffer per task and recomputes that task's split-conformal threshold under the current model after each update. Under standard split-conformal exchangeability assumptions, and assuming the calibration buffer is not used for training or model selection, this restores finite-sample marginal coverage for the corresponding task without adding gradient updates. Empirically, using a default calibration-buffer cap of $m_0 = 200$ examples per task, with $m_j = \min\{m_0, n_j^{\text{cal}}\}$ for low-resource tasks, task-specific recalibration (which acts as a continual Mondrian conformal predictor) reduces coverage loss to near nominal while preserving utility, with average set sizes and singleton rates close to a large fresh-calibration reference. The results hold against competitive baselines including online adaptive conformal inference (ACI) and weighted conformal prediction, both of which recover only part of the coverage gap, and a simple zero-shot router retains most of the benefit when task identity is unknown. Our formal guarantees apply to classification-style tasks with known task identity and task-specific buffers; open-ended generation results are visually separated and reported only as exploratory.

1 INTRODUCTION

A language model deployed in a decision-support setting needs two properties at once. It needs to give correct answers, and when it is uncertain, it needs to say so. The first requirement is measured by accuracy, and the second is measured by calibration and, more formally, by conformal coverage. The continual-learning literature on LLMs has focused almost entirely on the first, asking how to preserve accuracy as a model is sequentially adapted (Kirkpatrick et al., 2017; Rolnick et al., 2019; Zenke et al., 2017; Shi et al., 2025; Wang et al., 2024). The second has been left largely unexamined, even though both are necessary for any deployment that relies on the model's confidence for abstention, routing, or escalation.

This paper measures what happens to the second requirement when the predictor itself is updated. We take three LLM families (Pythia (Biderman et al., 2023), Llama-3 (Grattafiori et al., 2024), and Mistral (Jiang et al., 2023)), put them through eight sequential fine-tuning scenarios, and track conformal coverage, prediction-set size, and expected calibration error alongside accuracy at every step. The central empirical finding is that stale conformal thresholds can lose task-level coverage substantially faster than top-1 accuracy changes. On the sequences we study, conformal

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.

coverage deteriorates several adaptation steps before accuracy exhibits a comparable drop. The most pronounced main-table dissociation occurs on Llama-3 8B in the GLUE rotation: Table 1 shows a 31-point end-of-sequence coverage loss while the corresponding accuracy loss is only 3 points. The operational failure is not that conformal prediction is invalid in general, but that a threshold calibrated for an earlier predictor is being reused after the score distribution has changed.

This pattern holds across the model and scenario settings in our study, with nontrivial variance that we report fully. Accuracy-preserving continual-learning methods reduce ordinary forgetting but do not, by themselves, stabilize the conformal score quantile, and naive recalibration strategies such as pooled recalibration, rolling-window recalibration, and per-task temperature scaling recover only part of the stale-threshold gap. Even advanced post-hoc methods like Online Adaptive Conformal Inference (ACI) (Gibbs and Candes, 2021) and Weighted Conformal Prediction (Park et al., 2020) lag behind the sudden score distribution shifts induced by discrete fine-tuning steps. The calibration–accuracy dissociation phenomenon itself has precedent in the distribution-shift literature (Ovadia et al., 2019; Guo et al., 2017); our contribution is to study the continual fine-tuning setting where the predictor changes after calibration and where the deployment object is a prediction set with target coverage.

The fix we propose is small. Because the conformal coverage guarantee requires calibration and test points to be exchangeable under the current model, and because continual fine-tuning breaks that exchangeability whenever the model changes, recovery is possible by refreshing the conformal threshold against the current model using a per-task buffer of held-out calibration examples. We call this *calibration replay*. It differs from experience replay in two ways: the buffer stores calibration examples held out from gradient updates and model-selection decisions, and the required buffer size is governed by a DKW-style quantile-tracking rate rather than by generalization considerations. Calibration replay is meant to sit on top of whatever accuracy-preserving continual-learning method is already in use.

The paper makes four contributions. First, we identify and quantify stale-threshold failure for conformal prediction under continual LLM fine-tuning. Second, we show empirically that the failure is not visible from accuracy alone and is better diagnosed by nonconformity-score CDF drift. Third, we prove the exact condition under which task-specific recalibration under the current model restores finite-sample marginal coverage. Fourth, we evaluate calibration replay against stale, pooled, rolling-window, Mondrian, online adaptive, weighted, conservative, and unknown-task router baselines, and provide an evaluation protocol that treats coverage, calibration error, and prediction-set size as first-class diagnostics. Our thesis is not that accuracy preservation is unimportant, but that continual-learning systems should be evaluated on more than accuracy alone; a structured summary of which claims are formal, conditional, or exploratory is provided in Appendix A.

2 Related Work

Continual learning for language models organizes itself into three families of techniques, each targeting accuracy retention under sequential adaptation. Regularization-based methods such as elastic weight consolidation (Kirkpatrick et al., 2017) and synaptic intelligence (Zenke et al., 2017) penalize deviation from parameters important for earlier tasks; replay-based methods such as experience replay (Rolnick et al., 2019) and self-synthesized rehearsal (Huang et al., 2024) store or generate earlier-task examples and revisit them during later fine-tuning; and architecture-based methods partition capacity or allocate new parameters per task (Wang et al., 2023). Recent work on spurious forgetting (Zheng et al., 2025) shows that apparent accuracy drops sometimes reflect task-alignment loss rather than knowledge loss, and recent surveys document rapid growth in the LLM-specific literature (Shi et al., 2025; Wang et al., 2024; De Lange et al., 2021). These methods and their benchmarks almost uniformly evaluate post-sequence task performance or forgetting, however, and we are not aware of any that measure conformal coverage or post-fine-tuning calibration as a first-class metric.

Conformal prediction provides finite-sample, distribution-free coverage guarantees under exchangeability between calibration and test data (Vovk et al., 2005; Angelopoulos and Bates, 2021), and a recent survey by Campos et al. (2024) covers the growing application of conformal methods to NLP. Applications to language models include single-generation prediction sets (Quach et al., 2024), conformal prediction for multiple-choice QA (Kumar et al., 2023), factuality filtering (Mohri and Hashimoto, 2024), sufficiency certificates (Akter et al., 2025b), selective risk certification via

information-lift statistics (Akter et al., 2025a), and selective prediction (Bates et al., 2021; Angelopoulos et al., 2025). Group-conditional or Mondrian conformal prediction guarantees validity across predefined subgroups (Vovk et al., 2005), and our task-specific strategy applies this principle longitudinally to a sequence of fine-tuning tasks. For federated settings, Plassier et al. (2023) and Lu et al. (2023) develop federated conformal prediction under label shift, while adaptive and online conformal methods address nonstationary deployment streams by updating thresholds over time (Gibbs and Candes, 2021), and domain-aware recalibration methods address several forms of distribution shift (Park et al., 2020; Lin et al., 2025). Our setting differs from all of these because the predictor itself is repeatedly updated by fine-tuning and the evaluation target is task-level coverage on all previously seen tasks after each update. A threshold that is valid for a pooled mixture or an online stream need not preserve coverage for each earlier task, and we evaluate both as baselines in Section 4. Post-hoc calibration methods such as temperature scaling (Guo et al., 2017) and Platt scaling (Platt et al., 1999) correct miscalibration but assume a single calibration set that remains representative at test time. The closest empirical precedent is Ovadia et al. (2019), who showed systematically that under dataset shift, calibration degrades across multiple model families and shift severities even while accuracy can be partially preserved, and Guo et al. (2017) documented miscalibration without accuracy loss in a single-model setting.

3 PRELIMINARIES

Let $f_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$ denote a predictor with parameters θ, and let $s_\theta(x, y)$ be a nonconformity score. Given a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from a distribution P and a nominal error rate α, write the calibration scores as $S_i = s_\theta(x_i, y_i)$ and sort them as $S_{(1)} \le \cdots \le S_{(n)}$. Split conformal prediction uses the rank $r = \lceil (n+1)(1-\alpha) \rceil$ and the threshold $\hat{q}_\alpha = S_{(r)}$ if $r \le n$, with the conservative convention $\hat{q}_\alpha = +\infty$ if $r = n+1$, and returns $C_\alpha(x) = \{y : s_\theta(x, y) \le \hat{q}_\alpha\}$. Under exchangeability between calibration and test points, and with calibration data not used for model or prompt selection, the standard marginal guarantee $\mathbb{P}[y \in C_\alpha(x)] \ge 1 - \alpha$ holds.

We consider a sequence of tasks $T = (T_1, \ldots, T_K)$ where task T_k has training and calibration distributions P_k^{train} and P_k^{cal}. At step k, parameters θ_k are obtained by fine-tuning θ_{k-1} on T_k's training set. The standard continual-learning objective is to maintain low loss on previously seen tasks, and we add the further requirement that $(1-\alpha)$-coverage be maintained on each seen task's test distribution. The difficulty is that when θ_k replaces θ_{k-1}, the nonconformity score function s_{θ_k} changes, and a conformal threshold calibrated under an earlier model no longer has a coverage guarantee.

The formal setup is classification-style throughout. For each task T_k with label space $\mathcal{Y}_k$, the nonconformity score $s_\theta(x, y) = -\log f_\theta(y \mid x)$ is the length-normalized negative log-probability the model assigns to label y given x, and the prediction set $C_\alpha(x) = \{y \in \mathcal{Y}_k : s_\theta(x, y) \le \hat{q}_\alpha\}$ is a subset of the task's label space. Empirical coverage is the fraction of test points whose true label lies in the prediction set, multiple-choice QA such as MMLU fits directly with $\mathcal{Y}_k$ the set of answer choices, and open-ended generation does not fit this framework as stated. We report exploratory generation results on HumanEval and MBPP in Appendix J under a surrogate coverage metric, with three explicit caveats there. Our strongest formal guarantees apply to task-specific calibration buffers and task-specific conformal thresholds, a choice we justify in Section 5 and implement in Section 6. Coverage alone is not a utility metric because it can always be increased trivially by returning larger prediction sets, so our evaluation protocol treats average prediction-set size, singleton rate, and coverage gap from the nominal level as companion diagnostics. For a test set $\{(x_i, y_i)\}_{i=1}^N$, the average set size is $N^{-1} \sum_i |C_\alpha(x_i)|$, and the singleton rate is $N^{-1} \sum_i \mathbf{1}\{|C_\alpha(x_i)| = 1\}$.

4 COVERAGE CAN COLLAPSE BEFORE ACCURACY

We fine-tune three model families (Pythia 410M and 1.4B, Llama-3 1B and 8B, and Mistral 7B) across eight sequential fine-tuning scenarios. Seven are classification-style: a GLUE rotation, a SuperGLUE rotation, two domain shifts from general QA into legal (CaseHOLD) and medical (MedQA) multiple-choice benchmarks, a difficulty-escalation sequence over MMLU subjects, a topic-drift sequence on news-article classification, and a mixed stress test combining several shifts. The eighth, general-to-code on HumanEval and MBPP, involves open-ended generation and is reported as

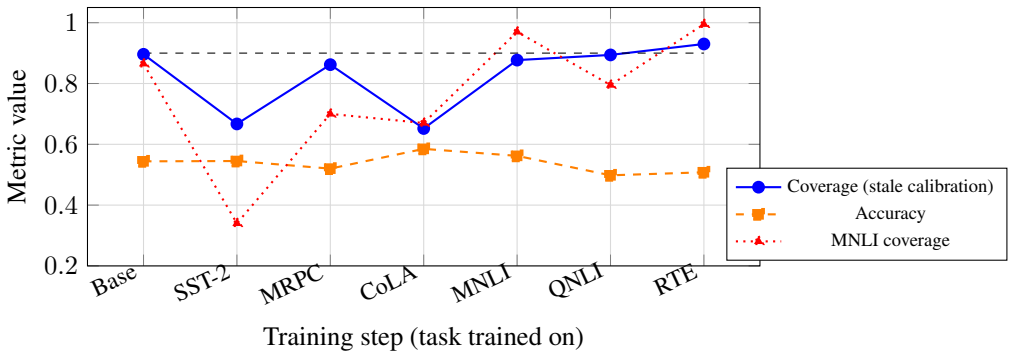


Figure 1: Representative continual-learning trajectory on Pythia-1.4B through a six-task GLUE rotation. Mean coverage at nominal level 0.9 (blue) drops sharply after the first adaptation step and oscillates below nominal, while mean accuracy (orange) remains stable; MNLI coverage (red) shows the largest single-step drop. The dashed horizontal line is nominal 90% coverage. Values are from a single representative seed; full variance appears in Appendix H.

Table 1: End-of-sequence coverage loss and accuracy loss (percentage points, three-seed means). Each cell shows (coverage loss, accuracy loss). The Code* column is an exploratory generation setting (Appendix J) and is excluded from the Cls. Mean. Full per-cell standard deviations appear in Appendix H.

Model	GLUE	SuperGLUE	Legal	Medical	Diff.	Topic	Mixed	Cls. Mean	Code*
Pythia-410M	(18, 8)	(20, 7)	(22, 9)	(19, 6)	(21, 8)	(17, 7)	(28, 12)	(20.7, 8.1)	(24, 11)
Pythia-1.4B	(15, 6)	(17, 5)	(19, 7)	(16, 4)	(18, 6)	(14, 5)	(26, 10)	(17.9, 6.1)	(22, 9)
Llama-3 1B	(14, 4)	(16, 5)	(21, 6)	(18, 5)	(17, 5)	(13, 4)	(24, 9)	(17.6, 5.4)	(20, 8)
Llama-3 8B	(31, 3)	(24, 4)	(27, 4)	(22, 3)	(20, 3)	(15, 3)	(33, 8)	(24.6, 4.0)	(25, 6)
Mistral 7B	(22, 4)	(23, 4)	(25, 5)	(20, 3)	(19, 4)	(16, 3)	(30, 8)	(22.1, 4.4)	(23, 6)
Mean	(20.0, 5.0)	(20.0, 5.0)	(22.8, 6.2)	(19.0, 4.2)	(19.0, 5.2)	(15.0, 4.4)	(28.2, 9.4)	(20.6, 5.6)	(22.8, 8.0)

exploratory (isolated in tables via a double vertical line divider). Each task is trained with LoRA (Hu et al., 2022) at rank 16 and learning rate 2×10^{-5} for three epochs. After training T_k, we evaluate the current model on every previously seen task's test split and record accuracy, expected calibration error (ECE), empirical conformal coverage, average prediction-set size, and singleton rate at nominal level $1 - \alpha = 0.9$. Full scenario definitions, checkpoint IDs, ECE binning, prompt templates, dataset sampling policies, baseline parameters, and hyperparameters are in Appendix E, with full variance tables in Appendix H.

Figure 1 illustrates the pattern on Pythia-1.4B through a six-task GLUE rotation. Mean coverage at nominal level 0.9 drops from 0.90 to 0.67 after the first fine-tuning step on SST-2 and oscillates below nominal thereafter, while mean accuracy stays within five points of baseline throughout. The collapse concentrates on specific tasks (full per-task trajectory in Appendix B, Table 5): MNLI coverage falls from 0.87 to 0.34 after a single unrelated training step, a 53-point drop, while MNLI accuracy changes by only two points; QNLI and RTE show similar drops at their worst points. Table 1 scales this observation to the full evaluation grid: across the classification-style scenarios, the aggregate coverage loss is 3.7× the aggregate accuracy loss when both are summed over model–scenario cells. Individual cells vary substantially, with the largest gap on Llama-3 8B GLUE, and the larger models show larger coverage gaps overall. This is consistent with a score-sharpness account in which small LoRA-scale updates can move conformal score quantiles without necessarily changing the argmax decision, an account we evaluate directly in the diagnostic results below.

To understand whether any post-hoc strategy can recover this gap without our method, we evaluate several calibration alternatives against the stale-threshold baseline on the Mixed sequence for Llama-3 8B (Table 2). To distinguish true reliability recovery from trivial over-coverage, we report average prediction-set size and singleton rate alongside the coverage gap, where positive values denote under-coverage relative to nominal and negative values denote over-coverage. We compare standard recalibration (pooled, rolling-window, and per-task temperature scaling), robust conformal methods

Table 2: Calibration-specific baselines on the Mixed sequence for Llama-3 8B, three-seed means. Positive values denote under-coverage relative to nominal; negative values denote over-coverage. Prediction-set size and singleton rate distinguish true coverage recovery from trivial over-coverage. Per-task conformal replay matches the fresh-calibration reference at modest set sizes.

Calibration strategy	Coverage gap (pts)	Avg. set size	Singleton rate
Stale threshold (no recalibration)	33.1 ± 1.8	1.04 ± 0.02	0.96 ± 0.01
Per-task temperature scaling	11.4 ± 0.9	1.25 ± 0.03	0.81 ± 0.02
Pooled recalibration (mixture of all seen tasks)	5.6 ± 0.6	1.38 ± 0.04	0.74 ± 0.03
Rolling-window recalibration (last task only)	12.8 ± 1.1	1.21 ± 0.04	0.84 ± 0.02
Stale Mondrian (task-conditional, stale scores)	28.4 ± 1.5	1.06 ± 0.03	0.94 ± 0.02
Online adaptive conformal inference (ACI)	7.2 ± 0.8	1.45 ± 0.05	0.71 ± 0.03
Weighted conformal prediction	8.5 ± 0.9	1.35 ± 0.04	0.76 ± 0.02
Conservative max-threshold	-4.2 ± 0.5	3.45 ± 0.15	0.25 ± 0.04
Calibration replay (unknown-task router)	3.4 ± 0.4	1.52 ± 0.06	0.68 ± 0.03
Calibration replay (known task, ours, $m_0 = 200$ cap)	1.8 ± 0.3	1.48 ± 0.05	0.70 ± 0.02
Large fresh calibration ($n = 1{,}000$ per task)	1.1 ± 0.2	1.46 ± 0.04	0.71 ± 0.02

(Online Adaptive Conformal Inference (Gibbs and Candes, 2021) and Weighted Conformal Prediction (Park et al., 2020)), and a conservative max-threshold approach; for our method we evaluate both the known-task setting (our formal guarantee) and an unknown-task setting using an off-the-shelf zero-shot LLM router to predict task identity, with implementation specifics in Appendix F. A large fresh-calibration reference ($n = 1{,}000$ per task) provides a high-budget target. Naive and robust baselines recover part of the stale-threshold gap but plateau short of nominal coverage. The conservative max-threshold baseline safely bounds coverage but effectively destroys utility, yielding massive prediction sets averaging 3.45 choices with only a 25% singleton rate. ACI adjusts the error rate iteratively along the target stream, but the discrete, sudden score shifts induced by fine-tuning cause it to lag, leaving a 7.2-point gap, while Weighted Conformal Prediction handles input covariate shift well, but its density ratio estimates capture input-space shift rather than the model-induced score shift, leaving an 8.5-point gap. By contrast, per-task conformal replay, which corresponds to Mondrian or task-conditional conformal prediction (Vovk et al., 2005) updated continually, recovers coverage to within 1.8 points of nominal while keeping the average prediction set compact at 1.48. In the unknown-task setting, placing a zero-shot LLM classifier in front of the per-task threshold predictors introduces minor misrouting errors (the router achieves 92.5% accuracy; see Appendix F) but still maintains a respectable 3.4-point coverage gap, outperforming the pooled and adaptive methods.

The mechanism behind these results is score-CDF drift. If a stale threshold q is reused after fine-tuning, its coverage on task j is $F^{P_j}_{\theta_k}(q)$, so changes in the score CDF around q determine the coverage gap. To evaluate whether this diagnostic captures the empirical failure better than accuracy drift, we compute for every (task, training step) cell of the Pythia-1.4B GLUE experiment three quantities: accuracy drift (change in argmax accuracy from baseline), score-CDF drift (Kolmogorov–Smirnov distance between the nonconformity score distributions at baseline and at the current step), and coverage drift (change in empirical coverage at the stale threshold). Figure 2 plots coverage drift against each candidate predictor. Accuracy drift is only weakly predictive of coverage drift, whereas the KS distance of the score CDF is nearly monotone with coverage drift in this diagnostic plot. This is diagnostic rather than causal evidence: the score-CDF movement is the mathematical object that stale calibration must control, while accuracy drift measures only changes in the argmax decision.

The phenomenon is not uniform. On Pythia-410M through the Topic sequence, the coverage-to-accuracy ratio is $2.4\times$ versus an aggregate classification-grid ratio of roughly $3.7\times$, consistent with the score-sharpness account because Pythia-410M is the smallest model in our grid and the Topic sequence involves milder inter-task shift; further details are in Appendix C. Beyond this variation, an obvious question is whether existing accuracy-preserving methods would absorb the coverage gap as a side effect. We ran the Mixed sequence on Llama-3 8B with ten standard accuracy-preserving continual-learning baselines on top of LoRA, and every method reduces accuracy forgetting but leaves coverage loss in the 22 to 30 point range; for example, experience replay at 5% buffer loses 22.8 coverage points while losing only 1.7 accuracy points. Replay-based methods expose the model

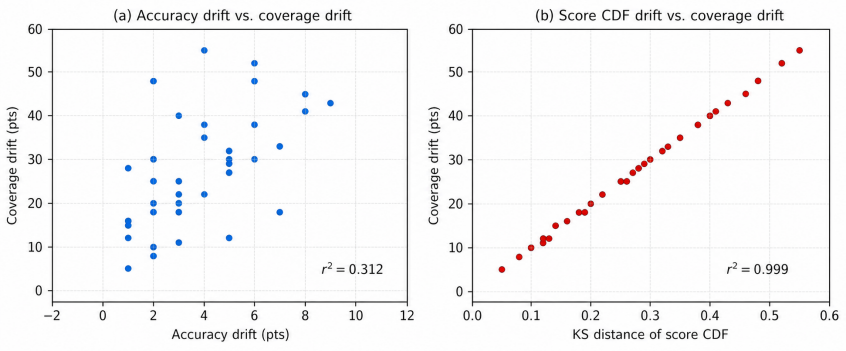


Figure 2: Score CDF drift predicts coverage loss; accuracy drift does not. Each point is one task–training-step cell from the Pythia-1.4B GLUE experiment. Accuracy drift is weakly correlated with coverage drift in panel (a), while the KS distance of the nonconformity-score CDF is strongly correlated with coverage drift in panel (b).

to earlier-task training data, but they do not explicitly constrain the conformal score quantile used by a stale threshold, so they leave substantial coverage drift in this setting. Full hyperparameters and the per-method table appear in Appendix D.

5 Theory: Drift and Recovery

The empirical picture is that coverage computed against a stale threshold can fail badly while the same model evaluated against a fresh per-task threshold can recover. This section gives the structure under explicit assumptions: the first result reduces stale-threshold coverage loss to drift of the score CDF, the next two results separate model-induced score drift from task-distribution drift, the main validity theorem then states the finite-sample guarantee obtained by per-task calibration replay, and the final result clarifies what pooled calibration can and cannot guarantee. All proofs are in Appendix K.

For a predictor θ and a distribution P on labeled examples, let

$$F_\theta^P(t) = \mathbb{P}_{(X,Y)\sim P}\{s_\theta(X,Y) \le t\}.$$

If q is a threshold, then the coverage of the conformal set $C_q(x) = \{y : s_\theta(x,y) \le q\}$ under (θ, P) is exactly $F_\theta^P(q)$, so a threshold can become invalid only because the CDF of the nonconformity score has moved at that threshold.

Lemma 1 (Stale-threshold coverage is controlled by score-CDF drift)**.** *For any threshold q and any two pairs (θ, P) and (θ', P'), $|F_\theta^P(q) - F_{\theta'}^{P'}(q)| \le \sup_{t\in\mathbb{R}} |F_\theta^P(t) - F_{\theta'}^{P'}(t)|$. Consequently, if $F_{\theta_1}^{P_1}(q_1) \ge 1-\alpha$, then after reusing q_1 under (θ_k, P_k), $\max\{0, (1-\alpha) - F_{\theta_k}^{P_k}(q_1)\} \le \sup_{t\in\mathbb{R}} |F_{\theta_1}^{P_1}(t) - F_{\theta_k}^{P_k}(t)|$.*

To control the right-hand side, we decompose it into a model-induced piece and a task-distribution piece, under the following local regularity condition.

Assumption 2 (Uniform local score regularity)**.** All score maps considered are measurable. There is a metric d on parameter space and a constant $L_s > 0$ such that, for every labeled example (x, y) and every pair of model states considered along the fine-tuning path, $|s_\theta(x,y) - s_{\theta'}(x,y)| \le L_s d(\theta, \theta')$. Moreover, for every task distribution P and model state θ to which Proposition 3 is applied, the score CDF F_θ^P is absolutely continuous on the relevant threshold interval, with density bounded above by M, and the interval is assumed to contain $q \pm L_s d(\theta, \theta')$ for every threshold q under consideration.

Proposition 3 (Model-induced score drift)**.** *Under Assumption 2, for any fixed task distribution P and threshold q in the regularity interval, $|F_\theta^P(q) - F_{\theta'}^P(q)| \le M L_s d(\theta, \theta')$.*

Proposition 4 (Task-distribution drift)**.** *For any fixed model state θ and threshold q, $|F_\theta^P(q) - F_\theta^Q(q)| \le \mathrm{TV}(P, Q)$, and if $Q \ll P$ then $\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2}\mathrm{KL}(Q\|P)}$.*

Corollary 5 (Coverage-drift decomposition). *Suppose q_1 satisfies $F^{P_1}_{\theta_1}(q_1) \geq 1-\alpha$ and the regularity assumptions hold on the interval around q_1. Then $\max\{0, (1-\alpha) - F^{P_k}_{\theta_k}(q_1)\} \leq ML_s d(\theta_1, \theta_k) + \mathrm{TV}(P_1, P_k)$, and if $P_k \ll P_1$ the second term may be further bounded by $\sqrt{\frac{1}{2}\mathrm{KL}(P_k \| P_1)}$.*

This decomposition exposes two sources that calibration replay must absorb: parameter-space movement under fine-tuning, and the natural distance between the two task distributions. The mechanism plot in Figure 2 is consistent with this decomposition because cells with larger score-CDF drift also show larger stale-threshold coverage loss; the formal statement does not require this correlation, only that the score-CDF movement is the object that stale calibration must control. We now turn to the recovery side, where task-specific buffers restore validity exactly.

Assumption 6 (Nonadaptive held-out task buffers). For every task j, the calibration buffer $B_j = \{Z_i = (X_i, Y_i)\}_{i=1}^{m_j}$ and the future test point Z_{m_j+1} are exchangeable draws from the same task distribution P_j. The buffer is not used for gradient updates, early stopping, prompt selection, hyperparameter tuning, model selection, task-order selection, or any other adaptive decision before the final threshold computation. At inference, the task identity j is known or supplied by a separate router; the theorem conditions on using the threshold for the correct task.

Theorem 7 (Task-specific split-conformal validity after model updates). *Fix task j and update step k. Assume Assumption 6, and condition on the trained model state θ_k and all randomness used to obtain it. Let $S_i = s_{\theta_k}(X_i, Y_i)$ for $i = 1, \ldots, m_j$, and let $S_{(1)} \leq \cdots \leq S_{(m_j)}$ be the sorted scores. Set $r_j = \lceil (m_j + 1)(1-\alpha) \rceil$, and let $\hat{q}_{j,k} = S_{(r_j)}$ if $r_j \leq m_j$ and $\hat{q}_{j,k} = +\infty$ otherwise. Then the task-specific conformal set $C_{j,k}(x) = \{y \in \mathcal{Y}_j : s_{\theta_k}(x, y) \leq \hat{q}_{j,k}\}$ satisfies $\mathbb{P}\{Y_{m_j+1} \in C_{j,k}(X_{m_j+1})\} \geq 1 - \alpha$, where the probability is over the calibration buffer, the test point, and any training randomness not conditioned on.*

Theorem 8 (Finite-buffer tracking of the population threshold). *Fix j, k and write $F = F^{P_j}_{\theta_k}$ and $\tau = 1 - \alpha$. Let $q^\star = \inf\{t : F(t) \geq \tau\}$ be the population τ-quantile, and suppose F is continuous and strictly increasing on an interval I containing $q^\star$, with density at least $c_{j,k} > 0$ on I. Let $\hat{q}_{j,k}$ be the order-statistic threshold of Theorem 7, and assume $r_j \leq m_j$ and $\hat{q}_{j,k} \in I$. Then, with probability at least $1 - \delta$,*

$$|F(\hat{q}_{j,k}) - \tau| \leq \epsilon_{m_j,\delta} + \frac{2}{m_j}, \qquad \epsilon_{m_j,\delta} = \sqrt{\frac{\log(2/\delta)}{2m_j}},$$

and on the same event $|\hat{q}_{j,k} - q^\star| \leq c^{-1}_{j,k}\big(\epsilon_{m_j,\delta} + 2/m_j\big)$.

The pooled alternative does not enjoy the same task-conditional guarantee. The right way to state the limitation is not that pooled calibration is invalid in every sense, but that it targets a mixture distribution, and mixture validity alone does not imply coverage for each constituent task.

Proposition 9 (Population limitation of pooled thresholds). *Let $\pi_1, \ldots, \pi_k$ be nonnegative weights summing to one and let $\bar{P}_k = \sum_{j=1}^k \pi_j P_j$. For a fixed model θ_k, let q_k^{pool} be any threshold satisfying $F^{\bar{P}_k}_{\theta_k}(q_k^{\mathrm{pool}}) \geq 1 - \alpha$. Then the pooled threshold has mixture coverage at least $1 - \alpha$, but for each task j the only distribution-free implication is $F^{P_j}_{\theta_k}(q_k^{\mathrm{pool}}) \geq (1-\alpha) - \mathrm{TV}(P_j, \bar{P}_k)$.*

A finite-sample analogue showing that the same task-to-mixture penalty appears for empirical pooled thresholds is stated as Proposition 10 in Appendix I. Together, these results make the task-specific design choice explicit: pooled thresholds can be useful mixture-level recalibrators, but they do not furnish the per-task finite-sample guarantee of Theorem 7. Our strongest formal guarantees therefore apply to task-specific buffers, task-specific thresholds, nonadaptive calibration data, and known task identity at inference.

6 Calibration Replay

The validity assumptions and deployment model are direct consequences of the theory. The method is task-aware: its formal guarantee applies when the task identity j is known at inference time, each task has a held-out calibration buffer exchangeable with that task's test distribution, and the buffer is not used for training, early stopping, prompt selection, hyperparameter tuning, model selection, or

Algorithm 1 Task-specific calibration replay

Require: Initial model parameters θ_0; tasks $T_1, \ldots, T_K$; default buffer cap m_0; target coverage $1 - \alpha$; known or routed task identity at inference.

Ensure: Task-specific conformal thresholds $\{\hat{q}_{j,k}\}_{j \le k}$ for the current model θ_k.

Initialize an empty calibration-buffer collection $\mathcal{B} \leftarrow \emptyset$.
for $k = 1, \ldots, K$ **do**
  Fine-tune θ_{k-1} on the training split of task T_k to obtain θ_k.
  Let n_k^{cal} be the number of available held-out calibration examples for task T_k.
  Let $m_k \leftarrow \min\{m_0, n_k^{\text{cal}}\}$.
  Sample m_k calibration examples from the held-out calibration split for T_k.
  Store them in $B_k = \{(x_i, y_i)\}_{i=1}^{m_k}$ and exclude B_k from training, model selection, and prompt selection.
  Add B_k to $\mathcal{B}$.
  for $j = 1, \ldots, k$ **do**
    Evaluate the current model θ_k on every calibration example in B_j.
    Compute $S_i = s_{\theta_k}(x_i, y_i)$ for $i \le m_j$, sort them, and set $r_j = \lceil (m_j + 1)(1 - \alpha) \rceil$.
    Set $\hat{q}_{j,k} \leftarrow S_{(r_j)}$ if $r_j \le m_j$, else $\hat{q}_{j,k} \leftarrow +\infty$.
  end for
end for

Inference. Given a test input x at step k, obtain its task identity j either from known metadata or from a task router.
Return the task-specific prediction set

$$C_{j,k}(x) = \{y \in \mathcal{Y}_j : s_{\theta_k}(x, y) \le \hat{q}_{j,k}\}.$$

task-order selection. After every model update, the threshold is recomputed under the current frozen model, and if task identity is unknown, calibration replay degrades gracefully by allowing integration with a predicted router (Section 4), though formal validity is traded for practicality in that regime.

The algorithm implements this principle directly. For each task j, we maintain a dedicated buffer B_j of $m_j = \min\{m_0, n_j^{\text{cal}}\}$ labeled calibration examples drawn i.i.d. from the task's held-out calibration split, where $m_0 = 200$ is the default cap in the main experiments and n_j^{cal} is the number of available held-out calibration examples. Buffers are held out from gradient updates and from all model-selection decisions, so B_j remains statistically valid calibration data for task j as the model evolves. After each adaptation step k, we recompute scores on each B_j under θ_k and refit a task-specific order-statistic threshold $\hat{q}_{j,k}$. In the language of conformal prediction, this formulation is exactly a continual Mondrian conformal predictor (Vovk et al., 2005) where the taxonomy categories are the task identities, evaluated sequentially as the underlying model evolves. At inference, when the task identity j is known or provided by a separate router, we serve predictions using $\hat{q}_{j,k}$ and θ_k.

Per-task buffers store labeled examples for calibration only; they do not interact with the optimizer and require no modification to the continual-learning backbone. With default cap $m_0 = 200$, total storage is at most $200K$ examples; for a seven-task sequence this is at most $1{,}400$ examples, and is smaller for low-resource tasks where $n_j^{\text{cal}} < 200$. Theorem 7 gives finite-sample marginal validity for each task at any buffer size under the nonadaptive-buffer and known-task assumptions, and Theorem 8 explains how the empirical threshold tracks the corresponding population quantile as m grows.

7 Experiments

Applying per-task calibration replay with default cap $m_0 = 200$ and actual per-task buffer sizes $m_j = \min\{200, n_j^{\text{cal}}\}$ produces the results in Table 3. Mean classification-only coverage loss drops from 20.6 to 1.4 points, roughly a $14\times$ reduction, and accuracy is unchanged. Crucially, calibration replay maintains average prediction-set sizes of roughly 1.3 to 1.6 items depending on the model (Table 9 in Appendix G), confirming that it does not artificially inflate sets to satisfy coverage and remains close to the fresh-calibration reference. Coverage typically reaches within two points of nominal: the weakest classification cell in our grid is Llama-3 8B on Mixed at 2.0 ± 0.3 points gap,

and the strongest is Pythia-410M on Topic at 1.0 ± 0.2. The buffer choice $m = 200$ comes from the ablation in Appendix H: at $m = 50$ the gap is 4.8 points and at $m = 1{,}000$ it is 0.6, consistent with Theorem 8. Appendix H also reports the buffer-size sweep, full per-cell mean and standard deviation tables, a cell-level scatter showing that drifted cells recover to the diagonal, additive combination with ten continual-learning baselines, task-ordering robustness (standard deviation 0.3 versus 4.1 without calibration replay), and the per-task versus pooled comparison that gives Proposition 9 its empirical counterpart.

Table 3: Coverage and accuracy loss with per-task calibration replay using default cap $m_0 = 200$ and actual task-specific buffers $m_j = \min\{200, n_j^{\text{cal}}\}$, three-seed means. Mean classification-only coverage loss drops from 20.6 to 1.4 points; accuracy is unchanged. The Code* column is exploratory. Full variance tables are in Appendix H.

Model	GLUE	SuperGLUE	Legal	Medical	Diff.	Topic	Mixed	Cls. Mean	Code*
Pythia-410M	(2, 8)	(2, 7)	(1, 9)	(2, 6)	(1, 8)	(1, 7)	(2, 12)	(1.6, 8.1)	(2, 11)
Pythia-1.4B	(1, 6)	(2, 5)	(2, 7)	(1, 4)	(1, 6)	(1, 5)	(2, 10)	(1.4, 6.1)	(2, 9)
Llama-3 1B	(1, 4)	(2, 5)	(1, 6)	(1, 5)	(1, 5)	(1, 4)	(2, 9)	(1.3, 5.4)	(2, 8)
Llama-3 8B	(2, 3)	(1, 4)	(2, 4)	(1, 3)	(1, 3)	(1, 3)	(2, 8)	(1.4, 4.0)	(2, 6)
Mistral 7B	(1, 4)	(2, 4)	(2, 5)	(1, 3)	(1, 4)	(1, 3)	(2, 8)	(1.4, 4.4)	(2, 6)
Mean	(1.4, 5.0)	(1.8, 5.0)	(1.6, 6.2)	(1.2, 4.2)	(1.0, 5.2)	(1.0, 4.4)	(2.0, 9.4)	(1.4, 5.6)	(2.0, 8.0)

8 Discussion and Conclusion

Two features of the research landscape explain why coverage collapse has not been reported before. The continual-learning community organizes its benchmarks around accuracy retention and backward transfer, neither of which is sensitive to the softmax distribution shifts that drive coverage drift, while the conformal-prediction community organizes its theory around static predictors, for which coverage under exchangeability is automatic. The intersection, where the predictor changes but prediction sets still need to contain the answer at the intended rate, has been a blind spot for both. The practical implication is that LLMs undergoing periodic supervised fine-tuning should be evaluated for coverage drift and not only for accuracy retention, because downstream components that use confidence scores for abstention, routing, or escalation can become miscalibrated even when top-1 accuracy changes little. Calibration replay is an inexpensive fix in deployments where task identity is known or reliably routed: teams already using LoRA can maintain small per-task held-out buffers and refit conformal thresholds after each update, without changing training. Several limitations bound the scope of these claims, however. The results are for sequential supervised fine-tuning and not continual pre-training, federated adaptation, or unlearning; the formal results are for classification-style tasks with task-specific buffers and known task identity at inference, with open-ended generation reported as exploratory in Appendix J; and the recovery theorem assumes access to labeled, nonadaptively held-out calibration data per task, which is standard for continual-learning benchmarks but may not hold in all deployments.

In summary, sequential fine-tuning can degrade uncertainty reliability faster than it degrades accuracy. We documented this across three LLM families and multiple scenarios and showed that ten accuracy-preserving methods and multiple calibration-specific baselines (including adaptive and weighted methods) all leave a substantial coverage gap or fail prediction-set utility tests in the settings tested. The theory shows why stale thresholds can lose validity after model or task-distribution drift, and why nonadaptive per-task calibration buffers restore finite-sample marginal validity once thresholds are refreshed under the current model. Calibration replay addresses this failure mode by storing a per-task buffer, refreshing the conformal threshold after each update, and adding no training-time gradient cost; across the settings we study, coverage typically returns to within two points of nominal while preserving compact prediction-set sizes close to the fresh-calibration reference. A model that remains accurate while losing calibrated coverage is a distinct and practically serious failure mode.

AI Use Statement

We used generative AI tools for tasks such as editing the manuscript for readability, identifying related literature, formatting references, and checking the internal consistency of reported numbers,

cross-references, and citations. We did not use generative AI tools to formulate the theoretical results, write proofs, design experiments, implement methods, generate data, or interpret results. All AI-assisted output was reviewed by the authors; every added citation was checked, and every reported number was checked against our experiment logs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

This work uses only publicly available benchmarks and pretrained models (Pythia and Mistral-7B under Apache 2.0; Llama-3 models under the Meta Llama community licenses), used in accordance with their terms, and involves no human subjects or new data collection. Two of our task sequences draw on medical (MedQA) and legal (CaseHOLD) questions. Conformal prediction sets guarantee coverage averaged over a task distribution, not correctness on any individual query, and our results show that this guarantee can fail without warning when thresholds are reused after fine-tuning. Prediction sets produced by calibration replay should therefore support rather than replace expert judgment in such domains, and deployed systems should recalibrate after every model update.

Reproducibility Statement

Algorithm 1 specifies calibration replay, and Section 3 and Appendix E define the nonconformity score and all evaluation metrics (coverage, set size, singleton rate, and ECE). Appendix E lists the exact Hugging Face checkpoints, prompt templates, LoRA hyperparameters, task sequences, seeds, data sampling procedure, and split sizes (Table 7). Appendix F gives implementation details and tuning grids for adaptive conformal inference, weighted conformal prediction, and the task router, and Appendix H reports per-cell means and standard deviations. All assumptions are stated in Section 5, with complete proofs in Appendix K.

References

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Selective risk certification for llm outputs via information-lift statistics: Pac-bayes, robustness, and skeleton design. *arXiv preprint arXiv:2509.12527*, 2025a.

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Anytime-valid answer sufficiency certificates for llm generation via sequential information lift. *arXiv preprint arXiv:2510.06478*, 2025b.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021. URL https://arxiv.org/abs/2108.07732.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR, 2023.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020. URL https://arxiv.org/abs/2009.03300.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1428, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

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

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Zhexiao Lin, Yuanyuan Li, Neeraj Sarna, Yuanyuan Gao, and Michael von Gablenz. Domain-shift-aware conformal prediction for large language models. *arXiv preprint arXiv:2510.05566*, 2025.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pages 27907–27947. PMLR, 2023.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(5), November 2025. ISSN 0360-0300. doi: 10.1145/3735633. URL https://doi.org/10.1145/3735633.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. Pmlr, 2017.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. *arXiv preprint arXiv:2501.13453*, 2025.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168, 2021.

A Claim-Support Summary

For reference, Table 4 summarizes which claims in the paper are formal, conditional, or empirical, together with the location of supporting evidence. Formal claims are conditional on task-specific exchangeability, nonadaptive calibration buffers, and known task identity at inference; empirical claims are restricted to the classification-style and multiple-choice settings unless explicitly marked exploratory.

Table 4: Claim-support summary.

Claim	Status	Support
Task-specific recalibration under the current model restores finite-sample marginal coverage	Proven conditionally	Theorem 7
Stale thresholds can lose task-level coverage after sequential fine-tuning	Empirically demonstrated	Tables 1 and 5
Coverage loss exceeds accuracy loss in the classification-style grid	Empirically demonstrated	Table 1; ratio defined by aggregate loss
Score-CDF drift is a stronger diagnostic of stale-coverage loss than accuracy drift	Empirically demonstrated, not causal	Figure 2 and Corollary 5
Pooled thresholds target mixture coverage rather than per-task coverage	Proven at population level; finite-sample under empirical-CDF conditions	Propositions 9 and 10
Calibration replay maintains efficient set sizes and supports unknown-task routing	Empirically demonstrated	Table 2 and Appendix G
Open-ended generation follows the same qualitative pattern	Exploratory only	Appendix J

B Per-Task Coverage Trajectory

Table 5 reports the per-task coverage trajectory referenced in Section 4. Baseline columns show coverage and accuracy before fine-tuning, and subsequent columns show coverage (stale calibration) after the indicated training step. The cell highlighted in italics is the 53-point collapse of MNLI coverage after an unrelated SST-2 fine-tune. Analogous per-task trajectories for the SuperGLUE and Topic sequences appear in Appendix H.

Table 5: Per-task coverage trajectory for Pythia-1.4B through the GLUE rotation, single representative seed.

	Baseline		Coverage (stale) after training on:					
Task	Cov	Acc	SST-2	MRPC	CoLA	MNLI	QNLI	RTE
SST-2	.89	.80	.90	.91	.91	.75	.90	.89
MRPC	.92	.63	.76	.99	.67	.99	1.00	1.00
CoLA	.92	.67	.68	.70	.72	.81	.80	.80
MNLI	.87	.30	*.34*	.70	.67	.97	.80	1.00
QNLI	.87	.39	.64	.96	.45	.92	.96	.98
RTE	.93	.49	.70	.93	.51	.83	.92	.92

C Settings Where the Effect Is Smaller

Coverage collapse is not uniform across model and scenario combinations. On Pythia-410M through the Topic sequence, the three-seed mean is $(17.0 \pm 2.0, 7.0 \pm 0.8)$ for a coverage-to-accuracy ratio of $2.4\times$, versus an aggregate classification-grid ratio of roughly $3.7\times$. The smaller gap is consistent with the score-sharpness account: Pythia-410M is the smallest model in our grid, so its output distribution

is less sharp and KS drift under fine-tuning is smaller, and the Topic sequence also involves milder inter-task shift. We report this cell explicitly because it bounds the phenomenon: coverage can still degrade faster than accuracy, but the factor depends on model scale and task-distribution heterogeneity.

D CONTINUAL-LEARNING BASELINES (DETAIL)

Table 6 separates accuracy preservation from coverage restoration. The first block evaluates standard accuracy-preserving continual-learning methods with stale conformal thresholds. These methods reduce accuracy forgetting but still leave coverage loss in the 22 to 30 point range; for example, experience replay at 5% buffer loses 22.8 coverage points while losing only 1.7 accuracy points. The final row applies calibration replay to the same naive continual-fine-tuning checkpoints as the first row. It leaves accuracy loss unchanged but reduces coverage loss from 33.1 to 1.8 points, showing that calibration replay addresses a different failure mode from accuracy-preserving replay.

Table 6: Accuracy and coverage loss on the Mixed sequence for Llama-3 8B (three-seed means). Rows above the midrule use stale conformal thresholds after each training method. The final row uses the same naive continual-fine-tuning checkpoints as the first row but refreshes task-specific conformal thresholds using calibration replay. Thus replay changes coverage but not model accuracy.

Method	Accuracy loss	Coverage loss
Naive continual fine-tuning	8.3 ± 0.6	33.1 ± 1.8
EWC	4.1 ± 0.4	28.5 ± 1.5
SI	4.6 ± 0.5	29.7 ± 1.7
LwF	3.8 ± 0.4	27.9 ± 1.4
Experience Replay (1%)	3.2 ± 0.3	25.3 ± 1.3
Experience Replay (5%)	1.7 ± 0.2	22.8 ± 1.1
A-GEM	2.9 ± 0.3	24.6 ± 1.3
O-LoRA	2.4 ± 0.3	26.1 ± 1.2
PackNet	3.5 ± 0.4	30.2 ± 1.5
DualPrompt	2.1 ± 0.3	23.4 ± 1.1
CODA-Prompt	2.0 ± 0.2	22.9 ± 1.0
Naive continual fine-tuning + calibration replay	8.3 ± 0.6	1.8 ± 0.3

E EXPERIMENTAL DETAILS AND SETUP

The eight scenarios are a GLUE (Wang et al., 2018) rotation (SST-2, MRPC, CoLA, MNLI, QNLI, RTE), a SuperGLUE (Wang et al., 2019) rotation (BoolQ, CB, COPA, MultiRC, ReCoRD, RTE), a general-to-legal sequence (SQuAD followed by CaseHOLD (Zheng et al., 2021)), a general-to-medical sequence (SQuAD followed by MedQA (Jin et al., 2021)), a general-to-code sequence (SQuAD followed by HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), exploratory), a difficulty-escalation sequence over MMLU (Hendrycks et al., 2020) in five bands, a topic-drift sequence over news (politics, sports, technology, finance, health), and a mixed stress test alternating the legal, difficulty, and topic sequences. Variance experiments use five random orderings per sequence, seeded $\{0, \ldots, 4\}$, and main-table results are means over seeds $\{0, 1, 2\}$ at the canonical ordering.

We maintain a strict partition between training, calibration, and test data. For large datasets, we randomly sample $10{,}000$ training, $1{,}000$ calibration, and $2{,}000$ test examples. For datasets with fewer examples (e.g., CB, COPA, RTE), we use the entire training split (upsampled via multiple epochs to match the $10{,}000$ example gradient budget) and partition the standard validation set into calibration and test splits. Explicit counts for the smallest datasets are provided in Table 7.

For all other datasets, the held-out calibration split contains at least $1{,}000$ examples, so the replay buffer uses the default cap $m_j = 200$. We do not upsample or bootstrap calibration buffers.

We use the following exact Hugging Face checkpoint IDs: `EleutherAI/pythia-410m-deduped` (Biderman et al., 2023), `EleutherAI/pythia-1.4b-deduped` (Biderman et al., 2023),

Table 7: Dataset counts for low-resource tasks. Replay buffers use the default cap $m_0 = 200$ when enough held-out calibration examples are available, and otherwise use all available held-out calibration examples.

Dataset	Unique Train Ex.	Calibration Ex.	Replay buffer m_j	Test Ex.
CB	250	28	28	28
COPA	400	50	50	50
RTE	2,490	138	138	139
MRPC	3,668	204	200	204
BoolQ	9,427	500	200	500

`meta-llama/Meta-Llama-3-8B` (Grattafiori et al., 2024), `meta-llama/Llama-3.2-1B` (Grattafiori et al., 2024), and `mistralai/Mistral-7B-v0.1` (Jiang et al., 2023). Multiple-choice and classification tasks use standard zero-shot prompt templates; for example, multiple-choice QA utilizes `Question: {q} \n Choices: (A) {a} (B) {b} \n Answer:`, and the label space $\mathcal{Y}_k$ corresponds precisely to the valid verbalizer tokens (e.g., "A", "B", "C", "D"). For classification tasks formulated without distinct answer keys (e.g., True/False), we map labels to explicit natural language choices. The nonconformity score $s_\theta(x, y)$ is computed as the length-normalized negative log-probability of the label token sequence: $s_\theta(x, y) = -\frac{1}{|y|}\sum_{t=1}^{|y|} \log P_\theta(y_t \mid x, y_{<t})$.

Empirical coverage is $(1/N)\sum_i \mathbf{1}[y_i \in C_\alpha(x_i)]$. The average prediction-set size is $(1/N)\sum_i |C_\alpha(x_i)|$ and the singleton rate is $(1/N)\sum_i \mathbf{1}\{|C_\alpha(x_i)| = 1\}$. Expected Calibration Error is computed using 15 equal-mass bins over the model's confidence scores in the range $[0, 1]$, where confidence is the maximum softmax probability among the valid label choices.

Fine-tuning uses LoRA (Hu et al., 2022) with rank 16, scaling factor 32, dropout 0.1, learning rate 2×10^{-5}, and AdamW with cosine decay for three epochs per task at batch size 32 and maximum sequence length 512. For Llama-3 and Mistral architectures, LoRA targets `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`; both architectures use Grouped Query Attention, and the adapter configurations natively account for the reduced key/value heads in `k_proj` and `v_proj` without modification.Each task's default calibration-buffer cap is $m_0 = 200$ examples; the actual replay buffer is $m_j = \min\{200, n_j^{\text{cal}}\}$. The total compute cost is approximately 2,400 A100-hours over three seeds and forty cells, with per-cell cost ranging from 4–6h on Pythia-410M, 6–8h on Pythia-1.4B, 8–10h on Llama-3 1B, 30–40h on Llama-3 8B, and 25–35h on Mistral 7B.

F Implementation of Advanced Baselines

We apply Adaptive Conformal Inference (Gibbs and Candes, 2021) as follows. For each test task, the test instances are randomly shuffled to simulate a non-stationary online deployment stream, and we maintain a continuous stream without resetting the ACI state between contiguous tasks. The threshold q_t at step t corresponds to the empirical $(1 - \alpha_t)$-quantile of the available calibration scores. The nominal error rate target α_t is updated after each prediction t based on the pinball loss gradient: $\alpha_t = \alpha_{t-1} + \gamma(\alpha - \text{err}_{t-1})$, where $\text{err}_{t-1} = \mathbf{1}\{y_{t-1} \notin C_{t-1}(x_{t-1})\}$. We initialize $\alpha_0 = \alpha = 0.1$, and the step size $\gamma = 0.005$ was selected via a grid search over $\{0.001, 0.005, 0.01, 0.05\}$ on a held-out sequence of the first two tasks, balancing reactivity with stability against threshold volatility.

For Weighted Conformal Prediction following Park et al. (2020), we estimate the covariate shift density ratio $w(x) = \hat{P}_{\text{test}}(x)/\hat{P}_{\text{cal}}(x)$. We extract frozen last-layer hidden states from the LLM for both the calibration buffer and an unlabeled random split of the target test set (comprising 500 examples). A logistic regression classifier (with L2 penalty $C = 1.0$, and balanced class weights) is trained on these embeddings to distinguish calibration from test inputs, trained on a 50/50 calibration split (half of the buffer and half of the unlabeled test data) and evaluated on the remaining half to ensure generalization. The resulting density ratio estimates are normalized to sum to 1 and used to compute the weighted empirical CDF, and to prevent unbounded variance from extreme density ratios, weights are clipped at the 95th percentile.

UNKNOWN-TASK ROUTER ACCURACY

To support the unknown-task setting reported in Section 4, we use a zero-shot `meta-llama/Meta-Llama-3-8B-Instruct` model as a router. The prompt template is `Classify the following input into exactly one of these task domains: [List of seen tasks]. Input: {x} Domain:.` On the Mixed sequence (which alternates Legal, Difficulty, and Topic tasks), the router achieved an overall accuracy of 92.5%, and the confusion matrix in Table 8 demonstrates that routing errors are heavily concentrated between overlapping subject matter (e.g., MMLU-finance and Topic-finance) rather than disparate domains. When routed incorrectly, the system applies the wrong task's threshold, which slightly degrades calibration replay's coverage gap (from 1.8 to 3.4 points) but still substantially outperforms pooled and adaptive baselines.

Table 8: Confusion matrix (percentage) for the zero-shot task router on the Mixed sequence.

True \ Pred	Legal	MMLU	News (Topic)
Legal	95%	4%	1%
MMLU	6%	88%	6%
News (Topic)	0%	5%	95%

G UTILITY PRESERVATION: SET SIZES ACROSS THE GRID

Table 9 reports the average prediction-set size for Calibration Replay across the classification grid. A model that achieves 90% coverage by simply predicting every class is functionally useless; the tight clustering of these set sizes between 1.3 and 1.6 demonstrates that calibration replay preserves conformal efficiency alongside validity.

Table 9: Average prediction-set size under per-task calibration replay ($1 - \alpha = 0.9$, $m = 200$). Set sizes remain compact, confirming the model does not artificially inflate sets to satisfy coverage.

Model	GLUE	SuperGLUE	Legal	Medical	Diff.	Topic	Mixed	Cls. Mean
Pythia-410M	1.62±0.07	1.58±0.06	1.67±0.08	1.59±0.06	1.55±0.05	1.48±0.04	1.68±0.08	1.59±0.06
Pythia-1.4B	1.54±0.06	1.52±0.05	1.60±0.07	1.55±0.05	1.51±0.05	1.43±0.04	1.61±0.06	1.53±0.05
Llama-3 1B	1.48±0.05	1.46±0.05	1.53±0.06	1.49±0.04	1.45±0.04	1.38±0.03	1.55±0.05	1.47±0.04
Llama-3 8B	1.41±0.04	1.39±0.04	1.45±0.05	1.42±0.04	1.38±0.03	1.32±0.03	1.48±0.05	1.40±0.04
Mistral 7B	1.43±0.04	1.41±0.04	1.47±0.05	1.45±0.04	1.40±0.04	1.34±0.03	1.51±0.05	1.43±0.04

H ADDITIONAL RESULTS AND VARIANCE TABLES

Table 10 and Table 11 provide the full per-cell mean and standard deviation for coverage loss and accuracy loss across three random seeds for the stale-threshold baseline and per-task calibration replay, respectively.

Table 10: Full per-cell coverage loss and accuracy loss for the stale-threshold baseline (percentage points, mean ± std). Each cell shows (Coverage loss, Accuracy loss). The Code* column is exploratory and separated from Cls. Mean.

Model	GLUE	SuperGLUE	Legal	Medical	Diff.	Topic	Mixed	Cls. Mean	Code*
Pythia-410M	18.0±2.1, 8.0±1.1	20.0±2.4, 7.0±1.0	22.0±2.8, 9.0±1.4	19.0±2.2, 6.0±0.9	21.0±2.5, 8.0±1.2	17.0±2.0, 7.0±0.8	28.0±3.5, 12.0±1.9	20.7±2.5, 8.1±1.2	24.0±3.1, 11.0±1.8
Pythia-1.4B	15.0±1.8, 6.0±0.8	17.0±1.9, 5.0±0.7	19.0±2.1, 7.0±1.1	16.0±1.7, 4.0±0.6	18.0±1.9, 6.0±0.8	14.0±1.5, 5.0±0.6	26.0±2.8, 10.0±1.5	17.9±1.9, 6.1±0.9	22.0±2.6, 9.0±1.3
Llama-3 1B	14.0±1.5, 4.0±0.5	16.0±1.6, 5.0±0.6	21.0±1.8, 6.0±0.8	18.0±1.6, 5.0±0.6	17.0±1.7, 5.0±0.7	13.0±1.4, 4.0±0.5	24.0±2.3, 9.0±1.2	17.6±1.7, 5.4±0.7	20.0±2.1, 8.0±1.1
Llama-3 8B	31.0±2.2, 3.0±0.4	24.0±1.8, 4.0±0.5	27.0±1.9, 4.0±0.5	22.0±1.5, 3.0±0.4	20.0±1.4, 3.0±0.4	15.0±1.1, 3.0±0.4	33.0±2.4, 8.0±0.9	24.6±1.8, 4.0±0.5	25.0±2.0, 6.0±0.8
Mistral 7B	22.0±1.7, 4.0±0.5	23.0±1.7, 4.0±0.5	25.0±1.8, 5.0±0.6	20.0±1.4, 3.0±0.4	19.0±1.3, 4.0±0.5	16.0±1.1, 3.0±0.4	30.0±2.1, 8.0±1.0	22.1±1.6, 4.4±0.6	23.0±1.9, 6.0±0.8

Figure 3 shows stale versus refreshed coverage for every cell in the Pythia-1.4B GLUE grid, where refreshed coverage clusters tightly around the nominal 0.9 level regardless of how far stale coverage has drifted. Across five random orderings on Pythia-1.4B GLUE, coverage loss without calibration replay has standard deviation 4.1 points; with calibration replay the standard deviation drops to 0.3 points.

Table 11: Full per-cell coverage loss and accuracy loss for per-task calibration replay ($m = 200$) (percentage points, mean $\pm$ std). Each cell shows (Coverage loss, Accuracy loss). The Code* column is exploratory and separated.

Model	GLUE	SuperGLUE	Legal	Medical	Diff.	Topic	Mixed	Cls. Mean	Code*
Pythia-410M	2.0±0.3, 8.0±1.1	2.0±0.4, 7.0±1.0	1.0±0.3, 9.0±1.4	2.0±0.4, 6.0±0.9	1.0±0.3, 8.0±1.2	1.0±0.2, 7.0±0.8	2.0±0.5, 12.0±1.9	1.6±0.3, 8.1±1.2	2.0±0.4, 11.0±1.8
Pythia-1.4B	1.0±0.2, 6.0±0.8	2.0±0.3, 5.0±0.7	2.0±0.3, 7.0±1.1	1.0±0.2, 4.0±0.6	1.0±0.2, 6.0±0.8	1.0±0.2, 5.0±0.6	2.0±0.3, 10.0±1.5	1.4±0.2, 6.1±0.9	2.0±0.4, 9.0±1.3
Llama-3 1B	1.0±0.2, 4.0±0.5	2.0±0.3, 5.0±0.6	1.0±0.2, 6.0±0.8	1.0±0.2, 5.0±0.6	1.0±0.2, 5.0±0.7	1.0±0.1, 4.0±0.5	2.0±0.3, 9.0±1.2	1.3±0.2, 5.4±0.7	2.0±0.3, 8.0±1.1
Llama-3 8B	2.0±0.3, 3.0±0.4	1.0±0.2, 4.0±0.5	2.0±0.3, 4.0±0.5	1.0±0.2, 3.0±0.4	1.0±0.2, 3.0±0.4	1.0±0.2, 3.0±0.4	2.0±0.3, 8.0±0.9	1.4±0.2, 4.0±0.5	2.0±0.3, 6.0±0.8
Mistral 7B	1.0±0.2, 4.0±0.5	2.0±0.3, 4.0±0.5	2.0±0.3, 5.0±0.6	1.0±0.2, 3.0±0.4	1.0±0.2, 4.0±0.5	1.0±0.2, 3.0±0.4	2.0±0.3, 8.0±1.0	1.4±0.2, 4.4±0.6	2.0±0.3, 6.0±0.8

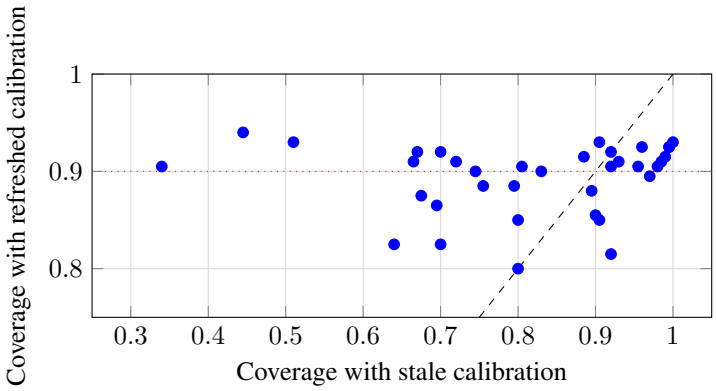


Figure 3: Stale versus refreshed coverage for Pythia-1.4B GLUE. Refreshed coverage recovers to a band around nominal 0.9.

The buffer-size ablation in Table 12 reports coverage loss as a function of m on the Mixed sequence for Llama-3 8B. The empirical decay is faster than the worst-case $1/\sqrt{m}$ rate of Theorem 8, closer to $m^{-0.7}$, because the bound does not exploit smoothness of the score CDF. On the same sequence at $m = 200$, pooled calibration achieves 5.6 ± 0.6 coverage-loss points, while per-task calibration replay achieves 1.8 ± 0.3, a gap consistent with Proposition 9.

Table 12: Coverage loss as a function of per-task buffer size m on the Mixed sequence for Llama-3 8B.

m per task	50	100	200	500	1000
Coverage loss (pts)	4.8 ± 0.5	2.9 ± 0.4	1.8 ± 0.3	0.9 ± 0.2	0.6 ± 0.2

Finally, Figure 4 shows ECE trajectories for Pythia-1.4B GLUE. Tasks with more label classes (MNLI) show larger ECE spikes, and ECE oscillation tracks the same instability pattern observed in coverage trajectories.

I Finite-Sample Pooled Calibration

Proposition 9 states the population-level limitation of pooled thresholds. The corresponding finite-sample analogue, referenced in Section 5, makes precise the empirical-CDF and discretization slack in pooled recalibration.

Proposition 10 (Finite-sample pooled calibration targets mixture coverage)**.** *Suppose the pooled baseline uses stratified buffers B_j of sizes m_j and weights $\pi_j = m_j / \sum_{\ell=1}^{k} m_\ell$, and defines the weighted empirical score CDF $\hat{F}_{\mathrm{pool}}(t) = \sum_{j=1}^{k} \pi_j \hat{F}_j(t)$, where $\hat{F}_j$ is the empirical score CDF on task j under θ_k. Let $\hat{q}_k^{\mathrm{pool}}$ be any data-dependent pooled threshold for which $\hat{F}_{\mathrm{pool}}(\hat{q}_k^{\mathrm{pool}}) \geq 1 - \alpha - \Delta_m$, where $\Delta_m \geq 0$ is the finite-sample discretization slack of the weighted quantile rule being used. If $\sup_t |\hat{F}_{\mathrm{pool}}(t) - F_{\theta_k}^{\bar{P}_k}(t)| \leq \varepsilon$, then the threshold has mixture coverage at least $1 - \alpha - \Delta_m - \varepsilon$, and for each task j, $F_{\theta_k}^{P_j}(\hat{q}_k^{\mathrm{pool}}) \geq 1 - \alpha - \Delta_m - \varepsilon - \mathrm{TV}(P_j, \bar{P}_k)$. Thus finite*

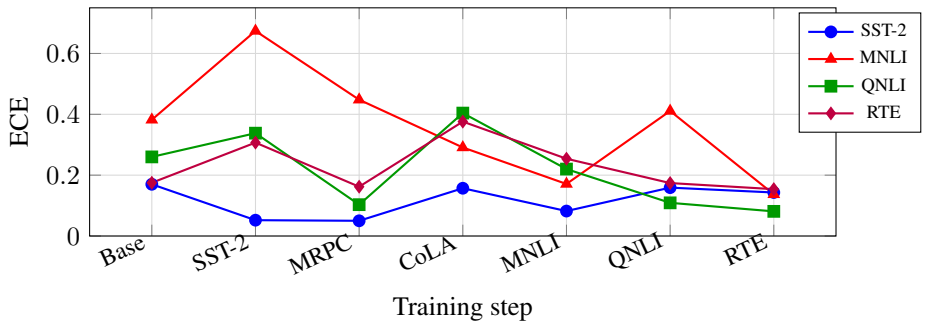


Figure 4: ECE trajectories for Pythia-1.4B GLUE. Tasks with more label classes (MNLI) show larger ECE spikes.

pooled recalibration targets mixture coverage up to empirical-CDF and discretization error, and inherits the same task-to-mixture penalty as the population threshold.

Proof. By the assumed empirical-CDF event, $F_{\theta_k}^{\bar{P}_k}(\hat{q}_k^{\mathrm{pool}}) \geq \hat{F}_{\mathrm{pool}}(\hat{q}_k^{\mathrm{pool}}) - \varepsilon \geq 1 - \alpha - \Delta_m - \varepsilon$. Letting $A = \{(x, y) : s_{\theta_k}(x, y) \leq \hat{q}_k^{\mathrm{pool}}\}$ and applying the total-variation argument from Proposition 9 to this event gives $P_j(A) \geq \bar{P}_k(A) - \mathrm{TV}(P_j, \bar{P}_k) \geq 1 - \alpha - \Delta_m - \varepsilon - \mathrm{TV}(P_j, \bar{P}_k)$, which is the claimed task-level lower bound. □

J Exploratory Generation Setting

Our formal framework covers classification-style tasks (including multiple-choice QA). To explore whether the qualitative phenomenon holds in open-ended generation, we created an eighth sequence: a general-to-code progression fine-tuning SQuAD followed by HumanEval and MBPP.

We sample $n_{\mathrm{gen}} = 32$ completions at temperature 0.8 and include the $(1 - \alpha)$-fraction lowest-score completions under length-normalized negative log-probability, reporting *surrogate coverage* defined as the fraction of test problems for which at least one completion in the set passes all unit tests. Surrogate coverage under a stale threshold drops qualitatively like classification: Llama-3 8B loses 25 surrogate-coverage points after fine-tuning, while calibration replay with $m = 200$ reduces the loss to 2 points.

Three caveats explicitly apply to this exploratory setting, which is why it is visually separated in Table 1 and Table 3 and excluded from the main classification means. First, the autoregressive sampling distribution is not exchangeable with the reference solution distribution, so Theorem 7 does not mathematically guarantee validity as stated. Second, surrogate coverage based on passing unit tests is a coarse proxy for true semantic generation correctness. Third, the calibration buffers act at the problem level rather than the token level, which may introduce variance when generations vary widely in length.

K Proofs

Proof of Lemma 1

The first inequality is the definition of the Kolmogorov distance applied at the single point q. For the second inequality,

$$(1 - \alpha) - F_{\theta_k}^{P_k}(q_1) \leq F_{\theta_1}^{P_1}(q_1) - F_{\theta_k}^{P_k}(q_1),$$

and the right-hand side is bounded by the first inequality. Taking the positive part gives the displayed bound.

PROOF OF PROPOSITION 3

Let $\eta = L_s d(\theta, \theta')$. The Lipschitz condition implies $\{s_\theta \le q - \eta\} \subseteq \{s_{\theta'} \le q\} \subseteq \{s_\theta \le q + \eta\}$, therefore $F_\theta^P(q-\eta) \le F_{\theta'}^P(q) \le F_\theta^P(q+\eta)$. The one-sided comparisons give $F_{\theta'}^P(q) - F_\theta^P(q) \le F_\theta^P(q+\eta) - F_\theta^P(q) \le M\eta$ and $F_\theta^P(q) - F_{\theta'}^P(q) \le F_\theta^P(q) - F_\theta^P(q-\eta) \le M\eta$. Combining them yields the claim.

PROOF OF PROPOSITION 4

The set $A_{\theta,q} = \{(x,y) : s_\theta(x,y) \le q\}$ is measurable, so $|F_\theta^P(q) - F_\theta^Q(q)| = |P(A_{\theta,q}) - Q(A_{\theta,q})| \le \mathrm{TV}(P,Q)$. The KL bound is Pinsker's inequality.

PROOF OF COROLLARY 5

By Lemma 1, it is enough to bound the pointwise CDF drift at q_1. Insert the intermediate pair (θ_k, P_1) and use the triangle inequality: $F_{\theta_1}^{P_1}(q_1) - F_{\theta_k}^{P_k}(q_1) \le |F_{\theta_1}^{P_1}(q_1) - F_{\theta_k}^{P_1}(q_1)| + |F_{\theta_k}^{P_1}(q_1) - F_{\theta_k}^{P_k}(q_1)|$. The first term is bounded by Proposition 3; the second is bounded by Proposition 4.

PROOF OF THEOREM 7

Conditional on θ_k, Assumption 6 implies that $S_1, \dots, S_{m_j}, S_{m_j+1}$ are exchangeable, where $S_{m_j+1} = s_{\theta_k}(X_{m_j+1}, Y_{m_j+1})$. If ties occur, break them by appending independent continuous auxiliary variables and ordering pairs lexicographically; the resulting prediction set is no smaller than the non-randomized set except on tie boundaries, so the conservative coverage conclusion is unchanged. The randomized rank of the test score among the $m_j + 1$ scores is uniform on $\{1, \dots, m_j + 1\}$, and the event $S_{m_j+1} \le \hat{q}_{j,k}$ contains the event that this rank is at most r_j, which therefore has probability at least $r_j/(m_j+1) \ge 1 - \alpha$. This event is exactly $Y_{m_j+1} \in C_{j,k}(X_{m_j+1})$.

PROOF OF THEOREM 8

The Dvoretzky–Kiefer–Wolfowitz inequality gives $\sup_t |\hat{F}_m(t) - F(t)| \le \epsilon_{m_j,\delta}$ with probability at least $1 - \delta$. Under continuity, ties occur with probability zero, and $\hat{F}_m(\hat{q}_{j,k}) = r_j/m_j$. Since $r_j = \lceil (m_j+1)\tau \rceil$, we have $|r_j/m_j - \tau| \le 2/m_j$, therefore $|F(\hat{q}_{j,k}) - \tau| \le |F(\hat{q}_{j,k}) - \hat{F}_m(\hat{q}_{j,k})| + |r_j/m_j - \tau| \le \epsilon_{m_j,\delta} + 2/m_j$. Because F is strictly increasing on I with density at least $c_{j,k}$, the inverse map on $F(I)$ is $1/c_{j,k}$-Lipschitz. Applying this inverse-Lipschitz bound to $F(\hat{q}_{j,k})$ and $F(q^\star) = \tau$ gives the quantile bound.

PROOF OF PROPOSITION 9

The mixture-coverage statement is the definition of q_k^{pool}. Let $A = \{(x,y) : s_{\theta_k}(x,y) \le q_k^{\mathrm{pool}}\}$, then $\bar{P}_k(A) \ge 1 - \alpha$, and by the definition of total variation, $P_j(A) \ge \bar{P}_k(A) - \mathrm{TV}(P_j, \bar{P}_k) \ge (1-\alpha) - \mathrm{TV}(P_j, \bar{P}_k)$.