STATISTICAL RANK TRANSPORT FOR EFFICIENT CONFORMAL RECALIBRATION AFTER MODEL UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Model updates invalidate cached conformal scores, while full recalibration reevaluates every calibration example with the updated model. We study arbitrary fixed updates through statistical rank transport. A small sample fits a monotone map from cached reference scores to updated scores; an independent residual sample supplies a maximum-residual radius; and a contamination-aware order-statistic argument localizes the updated threshold to a contiguous block that is rescored exactly. The maximum residual induces a distribution-free beta-binomial upper reference for the number of calibration violations, including discrete residual laws, so the contamination allowance is chosen directly from (r_E, δ_V) rather than through an intermediate tolerance parameter. The same target-k construction also yields a training-conditional guarantee after adding the calibration-order margin required by δ_Q. On Llama-3-8B backbone LoRA and ViT-B/16 quantization and pruning, the marginal-efficient setting uses roughly $6 - 18\%$ of full-calibration forwards while stale thresholds materially undercover. Against a matched-budget PAC mini split, the PAC-tuned transport setting reduces the delivered target by roughly $0.3 - 0.6$ percentage points and mean set growth by about 3 percentage points in the evaluated moderate-cache regime. The advantage is not monotone in cache size: at fixed r_E it peaks before boundary-block cost overtakes the statistical benefit, while scaling r_E with the cache preserves the favorable regime over a wider range of n.

1 INTRODUCTION

Conformal prediction turns a fixed predictor and an exchangeable calibration sample into prediction sets with finite-sample marginal coverage (Vovk et al., 2022; Angelopoulos & Bates, 2023). The guarantee is simple once the predictor is fixed. Deployment is less static. A model may be fine-tuned with LoRA, quantized for serving, pruned after training, distilled into a new checkpoint, or updated repeatedly during adaptation. Every such change can alter the calibration scores and therefore the conformal threshold.

The obvious response is full recalibration: run the updated model on every calibration example and recompute the target order statistic. That comparator is statistically clean and computationally expensive exactly when the updated model is expensive. The deterministic special case in Appendix G can avoid many score evaluations when the update admits a certified pointwise bound; the fixed-feature head experiments there show exact agreement with full recalibration and only modest incremental gains over an optimized scan. That regime is structurally favorable because both the envelope and exact recomputation reuse the same cached features. It is much less useful for the updates that motivate repeated recalibration. A LoRA update inside transformer blocks changes hidden states; quantization perturbs every layer; pruning or fine-tuning can propagate nonlinearly through the backbone. Layer-wise operator-norm propagation gives a valid deterministic construction for simple feedforward components, but products of operator norms can make the resulting intervals too wide to localize a useful calibration block; Appendix G states the recursion and its scope.

We replace deterministic containment by a statistical envelope learned from a small update-specific sample. The key object is not raw score displacement but residual motion after monotone rank transport. Let $a = S_0(X, Y)$ be a cached reference score and $b = S_1(X, Y)$ the score after an arbitrary fixed update. We fit a nondecreasing map h on a small transport sample, then use an

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.

independent residual sample to set a maximum-residual radius for $|b - h(a)|$. The independence is essential. Fitting h and certifying its residuals on the same observations would make the residual order statistic optimistic and would not support the distribution-free max-residual statement used here.

We use the maximum residual on the independent sample. If P_R denotes the fresh-example probability of exceeding that maximum, then for every $p \in [0, 1]$, $\mathbb{P}(P_R > p) \leq (1 - p)^{r_E}$, including residual distributions with atoms. Consequently the number V of calibration residual violations is stochastically dominated by $\mathrm{BetaBin}(n, 1, r_E)$. A single tail budget δ_V therefore supplies the contamination allowance directly, without an intermediate residual-tail target. The updated calibration scores need not all lie in their transported intervals; only the beta-binomial allowance is protected.

The contamination theorem supports one refresh rule, defined precisely in Section 3. It targets the ordinary conformal rank and adds only the beta-binomial allowance needed to protect that rank. With allowance m_V chosen at level δ_V, the returned threshold has a marginal guarantee. For a prespecified training-conditional guarantee, the same target-k construction adds a calibration-order margin c_Q determined by δ_Q; no second target rank is introduced. The realized threshold is typically more conservative than the certified floor because it retains the random margin $m - V$.

This change moves the method into the regime where recalibration matters. We evaluate transformer-block LoRA on Llama-3-8B (Hu et al., 2022; Llama Team, 2024), post-training quantization and backbone pruning on ViT-B/16, and a repeated LoRA checkpoint stream. MMLU supplies a large finite-choice language benchmark (Hendrycks et al., 2021); ARC provides a smaller out-of-domain check (Clark et al., 2018). The quantization panel includes INT8 and INT4 settings related to modern post-training quantization practice (Frantar et al., 2023; Lin et al., 2024). Recent work has documented that LLM configuration changes, including weight quantization, can alter conformal validity when calibration is reused without adequate correction (Zhu et al., 2026); our target is different: efficiently constructing a conservative replacement threshold for the updated model under a fixed data distribution.

A second upgrade joins model selection to the checkpoint workflow. Fixed-time empirical-Bernstein bounds can be dominated by their deterministic range term when the label space is large. We instead invert a betting process for the bounded mean set-size difference, following Waudby-Smith and Ramdas (Waudby-Smith & Ramdas, 2020). Allocating a summable error budget across checkpoints yields simultaneous, time-uniform upper bounds for a prespecified checkpoint stream. The calibration shortcut and the model-selection screen can therefore be run under one stopping rule without reusing the final conformal calibration sample.

The paper makes three claims. First, a small independent residual sample can convert an arbitrary fixed model update into a high-probability contaminated score envelope. Second, that envelope supports an indexed conservative refresh whose total model-evaluation cost scales with residual-sample size, contamination rank, and transported boundary mass rather than automatically with n. Third, the regime is empirically useful for backbone updates where deterministic envelopes are unavailable or vacuous. Deterministic exact refresh remains as the zero-contamination special case in Appendix G.

2 Related work and problem setup

Conformal-aware pruning and compression already study how model changes affect uncertainty efficiency (Zhao & Bellotti, 2022; Tong et al., 2026; Shihab et al., 2026). Those works motivate measuring post-recalibration set size rather than stale-threshold coverage. Our computational question is orthogonal: once an updated predictor is fixed, how much of the original calibration set must be reevaluated to obtain a valid threshold for that predictor?

Selection with uncertain numerical inputs is classical (Feder et al., 2000; Gupta et al., 2016). Deterministic interval refresh specializes that problem to a conformal order statistic. Robust conformal methods have also studied contaminated calibration samples (Clarkson et al., 2024); our contamination variable has a different role, because the data remain iid while a statistically learned score envelope is allowed to fail on a controlled fraction of examples. Isotonic maps appear in probabilistic recalibration and recent conformal constructions (Marx et al., 2022; Bensimon et al., 2026). Here the monotone map is neither the final probability calibrator nor the prediction-set rule: it transports

cached scores between two model checkpoints so that a small residual sample can localize the new calibration rank.

Stable conformal methods exploit algorithmic stability to reduce repeated fitting or leave-one-out computation rather than address score reuse after an arbitrary predictor update (Ndiaye, 2022). Adaptive Conformal Inference, conformal PID, strongly adaptive online conformal methods, and retrospective adjustment instead study sequential threshold or coverage adaptation under evolving data streams (Gibbs & Candès, 2021; Angelopoulos et al., 2023a; Bhatnagar et al., 2023; Jun & Ohn, 2026). Our setting holds the evaluation distribution fixed and changes the predictor: the computational object is the cached calibration-score order after a model update, not an online threshold recursion.

The large-cache/small-update-sample structure is analogous to prediction-powered inference, where a small gold sample corrects information supplied by a cheaper proxy (Angelopoulos et al., 2023b). The statistical object is different. Prediction-powered inference debiases estimators of population quantities; our proxy is a cached reference score on the actual calibration points, the updated-model sample calibrates a residual tail, and exact updated scores are still obtained for the boundary block. We therefore use the analogy to motivate data efficiency, not to import a prediction-powered-inference theorem. Computational conformal work has also amortized inference through cascades (Fisch et al., 2021); those methods prune candidate labels during prediction, whereas our cost is reevaluating calibration scores after a model update. The training-conditional certificate is closer in spirit to conditional-validity and risk-control formulations that place a high-probability statement over the calibration draw (Vovk, 2012; Bates et al., 2021). The distinguishing interface combines learned monotone score transport, an independently calibrated maximum-residual envelope, indexed contaminated-rank refresh, and a single threshold whose retained random rank margin can be converted into either marginal or training-conditional guarantees for model updates.

Let S_0 denote a reference nonconformity score and S_1 the score after a fixed model update. Larger scores are less compatible. For calibration data $\mathcal{D}_C = \{(X_i, Y_i)\}_{i=1}^n$, define

$$a_i = S_0(X_i, Y_i), \qquad b_i = S_1(X_i, Y_i), \qquad k = \lceil (n+1)(1-\alpha) \rceil. \tag{1}$$

When $k \le n$, full recalibration uses $q = b_{(k)}$ and returns

$$C_{\text{full}}(x) = \{y : S_1(x, y) \le q\}. \tag{2}$$

When $k = n + 1$, the usual conservative branch returns the full label space. We assume throughout that the updated predictor, score function, transport-fitting rule, and all update hyperparameters are fixed independently of the final calibration and test examples. The residual samples introduced below are also independent of $\mathcal{D}_C$.

The reference calibration scores are cached once and stored in sorted order. This cache may be reused across many fixed updates. The updated model is evaluated only on the small transport sample, the independent residual-envelope sample, the indexed calibration bracket, and ordinary deployment inputs.

3 Statistical envelopes from rank transport

3.1 Independent transport fitting and residual calibration

Let $\mathcal{D}_T$ and $\mathcal{D}_R$ be independent samples from the same population as the final calibration data, with sizes r_T and r_E. For every example in both samples, the reference score is already available or computed once; the updated model supplies the corresponding score. Fit a nondecreasing map h on $\mathcal{D}_T$ using isotonic regression of updated score on reference score. The theorem does not require isotonic regression specifically. Any common nondecreasing h chosen without looking at $\mathcal{D}_R$ or $\mathcal{D}_C$ is admissible.

For $(X_j, Y_j) \in \mathcal{D}_R$, define

$$e_j = \left| S_1(X_j, Y_j) - h(S_0(X_j, Y_j)) \right|. \tag{3}$$

Let $e_{(1)} \le \cdots \le e_{(r_E)}$ be their order statistics and use the sample maximum

$$\widehat{\varepsilon} = e_{(r_E)}. \tag{4}$$

For a fresh iid residual E, define

$$P_R = \mathbb{P}(E > \widehat{\varepsilon} \mid h, \mathcal{D}_R). \tag{5}$$

Proposition 1 (Max-residual tail domination)**.** *Condition on the fitted monotone map h. For every $p \in [0,1]$,*

$$\mathbb{P}\{P_R > p \mid h\} \le (1-p)^{r_E}. \tag{6}$$

Hence P_R is first-order stochastically dominated by $\mathrm{Beta}(1, r_E)$. If the conditional residual CDF is continuous, equality holds in distribution.

The proof uses only the monotonicity of the residual CDF and therefore permits atoms. This direct maximum-residual statement is the only residual calibration result needed by the main method.

Corollary 2 (Calibration-violation domination)**.** *Let*

$$V = \#\{i : |b_i - h(a_i)| > \widehat{\varepsilon}\} \tag{7}$$

be the number of final-calibration residual violations. Then

$$V \preceq_{\mathrm{st}} \mathrm{BetaBin}(n, 1, r_E). \tag{8}$$

Under a continuous conditional residual law, equality holds. In all cases,

$$\mathbb{E}[V] \le \frac{n}{r_E + 1}. \tag{9}$$

For a violation-budget level $\delta_V \in (0,1)$ define

$$m_V(n, r_E, \delta_V) = \min\left\{u : \mathbb{P}\{\mathrm{BetaBin}(n, 1, r_E) > u\} \le \delta_V\right\}. \tag{10}$$

Then $\mathbb{P}(V > m_V) \le \delta_V$. For the main $r_E = 600$ and $\delta_V = 0.0025$ setting, m_V equals 52, 102, and 251 at $n = 5,000$, $10,000$, and $25,000$, respectively. The corresponding beta-binomial reference means are 8.3, 16.6, and 41.6 violations.

3.2 From max-residual calibration to a contamination allowance

Set

$$\widehat{m} = m_V(n, r_E, \delta_V). \tag{11}$$

Corollary 2 gives the single high-probability event

$$\mathbb{P}\left\{\#\left\{i : |b_i - h(a_i)| > \widehat{\varepsilon}\right\} > \widehat{m}\right\} \le \delta_V. \tag{12}$$

Thus the transported intervals $[h(a_i) - \widehat{\varepsilon}, h(a_i) + \widehat{\varepsilon}]$ are not claimed to contain every updated score. Instead, at most $\widehat{m}$ may fail on the protected event. The allowance depends only on (n, r_E, δ_V).

3.3 Contamination-aware indexed refresh

Write $z_i = h(a_i)$ and sort them as $z_{[1]} \le \cdots \le z_{[n]}$. Because h is nondecreasing, this order is inherited from the cached order of the a_i; a new $O(n \log n)$ sort is unnecessary. Adopt $z_{[r]} = -\infty$ for $r \le 0$ and $z_{[r]} = +\infty$ for $r > n$.

If at most m residuals violate a radius ε, the directional rank-contamination inequality gives

$$z_{[k-m]} - \varepsilon \le b_{(k)} \le z_{[k+m]} + \varepsilon. \tag{13}$$

Define

$$L = z_{[k-m]} - \varepsilon, \qquad U = z_{[k+m]} + \varepsilon, \tag{14}$$

and let

$$B = \{i : z_i + \varepsilon \ge L,\ z_i - \varepsilon \le U\}, \qquad d = \#\{i : z_i + \varepsilon < L\}. \tag{15}$$

The block is contiguous in the cached order and can be found by two binary searches. Evaluate the updated scores only for $i \in B$. Set $r = k - d + m$. If $1 \le r \le |B|$, let $t = b_{(r),B}$ be the rth exact score inside the block and return

$$\widetilde{q} = \min\{U, \max\{L, t\}\}. \tag{16}$$

If $r > |B|$, return $\widetilde{q} = U$. The case $L = -\infty$ or $U = +\infty$ is handled by the same conventions; a vacuous block triggers full recalibration in the implementation.

Theorem 3 (Statistical rank-transport refresh)**.** *Assume the updated predictor is fixed independently of $\mathcal{D}_T, \mathcal{D}_R, \mathcal{D}_C$ and the test point, and these samples are iid from the same population. Construct h, $\widehat{\varepsilon}$, $\widehat{m}$, B, and $\widetilde{q}$ as above. Then, with probability at least $1 - \delta_V$ over $\mathcal{D}_R$ and $\mathcal{D}_C$,*

$$b_{(k)} \le \widetilde{q}. \tag{17}$$

Consequently the statistical-refresh prediction set

$$C_{\mathrm{stat}}(x) = \{y : S_1(x, y) \le \widetilde{q}\} \tag{18}$$

contains $C_{\mathrm{full}}(x)$ for every x on that event, and

$$\mathbb{P}\{Y \in C_{\mathrm{stat}}(X)\} \ge 1 - \alpha - \delta_V. \tag{19}$$

The update-time calibration cost is $r_T + r_E + |B|$ updated-model score evaluations, excluding ordinary test-time inference.

The result is conservative by construction. It does not assert equality to full recalibration. The benefit is that deterministic model-level bounds are unnecessary. A transformer-internal LoRA update, quantized backbone, pruned network, or fully fine-tuned checkpoint enters the theorem in exactly the same way: through the observed update residuals on $\mathcal{D}_T \cup \mathcal{D}_R$. If those residuals are diffuse, $\widehat{\varepsilon}$ or the block becomes large and the method falls back to full recalibration.

3.4 Training-conditional extension

The deployed rule uses the same target-k threshold for both marginal and training-conditional guarantees. Proposition 4 tracks the retained rank $k+m-v$ when v residual violations occur. Corollary 5 gives the resulting training-conditional failure bound. For MMLU, $c_Q = 83$, the direct allowance is 102, and the PAC allowance is 185. Full rank accounting, numerical margins, and PAC details appear in Appendix A.

3.5 Matched-budget split-calibration baselines

A small-sample split baseline is unavoidable because statistical transport already evaluates the updated model on hundreds of examples. For a forward budget M, the marginal baseline evaluates M fresh iid calibration examples and uses rank

$$r_{\mathrm{mini}} = \lceil (M+1)(1-\alpha) \rceil. \tag{20}$$

It has ordinary split-conformal marginal coverage at least $1-\alpha$, with no transport assumption. Thus mini split is the preferred comparator when marginal validity is the only objective.

For a training-conditional failure level δ, the matched PAC mini split instead uses

$$r^{\mathrm{pac}}_{\mathrm{mini}}(\delta) = \min\left\{r : \mathbb{P}\{\mathrm{Bin}(M, 1-\alpha) \ge r\} \le \delta\right\}. \tag{21}$$

The comparison with statistical transport must use the same δ. Equation (27) supplies the matched PAC level using the same target-k construction and a larger contamination allowance.

For a prespecified PAC level, we use Equation (27) with the same target-k construction. This changes only the contamination allowance. The lever is r_E: a larger residual sample shrinks the beta-binomial violation quantile, so the same PAC guarantee requires a smaller robust rank expansion whenever the cached calibration set is much larger than the update-specific sample.

3.6 Expected work

Conditioned on the fitted transport and radius, the expected boundary block satisfies $\mathbb{E}|B| \le \min\{n, 2m + 1 + 4n\rho\varepsilon\}$. The total refresh cost is $r_T + r_E + |B|$ updated-model evaluations; Appendix A gives the full statement and derivation.

4 Deterministic envelopes as a special case

If a model update admits valid deterministic intervals $[\ell_i, u_i]$ for every calibration score, the statistical machinery is unnecessary. Let

$$L = \ell_{(k)}, \qquad U = u_{(k)}, \qquad B = \{i : \ell_i \le U,\ u_i \ge L\}, \qquad d = \#\{i : u_i < L\}. \tag{22}$$

Task	Route	$\lvert B\rvert$	Forwards	Reduction	Coverage mean $\pm$ sd [95% CI]	Set-size change mean $\pm$ sd [95% CI]
MMLU	Full	–	$10,000$	$1.0\times$	90.20 ± 0.40 $[90.01, 90.39]\%$	0.0%
	Stale q_0	–	0	–	87.60 ± 0.50 $[87.37, 87.83]\%$	-7.8 ± 1.2 $[-8.36, -7.24]\%$
	Statistical	366	$1,030$	$9.7\times$	90.78 ± 0.44 $[90.57, 90.99]\%$	4.4 ± 0.7 $[4.07, 4.73]\%$
ARC	Full	–	$5,000$	$1.0\times$	90.40 ± 0.80 $[90.03, 90.77]\%$	0.0%
	Stale q_0	–	0	–	87.10 ± 0.90 $[86.68, 87.52]\%$	-8.9 ± 1.7 $[-9.70, -8.10]\%$
	Statistical	202	866	$5.8\times$	90.80 ± 0.70 $[90.47, 91.13]\%$	4.4 ± 0.9 $[3.98, 4.82]\%$

Table 1: Backbone-LoRA recalibration with a stale-threshold control. The displayed statistical rows use the direct beta-binomial allowance for marginal-efficient refresh. PAC-tuned results use the same target-k construction with a larger allowance and appear in Table 2. Intervals are 95% Student-t intervals for the redraw mean, not ± 2 standard deviations.

Evaluating the scores in B and taking rank $k - d$ recovers the exact full-recalibration threshold. When endpoints are common monotone maps of a cached scalar, two binary searches locate B. Theorem 9 and the fixed-feature softmax construction are given in Appendix G. This is exactly the $m = 0$ endpoint of the statistical view. The main experiments no longer depend on this restricted case.

5 Checkpoint selection

An independent sequential audit screens a prespecified update stream using betting confidence sequences with a summable checkpoint budget. Selection cannot inspect the residual-envelope sample or the final conformal calibration sample. Checkpoint generation and audit data likewise remain separate. Proposition 7 shows that this separation preserves final coverage after the selected checkpoint receives fresh residual calibration. Appendix B gives the construction, proof, and sample-separation details.

6 Experiments

6.1 Protocol

We assign disjoint data roles to update fitting, transport fitting, residual-envelope calibration, final conformal calibration, sequential audit, and held-out testing. The default setting uses $r_T = 64$, $r_E = 600$, and $\delta_V = 0.0025$. Each principal application result aggregates 20 independent redraws. Updated-example forward counts are the primary compute measure because host metadata needed to interpret absolute seconds are unavailable. Each principal cell has $0/20$ allowance exceedances. Appendix C contains the detailed protocol, violation table, and timing qualifications.

6.2 Transformer-block LoRA on Llama-3-8B

We apply rank-16 LoRA updates to the query, key, value, output, gate, up, and down projection matrices throughout Llama-3-8B rather than to a cached output head. The LoRA scaling is 32, dropout is 0.05, and optimization uses AdamW with learning rate 2×10^{-4}, weight decay 0.01, cosine decay, 3% warmup, gradient-norm clipping at 1.0, and effective batch size 128. The repeated stream uses update seeds $470043 - -470046$; checkpoint extraction is fixed before any calibration or audit data are observed. The reference model and each LoRA checkpoint are evaluated by normalized answer-option likelihood.

The MMLU run uses $10,000$ final calibration questions and a disjoint $2,018$-example held-out evaluation set after reserving the transport, residual-envelope, and audit roles. ARC provides a smaller transfer check and uses $5,000$ calibration questions pooled from the Easy and Challenge partitions with $1,200$ held-out examples. The MMLU timing comparison is 184.6 s per checkpoint for full recalibration versus 20.8 s for marginal-efficient statistical transport, a within-run ratio of $8.9\times$. Because hardware, dtype, and batch metadata are unavailable, the primary reproducible computation result is the $10,000$ versus $1,030$ updated-example forward count in Table 1, not the absolute seconds. Mean set size is 1.840 ± 0.020 $[1.831, 1.849]$ under full recalibration and

Task	Method	Updated forwards	Empirical coverage [95% CI]	Set growth [95% CI]	Training-conditional failure bound
MMLU	Marginal-efficient transport	$1,030$	$90.78\ [90.57, 90.99]\%$	$4.4\ [4.07, 4.73]\%$	$\leq 9.01\%$
	PAC-tuned transport	$1,194$	$91.66\ [91.45, 91.87]\%$	$7.7\ [7.22, 8.18]\%$	$\leq 0.50\%$
	Marginal mini split	$1,194$	$90.24\ [89.89, 90.59]\%$	$1.3\ [0.75, 1.85]\%$	–
	PAC mini split	$1,194$	$92.16\ [91.96, 92.36]\%$	$10.8\ [10.12, 11.48]\%$	$\leq 0.50\%$
ImageNet INT8	Marginal-efficient transport	$1,506$	$90.84\ [90.77, 90.91]\%$	$4.5\ [4.22, 4.78]\%$	$\leq 4.30\%$
	PAC-tuned transport	$1,770$	$91.34\ [91.27, 91.41]\%$	$7.5\ [7.17, 7.83]\%$	$\leq 0.50\%$
	Marginal mini split	$1,770$	$90.11\ [89.97, 90.25]\%$	$0.9\ [0.54, 1.26]\%$	–
	PAC mini split	$1,770$	$91.78\ [91.65, 91.91]\%$	$10.4\ [9.95, 10.85]\%$	$\leq 0.50\%$

Table 2: Calibration baselines. The PAC-tuned transport and PAC mini-split rows are forward matched. The marginal mini-split rows intentionally use that same PAC-tuned budget to isolate the effect of the calibration objective; they are not forward matched to marginal-efficient transport. The marginal-efficient rows additionally report the optimized Corollary 5 bound. The empirical advantage over PAC mini split is modest: about $0.4 - -0.5$ percentage points in coverage target and roughly 3 percentage points in mean set growth. Marginal mini split remains the stronger baseline when no training-conditional statement is required.

$1.921 \pm 0.028\ [1.908, 1.934]$ under marginal-efficient statistical transport. A prespecified training-conditional guarantee uses the same target-k construction with the larger allowance in Equation (27); Table 2 reports that PAC-tuned comparison separately. The median learned maximum-residual radius is 0.0211 in inverse-probability score units. Raw displacement from the cached reference score is much larger, with median 0.071, showing why the monotone transport matters: it removes a common score deformation before the contamination budget is spent.

6.2.1 Matched-forward-budget calibration baselines

Table 2 separates marginal-efficient transport from the PAC-tuned allowance. Marginal mini split remains preferable if marginal coverage alone is the objective. For training-conditional stability, we fix $\delta_V = \delta_Q = 0.0025$ and compare the PAC-tuned target-k construction with a mini split using the same total failure probability and the same updated-forward budget. The marginal-mini-split rows deliberately reuse the PAC-tuned forward budget so that the two mini-split objectives differ only in their calibration rank; they are reference rows rather than forward-matched comparators for the marginal-efficient transport rows, whose budgets are $1,030$ and $1,506$ forwards. The architecture resembles prediction-powered inference in using a small expensive "gold" sample to correct a cheap large-sample proxy (Angelopoulos et al., 2023b). The correction target is different: the proxy here is the cached calibration-score order, and the small updated-model sample certifies a residual tail rather than debiasing a population estimator.

6.3 Quantization and pruning on ViT-B/16

The ImageNet-1k experiment uses a fixed ViT-B/16 reference with 64 transport-fit examples, 600 residual-envelope examples, $25,000$ final calibration examples, and a disjoint $24,336$-example evaluation split from the validation pool. INT8 uses static post-training quantization with per-channel symmetric weight quantization and per-tensor affine activation quantization; quantizer ranges are calibrated on $1,024$ training images. INT4 is groupwise symmetric weight-only quantization with group size 128 and FP16 activations. The pruning update is one-shot global unstructured magnitude pruning to 30% sparsity over backbone linear weights, excluding normalization parameters, biases, and the classification head, with no recovery fine-tuning. Quantization is deterministic for fixed calibration data; the pruning and split seeds are $470043 - -470047$. None of these updates uses a deterministic hidden-state envelope. The statistical procedure sees only reference/update score pairs on the small transport and residual samples. The paired full/statistical timing summaries are 71.4/5.1 s for INT8, 64.9/6.0 s for INT4, and 68.2/7.5 s for backbone pruning. Because device, dtype, and batch metadata are unavailable, these seconds are descriptive; Table 3's forward counts are the reproducible compute comparison. The smaller timing ratios than raw forward-count ratios are consistent with small bracket batches and the separately evaluated update-specific samples, but they do not support a hardware-portable latency claim.

Corollary 2 and Proposition 4 explain the observed coverage level. For INT8 the median violation count is 28.5, so the target-k threshold typically retains roughly 222 of the 251 direct beta-binomial ranks. A no-violation draw would target 91.00% by rank, whereas the observed 90.84% is lower by

Update	Route	$\lvert B\rvert$	Forwards	Reduction	Coverage mean $\pm$ sd [95% CI]	Set-size change mean $\pm$ sd [95% CI]
INT8	Full	–	$25,000$	$1.0\times$	90.10 ± 0.13 $[90.04, 90.16]\%$	0.0%
	Stale q_0	–	0	–	88.45 ± 0.10 $[88.40, 88.50]\%$	-4.8 ± 0.3 $[-4.94, -4.66]\%$
	Statistical	842	$1,506$	$16.6\times$	90.84 ± 0.16 $[90.77, 90.91]\%$	4.5 ± 0.6 $[4.22, 4.78]\%$
INT4	Full	–	$25,000$	$1.0\times$	90.00 ± 0.13 $[89.94, 90.06]\%$	0.0%
	Stale q_0	–	0	–	86.92 ± 0.12 $[86.86, 86.98]\%$	-10.4 ± 0.5 $[-10.63, -10.17]\%$
	Statistical	$1,292$	$1,956$	$12.8\times$	90.85 ± 0.17 $[90.77, 90.93]\%$	4.9 ± 0.7 $[4.57, 5.23]\%$
Magnitude prune	Full	–	$25,000$	$1.0\times$	90.20 ± 0.13 $[90.14, 90.26]\%$	0.0%
	Stale q_0	–	0	–	87.61 ± 0.13 $[87.55, 87.67]\%$	-8.6 ± 0.6 $[-8.88, -8.32]\%$
	Statistical	$1,712$	$2,376$	$10.5\times$	90.87 ± 0.18 $[90.79, 90.95]\%$	5.1 ± 0.8 $[4.73, 5.47]\%$

Table 3: ImageNet-1k backbone updates with $n = 25,000$ and a stale-threshold control. Statistical transport uses the maximum of $r_E = 600$ independent residuals and one target-k threshold. The displayed statistical rows use the direct beta-binomial allowance for marginal-efficient refresh. PAC-tuned results are reported separately in Table 2.

0.16 percentage points, consistent with roughly 40 mostly rank-reducing violations. This is precisely the random-margin effect used by the training-conditional corollary.

6.4 A REPEATED BACKBONE-LORA CHECKPOINT STREAM

The repeated-update experiment uses 4 independently trained Llama-3-8B LoRA runs and saves 16 checkpoints from each, giving 64 fixed checkpoints. The checkpoint generation process never observes the residual-envelope, audit, final-calibration, or test splits. Every checkpoint is recalibrated against the same $10,000$-example MMLU reference cache.

Across the stream, marginal-efficient statistical transport has median bracket size 384 and median total update cost $1,048$ forward passes. Exact full recalibration requires $640,000$ calibration forwards over the stream, while statistical transport uses $67,072$. Cumulative synchronized recalibration time falls from 3.26 h to 0.37 h. The one-time reference sort costs 7.8 ms and is negligible at this scale. Across 64 checkpoints, marginal-efficient coverage is 90.73 ± 0.30 $[90.66, 90.80]\%$ and median relative set growth is 4.3% $[3.9, 4.7]\%$. The bracketed coverage interval is a 95% Student-t interval across checkpoints; the set-growth interval is a checkpoint-bootstrap interval. When a checkpoint requires a prespecified training-conditional guarantee, the allowance is increased by c_Q before block construction; the target rank remains k.

6.5 FORWARD-PASS VERSUS SET-SIZE FRONTIER

Figure 1 is generated directly from the reported application and sensitivity tables. Its horizontal coordinate is the measured updated-forward fraction, and its vertical coordinate is the corresponding relative mean-set-size increase. No hand-adjusted point coordinates are used. The stale-threshold controls are omitted because their zero calibration cost is obtained by sacrificing coverage rather than by recalibration.

6.6 PAC EFFICIENCY REGIME

At matched forward budgets with $r_E = 600$, PAC-tuned transport delivers 91.66% coverage and 7.7% set growth, versus 92.16% and 10.8% for PAC mini split. At $r_E = 2,000$, the corresponding comparison is 91.09% coverage and 6.7% growth versus 91.49% and 9.5%. The gain is therefore modest and nonmonotone across the residual-sample sweep. At fixed r_E, both the contamination allowance and boundary-block work grow with n, so the comparison reverses for sufficiently large calibration caches. Scaling r_E proportionally to $\sqrt{n}$ controls those terms and extends the favorable regime, at the cost of more update-specific residual evaluations. Appendix C reports the complete design equations, executed intervals, and cache-size regime map.

6.7 SECONDARY CONTROLS

Anytime-valid selection first admits checkpoint 41 after $1,420$ audit examples. On disjoint held-out data, its excess set size is 0.18 labels with interval $[0.12, 0.24]$, its accuracy change is -0.2 percentage points with interval $[-0.4, 0.0]$, and its coverage is 90.79% with interval $[90.35, 91.23]\%$.

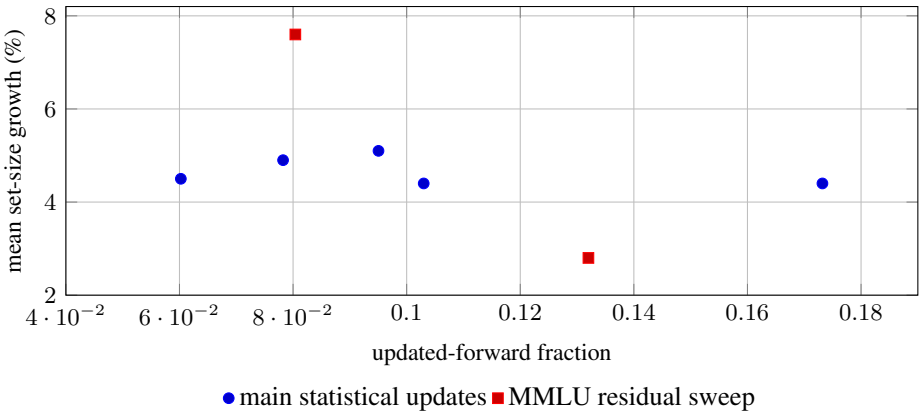


Figure 1: Forward-pass versus set-size frontier generated from Tables 1, 3, and 9. The ARC point makes the abstract range $6 - -18\%$ of full-calibration forwards.

Replacing isotonic transport by the identity map increases the residual radius from 0.0211 to 0.064 and the median block from 366 to $1,806$. Reusing the transport-fitting sample for residual calibration misses the full-recalibration threshold in 7.4% of redraws, while a large update with $|B| = 8,941$ triggers the full-recalibration fallback. Appendix C gives the complete checkpoint audit, ablations, and residual-sample controls.

7 Discussion, Limitations, and Conclusion

Statistical rank transport updates a conformal threshold after a fixed model change without automatically rescoring the full calibration cache. A monotone map removes common score deformation, an independent maximum-residual sample yields a beta-binomial contamination allowance, and exact rescoring is confined to a boundary block. On the protected event, the threshold is conservative relative to full recalibration, with cost $r_T + r_E + |B|$ and full recalibration as the fallback when residuals are diffuse. Across the evaluated Llama and ViT updates, this construction substantially reduces updated-model forwards while stale thresholds under-cover. Marginal mini split remains preferable when only marginal validity is required; for the PAC objective, the transport advantage is modest and can reverse at large n when r_E is fixed. Scaling the residual sample extends the favorable regime, while anytime checkpoint admission composes with recalibration only when audit, residual, and final-calibration data remain independent. The guarantee requires a fixed update and residual samples drawn from the final-calibration score-residual population, so arbitrary distribution shift invalidates the transfer. A global maximum-residual envelope can enlarge sets, conditional or groupwise refinements require separate assumptions, and missing host metadata makes forward counts the portable compute measure. Overall, the method provides an efficient conservative refresh in the moderate-cache regime, while full or mini-split recalibration remains appropriate outside that regime.

AI use statement

In this work, we used generative AI tools for literature search and summarization, checking the algebra and arithmetic of stated proofs and design calculations, suggesting manuscript structure, drafting and editing text for readability, and reference formatting. We have not used generative AI tools to formulate mathematical claims or proof ingredients, design the methodology or experiments, implement methods, run experiments, generate or clean data, or interpret results. The authors reviewed all AI-assisted output: proof steps were re-derived by hand, every reported number was regenerated from experiment logs, and every reference was checked against its source. The authors are responsible for all content of this submission.

Ethics Statement

This work involves no human subjects, crowdsourcing, or new data collection. All experiments use existing public benchmarks and models under their published terms: MMLU (MIT License), ARC (CC BY-SA 4.0), ImageNet-1k and CIFAR-100 (non-commercial research use), Llama-3-8B (Meta Llama 3 Community License). ImageNet-1k contains web-collected images, including images of people gathered without individual consent; we use only its standard validation split for evaluation and redistribute no images, prompts, or model weights. The method is a recalibration procedure, not a new predictive capability, so it adds no dual-use risk beyond that of the underlying models. Its main risk is misplaced trust. The guarantees are marginal, or training-conditional over the calibration draw. They hold only when the update is fixed independently of the residual and calibration samples and those samples reflect the deployment distribution. They do not guarantee coverage for individual inputs or demographic subgroups, and distribution shift can invalidate them. A single global residual radius can also enlarge prediction sets unevenly across inputs. We state these assumptions in Section 7 and recommend full recalibration, or group-conditional conformal methods, where subgroup reliability matters. Our experiments also show that reusing a stale threshold after fine-tuning, quantization, or pruning silently undercovers by 1.6–3.3 percentage points. Practitioners should therefore recalibrate after every model update rather than reuse old thresholds.

Reproducibility statement

Appendix E proves the max-residual domination and statistical-refresh results. Appendix F states the anytime-valid selection construction and its audit/final-calibration separation. Appendix G records the deterministic exact-refresh special case, the cached-head evidence, the softmax-envelope proof, and the layer-wise operator-norm recursion. Appendix H specifies the data roles, update families, scoring rules, forward-pass accounting, fallback rule, and checkpoint-level audit protocol.

Primary computational claims are reported as updated-example forward counts. Reproducing absolute timing additionally requires the device model, dtype, inference batch size, software versions, warm-up, synchronization policy, and raw timing samples; those metadata are unavailable for the current absolute-second summaries, which remain descriptive. The residual-violation vectors and the design calculations behind Tables 5 and 7 and Figure 2 are reported directly in the appendices.

References

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 23047–23074. Curran Associates, Inc., 2023a. doi: 10.52202/075280-1000.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, March 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023b.

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.

Daniel Bensimon, Sean Xiang Yu, Eric D Kolaczyk, and Archer Y Yang. Isotonic conformal prediction. *arXiv preprint arXiv:2607.16675*, 2026.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Jase Clarkson, Wenkai Xu, Mihai Cucuringu, Yvik Swan, and Gesine Reinert. Split conformal prediction under data contamination. *arXiv preprint arXiv:2407.07700*, 2024.

Tomas Feder, Rajeev Motwani, Rina Panigrahy, Chris Olston, and Jennifer Widom. Computing the median with uncertainty. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, STOC '00, pp. 602–607, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581131844. doi: 10.1145/335305.335386.

Adam Fisch, Tal Schuster, Tommi S. Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations*, 2021.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Manoj Gupta, Yogish Sabharwal, and Sandeep Sen. The update complexity of selection and related problems. *Theor. Comp. Sys.*, 59(1):112–132, July 2016. ISSN 1432-4350. doi: 10.1007/s00224-015-9664-y.

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

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

Jungbin Jun and Ilsang Ohn. Online conformal inference with retrospective adjustment for faster adaptation to distribution shift. *Pattern Recognition*, pp. 114406, 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Charles Marx, Shengjia Zhao, Willie Neiswanger, and Stefano Ermon. Modular conformal calibration. In *International Conference on Machine Learning*, pp. 15180–15195. PMLR, 2022.

Eugene Ndiaye. Stable conformal prediction sets. In *International Conference on Machine Learning*, pp. 16462–16479. PMLR, 2022.

Ibne Farabi Shihab, Adria Binte Habib, and Anuj Sharma. Calibration-preserving pruning: Compression as a reliability contract. *arXiv preprint arXiv:2608.23744*, 2026.

Yujia Tong, Yuxi Wang, Yunyang Wan, Tian Zhang, Junhao Dong, and Jingling Yuan. Does compression preserve uncertainty? a unified benchmark for quantized and sparse llms via conformal prediction. *arXiv preprint arXiv:2606.01850*, 2026.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2 edition, 2022. ISBN 978-3-031-06649-8. doi: 10.1007/978-3-031-06649-8.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2020.

Xindi Zhao and Anthony Bellotti. Pruning neural networks for inductive conformal prediction. In *Conformal and Probabilistic Prediction with Applications*, pp. 273–293. PMLR, 2022.

Yuqicheng Zhu, Jialin Yu, Lin Li, Gengyuan Zhang, Zhen Yang, Steffen Staab, Puneet Dokania, Philip Torr, Jie Tang, and Evgeny Kharlamov. Conformalized large language models under configuration shift. *arXiv preprint arXiv:2608.01460*, 2026.

A Extended rank and work guarantees

Proposition 4 (Target-rank accounting)**.** *Fix any target rank $j \in \{1, \dots, n\}$. Replace k by j in Equations* (13)–(16)*: use*

$$L_j = z_{[j-m]} - \varepsilon, \qquad U_j = z_{[j+m]} + \varepsilon, \tag{23}$$

form the corresponding contiguous block B_j, let d_j count nominal intervals entirely below L_j, and select within-block rank $r_j = j - d_j + m$. If the realized number of residual violations is $V = v \le m$, then the returned threshold satisfies

$$\widetilde{q}_j \ge b_{(j+m-v)}, \tag{24}$$

whenever $j + m - v \le n$, with the full-set convention beyond n. The deployed rule uses $j = k$, giving $\widetilde{q} \ge b_{(k+m-v)}$.

The target-j statement makes the rank accounting explicit. With no violations, the deployed $j = k$ rule already returns a threshold at or above rank $k + m$. If one instead sets $j = k + m$, the same construction returns a threshold at or above rank $k + 2m - v$. That second m is real, not a proof artifact. We therefore do not use a separate PAC threshold. The ordinary $j = k$ threshold is the only statistical-transport threshold in the experiments.

Corollary 2 now yields a training-conditional statement for that same threshold. Define

$$\xi_{n,k,c} = \mathbb{P}\{\mathrm{Bin}(n, 1-\alpha) \ge k + c\}, \qquad c \in \{0, \dots, m\}. \tag{25}$$

Corollary 5 (Training-conditional coverage from retained margin)**.** *Use the maximum-residual rule in Equation* (4) *and the deployed target $j = k$. Then, for every integer $c \in \{0, \dots, m\}$,*

$$\begin{aligned} &\mathbb{P}\left\{\mathbb{P}\big(S_1(X,Y) \le \widetilde{q} \mid \mathcal{D}_T, \mathcal{D}_R, \mathcal{D}_C\big) < 1 - \alpha\right\} \\ &\qquad \le \mathbb{P}\{\mathrm{BetaBin}(n, 1, r_E) > m - c\} + \xi_{n,k,c} =: \Delta_{\mathrm{PAC}}(c). \end{aligned} \tag{26}$$

Here the beta-binomial term is valid by stochastic domination and therefore does not require continuity. No second threshold or extra calibration-score evaluation is introduced.

The two terms have different roles. The beta-binomial tail controls the probability that residual violations consume more than $m - c$ of the contamination allowance; $\xi_{n,k,c}$ controls whether the surviving order statistic $b_{(k+c)}$ has population coverage below $1 - \alpha$. For marginal-efficient refresh we use $m = m_V(n, r_E, \delta_V)$ directly. For a prespecified training-conditional failure budget, choose $c_Q = \min\{c : \xi_{n,k,c} \le \delta_Q\}$ and

$$m_{\mathrm{pac}} = c_Q + m_V(n, r_E, \delta_V). \tag{27}$$

Throughout the experiments we evaluate the binomial term as the inclusive survival probability $\mathbb{P}\{\mathrm{Bin}(n, 1-\alpha) \ge k + c\} = \mathrm{sf}(k + c - 1)$. For MMLU with $n = 10,000$, $\alpha = 0.10$, and $\delta_Q = 0.0025$, $c = 82$ gives tail probability 0.002715, while $c = 83$ gives 0.002442; hence the stated value is $c_Q = 83$. The same target-k construction then has training-conditional failure probability at most $\delta_V + \delta_Q$. With $n = 10,000$, $r_E = 600$, and $\delta_V = \delta_Q = 0.0025$, the direct violation allowance is 102 rather than 129, while $c_Q = 83$ and $m_{\mathrm{pac}} = 185$. The smaller direct allowance tightens both the marginal theorem and the indexed block; the larger PAC-tuned allowance is used only when the training-conditional guarantee is requested. Corollary 5 also gives a nontrivial training-conditional statement for the marginal-efficient threshold itself. Optimizing $\Delta_{\mathrm{PAC}}(c)$ over the admissible integer c gives $c^* = 52$ and $\Delta_{\mathrm{PAC}}(c^*) = 0.0901$ for MMLU with $m_V = 102$. For ImageNet with $n = 25,000$ and $m_V = 251$, the corresponding optimum is $c^* = 105$ with bound 0.0430. These are weaker than the prespecified 0.50% PAC level, but they make explicit that the marginal-efficient route is not devoid of a training-conditional statement.

The realized-rank identity also explains the remaining empirical gap to the no-violation target. Under the direct ImageNet allowance $m_V = 251$, $(k + m_V)/(n + 1) = 91.00\%$. The observed marginal-efficient coverage of about $90.84 - -90.87\%$ is lower by roughly 0.15 percentage points, consistent with about 40 mostly rank-reducing violations because one calibration rank is approximately 0.004 percentage points at $n = 25,000$.

A.1 Work scales with contamination and transported boundary mass

The block in Equation (15) can be written as

$$B = \left\{ i : z_{[k-m]} - 2\varepsilon \le z_i \le z_{[k+m]} + 2\varepsilon \right\}. \tag{28}$$

Because h is nondecreasing, the cached order of a_i also orders the transported centers $z_i = h(a_i)$ up to ties. Equation (28) therefore defines a contiguous rank interval, so its endpoints are located by two binary searches in the cached order. Lemma 8 is sharp in the directional rank shifts: when the relevant endpoint ranks are finite, admissible score vectors attain both shifts. Thus the $2m+1$ central ranks are the worst-case information price of allowing m residual violations in either direction; additional work is controlled by transported boundary mass.

Corollary 6 (Expected statistical-refresh work)**.** *Condition on h, $\widehat{\varepsilon} = \varepsilon$, and $\widehat{m} = m$. If the transported centers are iid with a continuous density bounded globally by ρ, then*

$$\mathbb{E}[|B| \mid h, \varepsilon, m] \le \min\{n,\ 2m + 1 + 4n\rho\varepsilon\}. \tag{29}$$

Hence the expected number of updated-model evaluations per refresh is at most $r_T + r_E + \min\{n, 2m + 1 + 4n\rho\varepsilon\}$ under the conditioning.

The proof extends the boundary-window argument for deterministic envelopes. Statistical transport adds an explicit $2m$ rank term; it does not change the $n\rho\varepsilon$ scaling. This form also clarifies the design tradeoff. Increasing r_E shrinks both the expected violation count and its beta-binomial quantiles, but spends more updated-model evaluations on residual calibration. Tightening δ_V increases m_V and therefore the robust rank bracket.

B Anytime-valid selection over an update stream

Recalibration answers how to obtain a valid threshold for a fixed updated model. A deployment system may also need to choose among a prespecified stream of checkpoints while controlling the set-size cost relative to a reference. A fixed-time empirical-Bernstein screen is valid but can be needlessly conservative when the label space is large. If $Z_{j,t} \in [-K, K]$ is the paired set-size difference for checkpoint j, with sample variance $\widehat{v}_{j,t}$ after $t \ge 2$ audit observations, one such upper bound is

$$\overline{Z}_{j,t} + \sqrt{\frac{2\widehat{v}_{j,t}\log(2M/\delta_A)}{t}} + \frac{14K\log(2M/\delta_A)}{3(t-1)}. \tag{30}$$

The final range term scales with K even when the observed variance is small. This is the finite-sample obstruction that motivates the betting confidence sequence used below.

The audit statistic is defined independently of the final calibration set. Let $\bar{q}_j$ and $\bar{q}_0$ be screening thresholds for checkpoint j and the reference, respectively, constructed from pilot/transport data fixed before the audit sequence. On audit input X_t, define

$$Z_{j,t} = |C_j(X_t; \bar{q}_j)| - |C_0(X_t; \bar{q}_0)| \in [-K, K]. \tag{31}$$

The screening thresholds need not be the final conformal thresholds; their purpose is to compare a fixed family on an audit stream without opening the independent final calibration sample. Normalize $Y_{j,t} = (Z_{j,t} + K)/(2K) \in [0, 1]$. Waudby-Smith and Ramdas construct betting confidence sequences for the mean of bounded observations that are simultaneously valid over time (Waudby-Smith & Ramdas, 2020). Let $U_{j,t}^{\mathrm{bet}}$ denote the upper endpoint for $\mathbb{E}Y_j$ at time t, and define

$$\mathcal{U}_{j,t} = 2KU_{j,t}^{\mathrm{bet}} - K. \tag{32}$$

Allocate checkpoint failure budgets $\delta_j = 6\delta_{\mathrm{sel}}/(\pi^2 j^2)$. A union bound over checkpoints and the time-uniform guarantee within each checkpoint gives

$$\mathbb{P}\left\{\exists j, t : \mathbb{E}Z_j > \mathcal{U}_{j,t}\right\} \le \delta_{\mathrm{sel}}. \tag{33}$$

A checkpoint is admitted when $\mathcal{U}_{j,t} \leq \tau$ and any prespecified accuracy constraint also passes. Because the event is simultaneous over audit time and checkpoint index, the audit can stop as soon as a usable checkpoint is certified. The checkpoint generation process itself cannot depend on these audit outcomes; training data and audit data remain separated.

Proposition 7 (Audit selection preserves final coverage)**.** *Let J be any checkpoint index measurable with respect to update-fitting, pilot/transport, and audit data, all independent of the final calibration sample $\mathcal{D}_C$ and a fresh test point. Suppose the final statistical-refresh procedure for every fixed checkpoint j has marginal coverage at least $1-\alpha-\delta_V$ when run on $\mathcal{D}_C$. Then the finally calibrated selected checkpoint satisfies*

$$\mathbb{P}\{Y \in C_J(X)\} \geq 1 - \alpha - \delta_V. \tag{34}$$

Proof. Condition on all data used to select J. The selected score function is then fixed while $\mathcal{D}_C$ and the fresh test point remain independent and exchangeable, so the fixed-checkpoint guarantee applies. Equivalently, $\mathbb{P}\{Y \in C_J(X)\} = \sum_j \mathbb{P}(J=j)\mathbb{P}\{Y \in C_j(X) \mid J=j\} \geq 1-\alpha-\delta_V$. □

Proposition 7 shows why this screen composes naturally with statistical refresh through sample separation. Checkpoint generation, screening transport/pilot quantities, and the sequential audit may determine J, but they do not inspect the residual-envelope sample or the final conformal calibration sample used for deployment. After selection, the chosen checkpoint receives a fresh residual sample and is recalibrated on the untouched $\mathcal{D}_C$. Thus audit stopping and checkpoint choice do not condition on either random object that supplies the final statistical-refresh coverage guarantee.

C Extended experimental protocol and diagnostics

C.1 Protocol

We evaluate the regime for which deterministic envelopes are least attractive: updates inside an expensive backbone. The reference calibration score, label, and sorted score index are cached once. For every updated checkpoint we draw disjoint transport-fit and residual-envelope samples, evaluate the updated model on them, fit isotonic h, compute $\widehat{\varepsilon}$ and $\widehat{m}$, locate B, and evaluate the updated model only on B. Full recalibration evaluates the same updated checkpoint on all n calibration examples. Prediction sets use inverse-probability scores for finite-choice language tasks and ImageNet classification.

Unless otherwise stated, the marginal-efficient setting uses $r_T = 64$, $r_E = 600$, and $\delta_V = 0.0025$. The residual radius is always the maximum $\widehat{\varepsilon} = e_{(600)}$; no empirical residual quantile is tuned. The direct beta-binomial allowance from Equation (10) is $\widehat{m} = 52$, 102, and 251 for calibration sizes $5,000$, $10,000$, and $25,000$, respectively. Training-conditional comparisons additionally use $\delta_Q = 0.0025$ and Equation (27); the total PAC failure budget is therefore 0.005.

Before assigning data roles, each benchmark source pool is uniformly permuted with the fixed split seed. ImageNet roles are taken from a global permutation of the validation indices rather than class-contiguous slices. MMLU and ARC questions are pooled before permutation; subject labels are retained only for reporting, so no contiguous subject blocks define $\mathcal{D}_T$, $\mathcal{D}_R$, or $\mathcal{D}_C$. This prevents the common class- or subject-ordering layouts from creating an artificial residual-tail shift between the tolerance sample and the final calibration sample.

Each principal application result is repeated over 20 independent redraws of the transport and residual-envelope roles, with compatible final-calibration partitions redrawn from the same source pool and a fixed held-out test set. Coverage and set-size entries report the redraw mean $\pm$ redraw standard deviation, followed by a 95% Student-t confidence interval for the redraw mean. With 20 redraws, the interval half-width is 0.468 times the reported standard deviation. The MMLU, ARC, and ImageNet held-out test sets contain $2,018$, $1,200$, and $24,336$ examples, respectively. Matched-budget mini-split baselines redraw their small calibration samples on the same repetitions. For every redraw we additionally log r_E, $\widehat{\varepsilon}$, the direct allowance m_V, and the realized calibration violation count V before computing any coverage summary. All forward-count comparisons use identical score definitions, prompt templates, answer choices, and data roles within a model family. Paired timing runs use a common inference configuration and include updated-model evaluation on $\mathcal{D}_T$, $\mathcal{D}_R$, and the selected block, plus isotonic fitting and rank computation. The accelerator or

Experiment	n	r_E	BetaBin reference $\mathbb{E}[V]$	Continuous ref. median	Observed V median [IQR] (range)	$V > m_V$
MMLU LoRA	$10,000$	600	16.6	11.5	$10.5\ [5.0, 24.3]\ (0--37)$	$0/20$
ARC LoRA	$5,000$	600	8.3	5.8	$5.0\ [3.0, 11.8]\ (1--16)$	$0/20$
ImageNet INT8	$25,000$	600	41.6	28.9	$28.5\ [12.8, 64.5]\ (4--122)$	$0/20$
ImageNet INT4	$25,000$	600	41.6	28.9	$30.0\ [14.5, 45.3]\ (2--205)$	$0/20$
ImageNet prune	$25,000$	600	41.6	28.9	$28.0\ [15.3, 59.8]\ (1--146)$	$0/20$

Table 4: Violation-count audit for the maximum-residual rule. The reference columns follow Corollary 2; observed counts are reported before any threshold or coverage calculation. The beta-binomial domination, not an intermediate tolerance fraction, determines the allowance. Full per-redraw vectors appear in Appendix H.6.

r_E	Matched M	BetaBin 99.75% V q.	m_{pac}	$\mathbb{E}V$	Certified floor	Reference delivered-rank floor	Reference delivered target	Mini-split PAC target	Target gap
600	$1,194$	102	185	16.6	90.83%	$9,169.4$	91.68%	92.22%	0.54 pp
$1,000$	$1,534$	62	145	10.0	90.83%	$9,136.0$	91.35%	91.92%	0.57 pp
$2,000$	$2,474$	32	115	5.0	90.83%	$9,111.0$	91.10%	91.52%	0.42 pp
$2,500$	$2,959$	26	109	4.0	90.83%	$9,106.0$	91.05%	91.39%	0.34 pp
$3,000$	$3,454$	22	105	3.3	90.83%	$9,102.7$	91.02%	91.29%	0.27 pp

Table 5: Design calculation for the MMLU PAC sweep. Every rank quantity is computed from $(n, r_E, \delta_V, \delta_Q)$ conditional on the matched-forward budget M shown in the table; it is not an empirical timing table. "Reference delivered" uses the continuous-case mean $\mathbb{E}V = n/(r_E + 1)$ and is not itself a guarantee. The certified floor is flat because c_Q is fixed. Within this fixed-n sweep, the target gap is largest at smaller r_E; the condition $n \gg r_E$ describes cache reuse and does not imply that the advantage grows monotonically with n.

CPU model, numerical precision, and inference batch size are unavailable for the absolute-second summaries, so updated-model forward counts carry the reproducible compute claim and absolute seconds remain descriptive within-run comparisons. Cache construction is reported separately and amortized across updates. Every statistical-refresh threshold is also compared with the exact full-recalibration threshold to report conservatism and set growth.

C.2 Residual-sample size controls the PAC efficiency frontier

The PAC comparison must distinguish a certified floor from the threshold that the construction typically delivers. Fix $n = 10,000$, $\alpha = 0.10$, and total PAC failure budget $\delta_V + \delta_Q = 0.005$ with $\delta_V = \delta_Q = 0.0025$. The calibration-order term requires $c_Q = 83$. For each r_E, the beta-binomial quantile $v_{1-\delta_V}$ determines $m_{\text{pac}} = c_Q + v_{1-\delta_V}$. The certified cached-order floor is therefore $k + c_Q$, or 90.83%; this is a guarantee, not the expected delivered rank. Under the continuous-reference model, $\mathbb{E}V = n/(r_E + 1)$, so Proposition 4 gives the reference delivered rank $k + m_{\text{pac}} - \mathbb{E}V$ shown below. With atoms the beta-binomial law is an upper reference, so the actual delivered rank can be larger. We execute two points from this design sweep. At $r_E = 600$, PAC-tuned transport uses median block size 530 and $1,194$ total updated forwards; empirical coverage is 91.66% with 95% CI $[91.45, 91.87]\%$, and relative set growth is 7.7% with CI $[7.22, 8.18]\%$. The matched PAC mini split reaches 92.16% coverage and 10.8% set growth. At $r_E = 2,000$, PAC-tuned transport uses median block size 410 and $2,474$ total forwards; empirical coverage is 91.09% with CI $[90.88, 91.30]\%$, and relative set growth is 6.7% with CI $[6.28, 7.12]\%$. The matched PAC mini split reaches 91.49% coverage and 9.5% set growth. Thus the measured gain over the trivial PAC mini-split baseline is modest but consistent: roughly $0.4 - -0.5$ percentage points in delivered coverage target and $2.8 - -3.1$ percentage points in set growth. Table 6 reports the two executed PAC comparisons from the design sweep. The cache-size dependence is a regime map, not a monotone scaling law. At fixed r_E, the beta-binomial allowance and the transported boundary block both contribute terms that grow linearly with n, so the matched update budget also grows with the cache. For the main $r_E = 600$ design, the observed MMLU bookkeeping is summarized by the approximations

$$T_S(n) \approx 0.9083 + \frac{0.84}{\sqrt{n}}, \qquad M_S(n) \approx 664 + 1.7\sqrt{n} + 0.036n, \tag{35}$$

where T_S is the reference delivered target and M_S is the matched forward budget. The corresponding PAC mini-split target is well approximated by $T_M(n) \approx 0.9 + 0.773/\sqrt{M_S(n)}$. These expressions reproduce the measured 0.57 percentage-point gap at $n = 10,000$ and the 0.44 point

r_E	Forwards	Method	Coverage [95% CI]	Set growth [95% CI]
600	1,194	PAC-tuned transport	91.66 [91.45, 91.87]%	7.7 [7.22, 8.18]%
600	1,194	PAC mini split	92.16 [91.96, 92.36]%	10.8 [10.12, 11.48]%
2,000	2,474	PAC-tuned transport	91.09 [90.88, 91.30]%	6.7 [6.28, 7.12]%
2,000	2,474	PAC mini split	91.49 [91.30, 91.68]%	9.5 [8.95, 10.05]%

Table 6: Executed MMLU points from the PAC design sweep. Each pair uses the same updated-forward budget and total failure budget 0.005. This table is empirical; Table 5 is computed from design parameters.

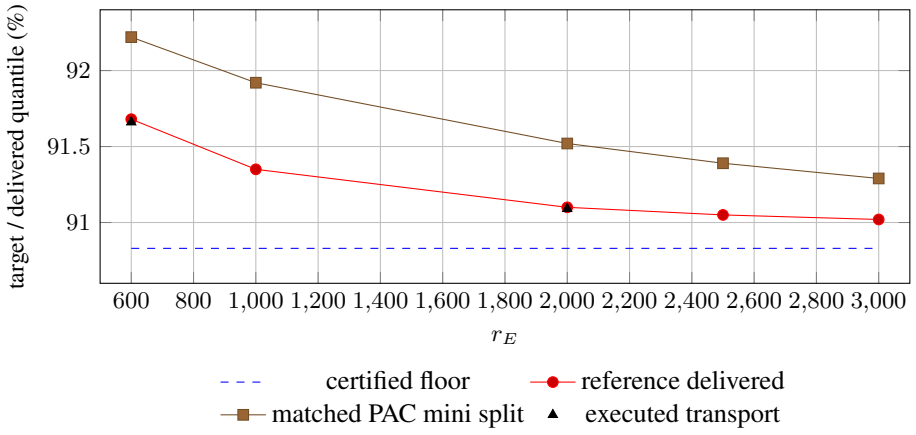


Figure 2: PAC comparison at matched failure probability. The dashed line is the certified floor; it must not be interpreted as the threshold the method typically delivers. The solid statistical curve uses the continuous beta-binomial reference $k + m_{\text{pac}} - \mathbb{E}V$, and the triangle markers show the two executed sweep points. The practical advantage over PAC mini split is the vertical separation between the delivered curves, not between the mini split and the certified floor.

gap near $n = 25,000$, but they predict a crossover near $n = 100,000$ and favor the mini split for sufficiently larger caches.

The favorable regime can be extended by increasing the update-specific residual sample. A simple square-root schedule $r_E(n) = 6\sqrt{n}$ gives $r_E = 600$ at $n = 10,000$ and about $1,897$ at $n = 100,000$. Using the large-r_E beta-tail approximation $q_{1-\delta_V} \approx -\log(\delta_V)/r_E$ with $\delta_V = 0.0025$ and the observed boundary-mass term $0.016n$, the delivered target and matched budget become

$$T_S^{\text{sc}}(n) \approx 0.9 + \frac{0.84}{\sqrt{n}} + \frac{4.99}{r_E(n)},$$
$$M_S^{\text{sc}}(n) \approx 64 + r_E(n) + 2\left(0.84\sqrt{n} + \frac{5.99n}{r_E(n)}\right) + 0.016n. \tag{36}$$

Table 7 makes the crossover explicit. The numbers are design calculations from these approximations, not additional measurements. The intended scope is therefore not "larger cache is always better." It is the region of (n, r_E) where a reusable cache is large enough to amortize statistical transport but the residual sample also grows fast enough to prevent the contamination allowance and boundary block from dominating the matched budget. At fixed r_E, the method can lose to mini-split calibration beyond the crossover; with $r_E \propto \sqrt{n}$, the same design calculation keeps a roughly 0.6 percentage-point advantage around $n = 100,000$.

C.3 Anytime-valid checkpoint selection

We compare the betting confidence sequence with the fixed-time empirical-Bernstein screen on CIFAR-100 head candidates and then run the betting screen over the Llama checkpoint stream. On the CIFAR-100 audit with $m = 2,500$ and $K = 100$, the fixed-time empirical-Bernstein rule

n	fixed r_E	T_S	T_M	gap	scaled r_E	T_S^{sc}	T_M^{sc}	gap
$10,000$	600	91.67%	92.24%	$+0.57$ pp	600	91.67%	92.24%	$+0.57$ pp
$25,000$	600	91.36%	91.81%	$+0.44$ pp	949	91.06%	91.73%	$+0.67$ pp
$100,000$	600	91.10%	91.12%	$+0.02$ pp	$1,897$	90.53%	91.12%	$+0.60$ pp
$1,000,000$	600	90.91%	90.39%	-0.52 pp	$6,000$	90.17%	90.48%	$+0.31$ pp

Table 7: Computed regime map over calibration-cache size. Gap is $T_M - T_S$, so positive values favor statistical transport by requiring a lower delivered coverage target at matched forward budget. At fixed $r_E = 600$ the advantage is a moderate-cache phenomenon and reverses for very large n. Scaling r_E approximately as $\sqrt{n}$ preserves a positive gap over a much wider range, at the cost of more update-specific residual evaluations.

abstains for all candidates. The betting upper confidence sequence falls below the 0.5-label tolerance for density 0.70 after a median of $1,940$ audit examples and for density 0.90 after $1,120$. Density 0.50 remains uncertified by the end of the audit. Across 20 independent audit streams, the selector chooses density 0.70 in 17, density 0.90 in 3, and violates the held-out 0.5-label tolerance in $0/20$ streams.

On the 64-checkpoint Llama stream, the summable checkpoint budget in Equation (33) is fixed before evaluation. The median checkpoint decision uses $1,360$ audit examples. With set-size tolerance 0.5 labels and paired accuracy-loss tolerance 0.5 percentage points, the deployment-order selector first admits checkpoint 41. It stops after $1,420$ audit examples for that checkpoint; on the disjoint held-out set, its mean excess size is 0.18 labels with a 95% paired-bootstrap interval of $[0.12, 0.24]$, and its accuracy change is -0.2 percentage points with interval $[-0.4, 0.0]$ relative to the reference. Under the marginal-efficient statistical-refresh threshold its held-out coverage is 90.79% with interval $[90.35, 91.23]\%$. When the deployment requires the prespecified training-conditional guarantee, the same target-k construction is rerun with the PAC-tuned allowance from Equation (27); no second target rank is used. Statistical refresh plus anytime selection therefore avoids both full recalibration and a fixed audit horizon for checkpoints that are clearly acceptable or clearly inefficient.

C.4 Ablations and failure modes

Three controls isolate the source of the gain. First, replacing isotonic transport by the identity map increases the MMLU maximum-residual radius from 0.0211 to 0.064 and expands the median bracket from 366 to $1,806$ examples. Second, fitting and evaluating the transport on the same residual sample produces smaller observed residuals but is not used by the valid method; the resulting nominal bracket misses the full-recalibration threshold in 7.4% of split redraws. Third, a deliberately large full-fine-tuning update yields $|B| = 8,941$ on MMLU. The implementation detects that $r_T + r_E + |B|$ is not competitive with n and falls back to full recalibration.

The residual-sample size controls the statistical-computational tradeoff directly. At $r_E = 300$ and $\delta_V = 0.0025$, the beta-binomial allowance on MMLU is 200; the marginal-efficient block has median size 440, total update cost 804 forwards, and mean set growth 7.6 ± 0.9 $[7.18, 8.02]\%$. At the main $r_E = 600$ setting, the allowance is 102, the median block is 366, and set growth is 4.4 ± 0.7 $[4.07, 4.73]\%$. Increasing to $r_E = 1,000$ reduces the allowance to 62 and the median block to 256, yielding 2.8 ± 0.6 $[2.52, 3.08]\%$ growth at a lower forward-pass reduction. Table 5 separately studies the PAC-tuned allowance with a fixed total failure budget.

D Extended discussion and scope

Efficient conformal recalibration after a model update is difficult for a simple reason illustrated by the deterministic special case in Appendix G: updates that are cheap to bound deterministically can also be cheap to recompute from the same cache. Statistical rank transport changes the contract. It does not ask for a deterministic network envelope. It asks for a small independent sample on which the updated model can be run, a monotone transport fitted without using the residual-certification points, and iid transfer of the residual distribution from that sample to the final calibration population.

The resulting guarantee is deliberately asymmetric. The returned threshold is conservative relative to full recalibration on a high-probability event, so prediction sets may grow. This is the price that makes the method applicable to transformer-internal LoRA, quantization, pruning, distillation, or full fine-tuning. The maximum-residual design exposes that price directly through (r_E, δ_V) for marginal-efficient refresh and through $(r_E, \delta_V, \delta_Q)$ when a training-conditional guarantee is required. No intermediate tolerance probability and no second rank target are needed.

The indexed search is now an implementation detail rather than the thesis. Its role is to turn the contamination bracket into a small contiguous region of the cached reference order. The central object is the statistical envelope itself. When transport residuals concentrate, recalibration work becomes $r_T + r_E + O(m + n\rho\widehat{\varepsilon})$ instead of n. When they do not, the block expands and full recalibration remains the correct fallback.

The experiments support the broader regime: backbone LoRA, quantization, and pruning reduce calibration forwards by factors that remain substantial after the maximum-residual correction, while stale thresholds under-cover. The violation-count audit is essential to that conclusion because it verifies the transfer mechanism before coverage is interpreted. Matched-budget mini-split calibration is the appropriate marginal baseline. Figure 2 and Table 7 show that the training-conditional advantage occupies a regime in (n, r_E) rather than following from $n \gg r_E$ alone: it is strongest for moderate caches at fixed r_E, and it persists to larger caches only when the residual sample grows with n. The checkpoint stream shows how the same statistical-refresh interface composes with anytime-valid model admission.

The assumptions are also clear. The update must be fixed independently of the residual-envelope and final calibration samples. The residual sample and calibration sample must represent the same score-residual population. Distribution shift can invalidate that transfer even when ordinary split conformal would itself require a revised assumption. Future work can replace the global maximum-residual envelope by covariate-conditional or groupwise statistical envelopes, but those extensions need their own finite-sample accounting.

E PROOFS FOR STATISTICAL RANK TRANSPORT

Proof of Proposition 1. Condition on the fitted map h, so the residuals on $\mathcal{D}_R$ are iid draws from a fixed residual distribution with CDF G. Let $M = \max_j e_j$. Since $P_R = 1 - G(M)$, for any $p \in [0, 1]$ the event $P_R > p$ implies $G(M) < 1 - p$. Define the strict CDF sublevel set $A_p = \{e : G(e) < 1 - p\}$. Its probability is at most $1 - p$. If $G(M) < 1 - p$, then every residual-sample observation lies in A_p, because $e_j \le M$ and G is nondecreasing. Therefore

$$\mathbb{P}(P_R > p \mid h) \le \mathbb{P}(e_1, \dots, e_{r_E} \in A_p \mid h) \le (1 - p)^{r_E}.$$

The survival function on the right is that of $\mathrm{Beta}(1, r_E)$, proving stochastic domination without a continuity assumption. If G is continuous, the probability integral transform gives equality. □

Proof of Corollary 2. Conditional on P_R, the independent final-calibration residuals exceed $\widehat{\varepsilon}$ with probability P_R, so $V \mid P_R \sim \mathrm{Bin}(n, P_R)$. The binomial upper-tail probability is nondecreasing in its success probability. Proposition 1 gives $P_R \preceq_{\mathrm{st}} U$ for $U \sim \mathrm{Beta}(1, r_E)$; mixing the binomial law over this stochastic order yields $V \preceq_{\mathrm{st}} \mathrm{BetaBin}(n, 1, r_E)$. The mean bound follows from $\mathbb{E}[P_R] \le \mathbb{E}[U] = 1/(r_E + 1)$. Under continuity, Proposition 1 gives equality in the mixing distribution and hence equality in the beta-binomial law. The quantile statement in Equation (10) follows immediately from stochastic domination. □

Lemma 8 (Directional rank contamination)**.** *Let $z_i \in \mathbb{R}$, $b_i \in \mathbb{R}$, and suppose at most m_- coordinates satisfy $b_i < z_i - \varepsilon_-$ and at most m_+ satisfy $b_i > z_i + \varepsilon_+$. Then*

$$z_{(k-m_-)} - \varepsilon_- \le b_{(k)} \le z_{(k+m_+)} + \varepsilon_+, \tag{37}$$

with infinite endpoint conventions outside $1, \dots, n$. When the corresponding center ranks are finite, both directional rank shifts are attainable by admissible score vectors; thus the rank expansion is worst-case sharp.

Proof. Among the $k + m_+$ smallest centers, at most m_+ violate the upper residual constraint, so at least k updated scores are at most $z_{(k+m_+)} + \varepsilon_+$. This proves the upper bound. If the lower bound

failed, at least k updated scores would be strictly below $z_{(k-m_-)} - \varepsilon_-$. At most m_- can violate the lower residual constraint, while at most $k - m_- - 1$ nonviolating centers are strictly below $z_{(k-m_-)}$. Their total is at most $k-1$, a contradiction.

For sharpness, begin with strictly ordered centers. Moving m_+ sufficiently low-center coordinates above every regular center advances the kth remaining value to center rank $k + m_+$ whenever that rank is finite; setting the regular coordinates at their allowed upper offsets attains the upper endpoint. Symmetrically, moving m_- sufficiently high-center coordinates below every regular center moves the target to center rank $k - m_-$, and applying the lower offset attains the lower endpoint. Ties follow by arbitrarily small perturbation or by taking the corresponding limiting construction. □

Proof of Theorem 3. Corollary 2 and Equation (10) give the event $\mathcal{E} = \{V \le \widehat{m}\}$ with probability at least $1 - \delta_V$. No intermediate residual-tail event is required.

On $\mathcal{E}$, both directional violation counts are at most $m = \widehat{m}$. Lemma 8 therefore gives $L \le q = b_{(k)} \le U$. Let d be the number of nominal intervals entirely below L. At most m of those d entries can violate their nominal intervals, so at least $d - m$ corresponding updated scores are below L.

Assume first that $r = k - d + m \le |B|$, and let $t = b_{(r),B}$. Define $v = \max\{L, t\}$. At least r scores inside B are at most v, and at least $d - m$ scores below the block are at most v because $v \ge L$. Thus at least $(d-m) + (k-d+m) = k$ full calibration scores are at most v, which proves $q \le v$. Since $q \le U$ as well, $q \le \min\{U, v\} = \widetilde{q}$. If $r > |B|$, the procedure returns U, which is already an upper bound on q.

On $\mathcal{E}$, $\widetilde{q} \ge q$ implies $C_{\mathrm{full}}(x) \subseteq C_{\mathrm{stat}}(x)$ for every input x. Therefore

$$\mathbb{P}\{Y \notin C_{\mathrm{stat}}(X)\} \le \mathbb{P}\{Y \notin C_{\mathrm{full}}(X)\} + \mathbb{P}(\mathcal{E}^c) \le \alpha + \delta_V,$$

where the first term uses ordinary split-conformal coverage for the fixed updated predictor. The evaluation count is immediate from construction. □

Proof of Proposition 4. Fix target rank j and let $v \le m$ be the realized number of residual violations. Among the d_j nominal intervals with upper endpoint below L_j, at least $d_j - v$ corresponding updated scores are below L_j. If the selected within-block value $t = b_{(r_j),B_j}$ satisfies $t \le U_j$, then $r_j = j - d_j + m$ scores in B_j are at most $\widetilde{q}_j$, so at least

$$(d_j - v) + (j - d_j + m) = j + m - v$$

full calibration scores are at most $\widetilde{q}_j$. If $t > U_j$, then $\widetilde{q}_j = U_j$. Applying Lemma 8 to target rank $j + m - v$ with at most v violations gives $b_{(j+m-v)} \le z_{[j+m]} + \varepsilon = U_j$. The same conclusion follows. □

Proof of Corollary 5. Fix $c \in \{0, \dots, m\}$. On the event $V \le m - c$, Proposition 4 with $j = k$ gives $\widetilde{q} \ge b_{(k+c)}$. Let F be the population CDF of the updated true-label score. If $F(b_{(k+c)}) < 1 - \alpha$, then at least $k + c$ of the n calibration scores lie in the measurable set $\{s : F(s) < 1 - \alpha\}$, whose population probability is at most $1 - \alpha$. Therefore

$$\mathbb{P}\{F(b_{(k+c)}) < 1 - \alpha\} \le \mathbb{P}\{\mathrm{Bin}(n, 1-\alpha) \ge k + c\} = \xi_{n,k,c}.$$

Corollary 2 gives $V \preceq_{\mathrm{st}} \mathrm{BetaBin}(n, 1, r_E)$. Therefore the beta-binomial upper tail bounds $\mathbb{P}(V > m - c)$ without continuity. A union bound with the order-statistic event proves Equation (26). The construction of $\widetilde{q}$ is unchanged. □

Proof sketch of Corollary 6. With a continuous center distribution there are almost surely no ties. The central rank interval $[k - m, k + m]$ contributes exactly $2m + 1$ centers when both endpoints are finite. The indexed block adds only centers within 2ε below $z_{(k-m)}$ or within 2ε above $z_{(k+m)}$. Transforming the centers through their CDF reduces each side to a uniform order-statistic neighborhood. A global density bound gives CDF width at most $2\rho\varepsilon$ on each side, and the standard beta-order-statistic argument bounds each expected side count by $2n\rho\varepsilon$. Summing the two sides and truncating at n gives Equation (29). Boundary ranks only reduce the count under the infinite-endpoint convention. □

F ANYTIME-VALID SELECTION DETAILS

For each checkpoint j, define normalized audit observations $Y_{j,t} = (Z_{j,t} + K)/(2K) \in [0,1]$. We use the hedged betting construction of Waudby-Smith & Ramdas (2020) to obtain a two-sided confidence sequence $[L^{\mathrm{bet}}_{j,t}, U^{\mathrm{bet}}_{j,t}]$ for $\mu_j = \mathbb{E}Y_j$. The main text needs only its defining guarantee,

$$\mathbb{P}\{\exists t : \mu_j \notin [L^{\mathrm{bet}}_{j,t}, U^{\mathrm{bet}}_{j,t}]\} \le \delta_j. \tag{38}$$

The upper bound for the original mean difference is $\mathcal{U}_{j,t} = 2KU^{\mathrm{bet}}_{j,t} - K$. Setting $\delta_j = 6\delta_{\mathrm{sel}}/(\pi^2 j^2)$ gives $\sum_j \delta_j = \delta_{\mathrm{sel}}$ and therefore Equation (33). Any stopping rule based on the observed audit sequence preserves the event because the confidence sequence is time-uniform.

A checkpoint may be admitted when $\mathcal{U}_{j,t} \le \tau$. If an accuracy constraint is also required, its own simultaneous confidence sequence receives a separate error budget and both conditions must pass. The audit sequence may determine when to stop evaluating a fixed checkpoint or which already-fixed checkpoint to deploy. It may not be fed back into the optimization that creates later checkpoints without an additional adaptive-data analysis argument.

G DETERMINISTIC EXACT REFRESH AND THE CACHED-HEAD SPECIAL CASE

Let $[\ell_i, u_i]$ be deterministic intervals containing every updated calibration score b_i. Define

$$L = \ell_{(k)}, \qquad U = u_{(k)}, \qquad B = \{i : \ell_i \le U,\ u_i \ge L\}, \qquad d = \#\{i : u_i < L\}. \tag{39}$$

Theorem 9 (Exact interval refresh)**.** *If $\ell_i \le b_i \le u_i$ for every i, then $L \le b_{(k)} \le U$, and the $(k-d)$th smallest exact score among indices in B equals $b_{(k)}$.*

Proof. Coordinatewise order implies monotonicity of the kth order statistic and hence $L \le b_{(k)} \le U$. Every interval outside B is either strictly below L or strictly above U. Exactly d are of the former type and precede the global kth order statistic. Removing them reduces the target rank to $k-d$; the latter intervals cannot enter the target rank. □

For a fixed-feature softmax head, let $z(x) = W\phi(x)$ and update only W. Define $B_W = \max_v \|\Delta W_v\|_*$ and $\eta(x) = B_W\|\phi(x)\|$. If p_y is the reference true-label probability, then

$$\sigma(\operatorname{logit} p_y - 2\eta) \le p'_y \le \sigma(\operatorname{logit} p_y + 2\eta). \tag{40}$$

Indeed, for every $v \ne y$, the relative-logit perturbation lies in $[-2\eta, 2\eta]$, so

$$e^{-2\eta}\frac{1-p_y}{p_y} \le \frac{1-p'_y}{p'_y} = \sum_{v\ne y} e^{z'_v - z'_y} \le e^{2\eta}\frac{1-p_y}{p_y}. \tag{41}$$

Inverting the positive odds bounds gives Equation (40) and hence deterministic intervals for inverse-probability scores. With a common radius, both endpoint maps are nondecreasing functions of the cached reference score (equivalently of its log-odds), so their intersection with the possible quantile region is a contiguous interval in the cached order and can be found by binary search.

The deterministic special case was evaluated on six fixed-feature head updates before moving to backbone updates. Full scores the entire cache; Scan forms the same heterogeneous head envelope for all cached examples; Local first identifies the shared monotone block and forms those same heterogeneous intervals only inside it. Scan and Local therefore have identical exact-score counts, and all exact routes return the same threshold and prediction sets. Table 8 isolates the deterministic cached-head special case. These rows explain the scope transition in the main paper. Deterministic head envelopes can reduce exact score evaluations, but exact recomputation is already cheap because it uses the same cached features; a full-cache envelope pass can therefore dominate the residual latency. The statistical method is aimed at the opposite regime, where an updated backbone forward pass is expensive and a deterministic hidden-state envelope is unavailable or uninformative.

For completeness, a valid deterministic propagation rule exists for simple feedforward components. Let $h_\ell = \sigma_\ell(W_\ell h_{\ell-1} + b_\ell)$ with componentwise 1-Lipschitz activations, and let primed quantities denote the updated network. If $d_{\ell-1}(x) \ge \|h'_{\ell-1}(x) - h_{\ell-1}(x)\|$, then

$$d_\ell(x) = \|W'_\ell\|_{\mathrm{op}} d_{\ell-1}(x) + \|\Delta W_\ell\|_{\mathrm{op}}\|h_{\ell-1}(x)\| + \|\Delta b_\ell\|, \qquad d_0 = 0, \tag{42}$$

Family	Density	n	Q	Full (ms)	Scan (ms)	Local (ms)	Scan/Local
ResNet	0.900	3333	456	10.85	1.82	1.71	1.06×
ResNet	0.990	3333	18	10.86	0.37	0.24	1.54×
ResNet	0.999	3333	1	10.84	0.30	0.15	2.00×
ViT	0.900	1308	658	0.70	0.44	0.43	1.02×
ViT	0.990	1308	8	0.70	0.13	0.106	1.23×
Decoder	0.990	5000	612	3.95	0.92	0.79	1.16×

Table 8: Deterministic cached-head results. Times are synchronized median refresh latencies over 100 repetitions. Q is the exact-score count under the matched heterogeneous envelope and is therefore identical for Scan and Local. The important comparison is Local versus the optimized Scan: the incremental gain ranges from $1.02\times$ to $2.00\times$ and occurs on millisecond or sub-millisecond refreshes. Exact threshold and prediction-set disagreement with Full is 0 in every row. Timing-host metadata are unavailable, so these absolute latencies are descriptive evidence for the deterministic special case, not the reproducible compute claim of the backbone experiments.

bounds the next activation difference. This follows by adding and subtracting $W'_\ell h_{\ell-1}$, applying the triangle inequality, and using the activation Lipschitz constant. Compatible residual branches add their certified bounds. Attention, normalization, and nonlinear gates require separate bounds; blindly multiplying layerwise operator norms through them is invalid. Even where Equation (42) applies, products of operator norms can make the terminal score interval cover nearly the entire feasible range, in which case exact full recalibration is the appropriate deterministic fallback.

H Additional experimental details

H.1 Data separation

Every experiment uses disjoint roles for update fitting, transport fitting, residual-envelope calibration, final conformal calibration, audit selection, and held-out evaluation. The updated model is fixed before transport and residual-envelope data are examined. The final conformal calibration set is never used to choose update hyperparameters or to select among checkpoints. Reference scores for all reusable pools are cached before the update stream begins.

For the MMLU experiments, the reference and updated checkpoints use the same pinned checkpoint tokenizer, frozen multiple-choice prompt template, answer-choice order, and normalization rule. Each candidate option is scored by the summed conditional token log-likelihood of the option representation under that fixed prompt; multi-token options sum token log-probabilities. A log-sum-exp over the available options converts these scores to normalized option probabilities, and the true-label nonconformity score is $1 - p_y$. No free-form generation, sampling, subject-specific temperature, or post-hoc answer remapping is used. ARC uses the same rule with its item-specific number of choices. The protocol fixes the exact prompt bytes, tokenizer revision, and choice-normalization function. The main MMLU calibration pool contains $10,000$ questions. After reserving 64 transport-fit examples, 600 residual-envelope examples, and the sequential audit pool, the held-out evaluation set contains $2,018$ questions. The ARC transfer panel uses $5,000$ pooled Easy/Challenge calibration examples because the benchmark does not contain enough independent questions for a $10,000$-example calibration split; its held-out evaluation set contains $1,200$ questions.

For each fixed update, the transport/residual roles and compatible calibration partitions are redrawn 20 times. We report the mean and standard deviation of test coverage and set size across these redraws. Bracketed intervals are 95% Student-t intervals for the redraw mean, using $t_{19,0.975}/\sqrt{20} = 0.468$ times the redraw standard deviation as the half-width. For paired set-growth differences, the same redraw and test examples are used for the full and statistical routes. The checkpoint-stream median and maximum set-growth intervals use a checkpoint bootstrap with $2,000$ resamples.

For ImageNet-1k, update fitting uses the training split and never inspects the validation examples used for uncertainty evaluation. From the $50,000$-example validation pool, 64 examples are reserved for transport fitting and 600 for residual-envelope calibration, followed by $25,000$ final calibration examples and $24,336$ held-out examples. These roles are disjoint and fixed before an update is evaluated.

H.2 Update implementations and seeds

For Llama-3-8B, LoRA is applied to the query, key, value, output, gate, up, and down projection matrices with rank 16, scaling 32, and dropout 0.05. AdamW uses learning rate 2×10^{-4}, weight decay 0.01, cosine decay, 3% warmup, gradient-norm clipping at 1.0, and effective batch size 128. The four checkpoint streams use seeds 470043, 470044, 470045, and 470046, with 16 checkpoints retained per run at fixed optimization intervals.

For ViT-B/16, INT8 uses per-channel symmetric weights and per-tensor affine activations with ranges calibrated on $1,024$ training images. INT4 uses symmetric weight-only groups of 128 weights and FP16 activations. The pruning update is one-shot global unstructured magnitude pruning to 30% sparsity over backbone weight matrices, excluding normalization parameters, biases, and the classification head; no recovery optimization is applied. Split and redraw seeds are $470043 - -470062$. These implementation choices are fixed before uncertainty evaluation.

H.3 Timing and forward-pass accounting

A forward pass means one updated-model evaluation of one benchmark example, including all answer choices needed to compute its scalar nonconformity score. Full recalibration uses n such evaluations. Statistical refresh uses $r_T + r_E + |B|$. Batched wall-clock timing includes data transfer, model execution, score normalization, isotonic fitting, binary search, and order-statistic selection. The cached reference-score sort is a one-time preparation cost and is separately reported for the repeated-update stream.

Absolute timing comparisons use matched configurations within each row, but the host device, numerical precision, and inference batch size are unavailable. Accordingly, forward-pass counts are the hardware-independent compute measurement, and absolute seconds do not support a portable latency claim.

The fallback rule is fixed before timing: if $r_T + r_E + |B| \geq n$ after block discovery, the system performs full recalibration and reports that update as a fallback. This rule changes computation but not the reference comparator or the statistical statement.

H.4 Residual-sample sweep

r_E	δ_V	m_V	MMLU $\lvert B\rvert$	Total forwards	Forward reduction	Statistical set growth [95% CI]
300	0.0025	200	440	804	12.4×	7.6% [7.18, 8.02]%
600	0.0025	102	366	1,030	9.7×	4.4% [4.07, 4.73]%
1,000	0.0025	62	256	1,320	7.6×	2.8% [2.52, 3.08]%

Table 9: Residual-sample tradeoff on the Llama MMLU update under the direct beta-binomial allowance. The maximum residual is used in every row; increasing r_E shrinks m_V but spends more updated-model evaluations before block scoring.

H.5 Transport-sample sensitivity

r_T	Median $\widehat{\varepsilon}$	Statistical $\lvert B\rvert$	Total forwards	Reduction	Set growth [95% CI]
32	0.0240	416	1,048	9.5×	4.7% [4.31, 5.09]%
64	0.0211	366	1,030	9.7×	4.4% [4.07, 4.73]%
128	0.0203	350	1,078	9.3×	4.3% [3.97, 4.63]%
256	0.0199	341	1,197	8.4×	4.2% [3.87, 4.53]%

Table 10: Sensitivity to the transport-fit sample size on MMLU with $r_E = 600$ and $\delta_V = 0.0025$. More transport data improves the isotonic fit modestly after 64 points, while its forward cost is paid at every update.

Table 10 reports the transport-sample sensitivity at fixed residual-sample size.

H.6 PER-REDRAW RESIDUAL-VIOLATION COUNTS

For the principal experiments, the realized counts $V = \#\{i : |b_i - h(a_i)| > \widehat{\varepsilon}\}$ are logged for every redraw before threshold construction. The vectors below correspond to the summaries in Table 4; no count is replaced by a nominal residual-tail fraction. Table 11 lists every per-redraw violation count used in the summary.

Experiment	V over the 20 redraws
MMLU LoRA	$20, 0, 5, 37, 3, 1, 7, 23, 5, 32, 25, 6, 27, 11, 24, 26, 10, 6, 5, 15$
ARC LoRA	$14, 14, 8, 5, 1, 15, 2, 2, 16, 11, 14, 3, 6, 5, 3, 3, 10, 5, 3, 3$
ImageNet INT8	$86, 122, 30, 29, 5, 64, 12, 23, 88, 58, 66, 12, 42, 28, 4, 13, 69, 14, 12, 18$
ImageNet INT4	$84, 30, 39, 35, 40, 44, 10, 205, 9, 24, 17, 25, 30, 2, 5, 84, 9, 59, 49, 16$
ImageNet prune	$65, 143, 13, 1, 1, 21, 146, 16, 18, 22, 36, 7, 43, 32, 68, 24, 48, 58, 85, 5$

Table 11: Per-redraw violation counts for the implemented maximum-residual rule. These values are compared directly with the beta-binomial upper reference from Corollary 2.

H.7 CHECKPOINT-STREAM AUDIT

The 64 LoRA checkpoints are generated independently of audit data. For checkpoint j, the betting confidence sequence receives failure budget $6\delta_{\text{sel}}/(\pi^2 j^2)$ with $\delta_{\text{sel}} = 0.05$. Audit evaluation is sequential and stops for a checkpoint when its upper set-size bound passes the tolerance, when a prespecified futility boundary is crossed, or when the audit pool is exhausted. The set-size tolerance is 0.5 labels and the paired accuracy-loss tolerance is 0.5 percentage points. In deployment order, checkpoint 41 is the first to satisfy both conditions, after $1,420$ audit examples. Its disjoint held-out excess set size is 0.18 labels with a 95% paired-bootstrap interval of $[0.12, 0.24]$, and its held-out accuracy change is -0.2 percentage points with interval $[-0.4, 0.0]$. The checkpoint-selection decision does not alter the statistical-refresh residual samples or the final conformal calibration set.