AUDIT-CERTIFIED CONFORMAL ROUTING FOR RELIABLE HYBRID TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid post-training combines scalar-reward updates with distillation, but noisy reward estimates can route examples into wrong-direction updates. We study a split-conformal sign gate with an independent audit of wrong-sign risk among routed examples. A single global conformal radius induces exactly a magnitude threshold; its contribution is a finite-sample marginal interpretation of that threshold, not a different sample ranking. Our evaluation separates calibration, audit, routing diagnosis, policy fitting, and final testing. On GSM8K, the global gate obtains 50.0 ± 0.5 accuracy, while studentized routing obtains 50.7 ± 0.5, versus 49.8 ± 0.5 for an equally audited magnitude gate; sample-adaptive conformal variants are also evaluated on MATH, HumanEval, and two additional 3B backbones. A separate group-aware construction addresses the sign of the centered advantage used by GRPO. These guarantees concern routing on the stated distribution, not optimization gains or robustness to arbitrary policy drift.

1 INTRODUCTION

Large language model post-training increasingly combines scalar reward feedback and distillation from a teacher model (Yang et al., 2026b; Ko et al., 2026; Yang et al., 2026c). A central choice is which examples receive a reward-based update and which instead use distillation (Li et al., 2026; Yang et al., 2026a). This choice is consequential when reward estimates are noisy or shifted, as in verifier- and process-reward settings (Lightman et al., 2023; Wang et al., 2024).

We use split-conformal intervals (Vovk et al., 2005; Angelopoulos & Bates, 2021) to certify the reward sign relative to a margin, followed by an independent audit of the conditional wrong-sign rate among routed examples. With one global radius, the rule is algebraically a threshold on $|\hat{r}|$. Its finite-sample calibration gives that threshold a marginal error interpretation; it does not change the ordering of examples. Studentized and conformalized-quantile intervals allow the width to depend on the example and test whether this added adaptivity matters empirically.

Our contributions are a marginal wrong-sign guarantee with an explicit fixed-α protocol, an independent finite-family audit for conditional routed risk, comparisons against an equally audited magnitude gate and fresh held-out routing measurements, and a separate group-aware certificate for the centered advantage used by GRPO. The statistical claims concern routing under their stated distributional assumptions; downstream accuracy is an empirical outcome.

2 PROBLEM SETUP

Let $X \in \mathcal{X}$ denote a prompt and let $Y \in \mathcal{Y}$ denote a sampled completion. A labeled example is represented by

$$Z = (X, Y, R), \qquad R = r^\star(X, Y) \in \mathbb{R},$$

where R is a trusted signed target available on calibration or audit data. In exact-match math tasks such as GSM8K and MATH (Cobbe et al., 2021; Hendrycks et al., 2021), R can be a centered correctness utility; in code tasks such as HumanEval (Chen et al., 2021), it can be a centered unit-test pass rate. The sign of R determines whether the scalar reward update is directionally reliable. We use a nonnegative margin $\tau \geq 0$: values $R \geq \tau$ support a positive reward update, values $R \leq -\tau$ support a negative reward update, and values in $(-\tau, \tau)$ form an indifference region.

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.

A reward estimator

$$\hat{r} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

assigns a scalar score to a prompt–completion pair. The hybrid training loop has two update branches: a reward branch, which uses $\hat{r}$ inside a reinforcement-style objective such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024), and a distillation branch, which uses a teacher or reference-policy objective. The router emits

$$b(x, y) \in \{-1, 0, +1\},$$

where $+1$ denotes a positive reward update, -1 denotes a negative reward update, and 0 denotes a distillation update. The binary reward-routed event is

$$A(x, y) = \mathbf{1}\{b(x, y) \neq 0\}.$$

Assumption 1 (Proper splitting and fixed reward estimator)**.** The data used to train $\hat{r}$ are independent of the calibration and audit data used below. Equivalently, all guarantees are conditional on a fixed measurable function $\hat{r}$ that was produced without using the calibration and audit labels. If the same labeled pool is used for reward-model training and calibration, the construction is implemented with cross-fitting so that every residual is evaluated out of sample.

Assumption 2 (Exchangeable calibration distribution)**.** For the base guarantee, the calibration examples $Z_1, \ldots, Z_n$ and the future example Z_{n+1} are exchangeable conditional on $\hat{r}$. A common sufficient condition is i.i.d. sampling from a fixed distribution P_0 over (X, Y, R) after $\hat{r}$ has been fixed.

3 Conformal routing

The protocol uses prompt-ID-disjoint reward-model fitting, conformal calibration, risk audit, fresh routing test, policy fitting, optimization-development, and final evaluation roles (Table 1). The primary $\alpha = 0.15$ is fixed before evaluation; $\hat{q}_\alpha$ is derived only from calibration residuals. Only τ is selected by the independent audit, never by final-test accuracy.

Table 1: Prompt-ID-disjoint source-pool accounting. The 3B code extension repartitions APPS as stated in Appendix E; HumanEval is final test only.

Source pool	RM fit	Calib.	Audit	Routing test	Policy fit	Opt. dev.	Final test
GSM8K train (7,473)	3,000	800	600	600	2,223	250	GSM8K test (1,319)
MATH train (7,500)	2,500	1,000	750	750	2,250	250	MATH test (5,000)
APPS train, fixed 3,000 IDs	1,200	400	400	400	500	100	HumanEval (164)

Given a calibration set $\mathcal{D}_{\text{cal}} = \{Z_i = (X_i, Y_i, R_i)\}_{i=1}^n$, define residual scores

$$S_i = |R_i - \hat{r}(X_i, Y_i)|.$$

Following the standard split-conformal construction (Vovk et al., 2005; Angelopoulos & Bates, 2021), for a target marginal error level $\alpha \in (0, 1)$, set

$$k_\alpha = \lceil (n+1)(1-\alpha) \rceil, \qquad \hat{q}_\alpha = \begin{cases} S_{(k_\alpha)}, & k_\alpha \leq n, \\ +\infty, & k_\alpha = n+1, \end{cases} \tag{1}$$

where $S_{(1)} \leq \cdots \leq S_{(n)}$ are the ordered residuals. The $+\infty$ convention handles the finite-sample boundary case $\alpha < 1/(n+1)$ and makes the router abstain from reward updates at unattainably strict levels.

For a new pair (x, y), define the split-conformal interval

$$\mathcal{C}_\alpha(x, y) = [\hat{r}(x, y) - \hat{q}_\alpha,\ \hat{r}(x, y) + \hat{q}_\alpha]. \tag{2}$$

The conformal router is

$$b_{\alpha,\tau}(x, y) = \begin{cases} +1, & \hat{r}(x, y) - \hat{q}_\alpha \geq \tau, \\ -1, & \hat{r}(x, y) + \hat{q}_\alpha \leq -\tau, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

3.1 TRAINING OBJECTIVE INDUCED BY THE ROUTER

The router chooses the update branch; the reinforcement and distillation objectives are specified separately. We use exclusive branches in the primary comparison to isolate the routing decision. An auxiliary distillation loss on reward-routed examples can be added without changing the gate or its certificate (Table 10). In our experiments, the reward branch follows a GRPO-style objective (Shao et al., 2024). For a minibatch, define

$$B_R = \{i : b_i \in \{-1, +1\}\}, \qquad B_D = \{i : b_i = 0\}.$$

Let

$$\mathcal{L}_R(\theta) = \begin{cases} |B_R|^{-1} \sum_{i \in B_R} \mathcal{L}_{\mathrm{GRPO}}(x_i, y_i, \hat{r}_i; \theta), & |B_R| > 0, \\ 0, & |B_R| = 0, \end{cases} \tag{4}$$

and

$$\mathcal{L}_D(\theta) = \begin{cases} |B_D|^{-1} \sum_{i \in B_D} \ell_D(x_i, y_i; \theta), & |B_D| > 0, \\ 0, & |B_D| = 0. \end{cases} \tag{5}$$

For tokenizer-compatible Qwen student–teacher pairs, ℓ_D is the token-level KL from the frozen Qwen teacher. For Llama students, the teacher generates text, which is retokenized with the Llama tokenizer; ℓ_D is teacher-forced cross-entropy on that sequence. A cross-vocabulary token-level KL is not used. Each update minimizes

$$\mathcal{L}(\theta) = \lambda_R \mathcal{L}_R(\theta) + \lambda_D \mathcal{L}_D(\theta), \tag{6}$$

where λ_R and λ_D control the relative scale of reward and distillation updates. In the implementation used for the experiments, routing occurs before advantage normalization; reward advantages are normalized over the active reward-routed set, and empty branch terms are skipped as in Eqs. (4)–(5).

4 THEORY

The results below certify the statistical behavior of the router. They do not imply that any particular optimizer must improve downstream accuracy.

4.1 FINITE-SAMPLE MARGINAL ROUTING GUARANTEE

Theorem 3 (Split-conformal coverage)**.** *Under Assumptions 1 and 2, the interval in Eq.* (2) *satisfies*

$$\mathbb{P}\big(R_{n+1} \in \mathcal{C}_\alpha(X_{n+1}, Y_{n+1})\big) \geq 1 - \alpha,$$

conditional on the fixed estimator $\hat{r}$. If the residual scores $S_1, \ldots, S_{n+1}$ are almost surely distinct, then

$$\mathbb{P}\big(R_{n+1} \in \mathcal{C}_\alpha(X_{n+1}, Y_{n+1})\big) \leq 1 - \alpha + \frac{1}{n+1}.$$

Definition 4 (Wrong-sign reward routing)**.** For a future example $Z = (X, Y, R)$ and router $b_{\alpha,\tau}$, define

$$W_{\alpha,\tau}(Z) = \{b_{\alpha,\tau}(X, Y) = +1,\ R < \tau\} \cup \{b_{\alpha,\tau}(X, Y) = -1,\ R > -\tau\}. \tag{7}$$

Proposition 5 (Marginal wrong-sign routing control)**.** *Under Assumptions 1 and 2,*

$$\mathbb{P}\big(W_{\alpha,\tau}(Z_{n+1})\big) \leq \alpha.$$

Moreover, if $Z_{n+1}, \ldots, Z_{n+T}$ are future examples such that each Z_{n+t} is marginally exchangeable with the calibration set conditional on $\hat{r}$, then

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{W_{\alpha,\tau}(Z_{n+t})\}\right] \leq \alpha T.$$

4.2 AUDIT-CERTIFIED ROUTING FOR CONDITIONAL ROUTED RISK

Marginal wrong-sign control does not bound risk conditional on routing. An independent audit provides a finite-family guarantee (Angelopoulos et al., 2025; 2022; Bates et al., 2021). Fix α and $\hat{q}_\alpha$ before inspecting audit labels. For the primary experiment, let $\Lambda = \{\tau_1, \ldots, \tau_M\} = \{0, 0.05, 0.10\}$ and $M = 3$. More generally, the argument applies to any finite family of candidate routers fixed independently of the audit set; selecting a radius from that family on audit data would retain the audit guarantee but would not automatically retain the single-radius marginal conformal guarantee.

For each candidate $\lambda \in \Lambda$, on m i.i.d. audit examples count $a_\lambda = \sum_i \mathbf{1}\{A_\lambda(Z_i) = 1\}$ routed and $w_\lambda = \sum_i \mathbf{1}\{W_\lambda(Z_i)\}$ wrong-sign examples. At familywise failure probability δ, define the one-sided exact Clopper–Pearson upper bound

$$U(\lambda) = \begin{cases} 1, & a_\lambda = 0 \text{ or } w_\lambda = a_\lambda, \\ \mathrm{Beta}^{-1}(1 - \delta/M; w_\lambda + 1, a_\lambda - w_\lambda), & \text{otherwise}, \end{cases} \tag{8}$$

where $\mathrm{Beta}^{-1}(u; a, b)$ is the u quantile of a beta distribution. Select the candidate with the largest audit routed count subject to $U(\lambda) \le \beta$; use distillation only if none is feasible. For the 1–1.5B primary experiments $\beta = 0.25$ and $\delta = 0.05$ are prespecified, not selected by accuracy. The separately reported 3B extension uses the task-specific budgets listed in Appendix E.1.

Theorem 6 (Finite-sample routed-risk audit)**.** *Conditional on the fixed candidate family, suppose the m audit examples are i.i.d. from P_0 and independent of candidate construction. With probability at least $1 - \delta$ over the audit set, simultaneously for every candidate with $U(\lambda) \le \beta$ and $p_A(\lambda) > 0$,*

$$\mathbb{P}_{P_0}(W_\lambda(Z) \mid A_\lambda(Z) = 1) \le \beta. \tag{9}$$

The adaptively selected nonempty router inherits this bound. This is an audit guarantee on P_0, separate from split-conformal marginal coverage.

For the fixed conformal radius, the marginal claim also survives selection of τ on the audit split: for every realized margin $\tau \ge 0$, $W_{\alpha,\tau}(Z) \subseteq \{R \notin \mathcal{C}_\alpha(X, Y)\}$ pointwise. Thus even a data-dependent choice of margin cannot increase the marginal wrong-sign probability beyond conformal miscoverage. The sensitivity sweep over α in Appendix B.1 does not select the primary model.

4.3 EXAMPLE-ADAPTIVE INTERVALS

The global gate in Eq. (3) is a magnitude gate: up to score-zero and boundary tie handling, $A_{\alpha,\tau}(x, y) = 1$ exactly when $|\hat{r}(x, y)| \ge \hat{q}_\alpha + \tau$. An explicit cutoff $c = \hat{q}_\alpha + \tau$ and a correctly implemented top-$|\hat{r}|$ gate at the induced fraction yield identical masks when score ties are broken by immutable example ID. On GSM8K, global conformal routing reaches 50.0 ± 0.5. Seed-paired contrasts give studentized minus audit-tuned magnitude +0.9 [+0.3, +1.5] and studentized minus uncertainty threshold +1.6 [+1.0, +2.2].

To study example-dependent widths, fit a positive scale estimate $\hat{s}(x, y)$ on reward-model fitting data and floor it consistently as $s(x, y) = \max\{\hat{s}(x, y), s_{\min}\}$. Studentized calibration uses $S_i^{\mathrm{stu}} = |R_i - \hat{r}_i|/s_i$ and the same finite-sample quantile rule as Eq. (1). Its interval is $[\hat{r} - \hat{q}_\alpha^{\mathrm{stu}} s, \hat{r} + \hat{q}_\alpha^{\mathrm{stu}} s]$, and the corresponding endpoints replace those of Eq. (3). For conformalized quantile regression, fit lower and upper quantile functions $\hat{\ell}, \hat{u}$ on reward-model fitting data, calibrate $S_i^{\mathrm{CQR}} = \max\{\hat{\ell}_i - R_i, R_i - \hat{u}_i\}$, and route with the expanded interval $[\hat{\ell} - \hat{q}_\alpha^{\mathrm{CQR}}, \hat{u} + \hat{q}_\alpha^{\mathrm{CQR}}]$. Neither scale nor quantile models use calibration, audit, routing-test, or final-test labels during fitting. Both variants retain the usual marginal split-conformal argument when their fitting and calibration assumptions hold.

4.4 GROUP-NORMALIZED DIRECTION

The raw-sign event in Proposition 5 concerns the pre-normalization reward; group centering can reverse the sign of a GRPO advantage. For a group G of eight completions and its raw-gate active set A_G, define for each $i \in A_G$

$$A_i^\star = R_i - \frac{1}{|A_G|} \sum_{j \in A_G} R_j, \qquad \hat{A}_i = \hat{r}_i - \frac{1}{|A_G|} \sum_{j \in A_G} \hat{r}_j. \tag{10}$$

Split the conformal-calibration IDs into raw-residual and group-advantage halves before inspection. Compute the raw radius on the first half and lock the margin on the raw audit before opening the independent group-calibration half. With that active-set rule fixed, use $S_G = \max_{i \in A_G} |A_i^\star - \hat{A}_i|$ and its group-level split-conformal quantile q_G. A group uses reward only if $|A_G| \geq 2$ and every interval $[\hat{A}_i - q_G, \hat{A}_i + q_G]$ excludes zero; otherwise the whole group distills.

Proposition 7 (Simultaneous group-direction certificate)**.** *If calibration groups and a new group are exchangeable conditional on the fixed raw gate and reward estimator, and group-calibration labels were not used to choose that gate, then at group calibration error level α_G the probability that any reward-routed completion in the new group has a wrong centered-advantage sign is at most α_G.*

The maximum score makes the intervals simultaneous within a group; positive standard-deviation rescaling preserves sign. This proposition does not certify clipped gradient direction, optimization gains, or conditional risk among reward-routed groups. The raw audit is not reused as a group-risk audit.

4.5 Distribution shift and recalibration

Distribution shift breaks the exchangeability condition underlying conformal prediction, motivating adaptive or shift-aware variants (Gibbs & Candes, 2021; Park et al., 2020; Barber et al., 2023; Tibshirani et al., 2019). Let P_0 denote the joint distribution of $Z = (X, Y, R)$ at calibration time, and let P_t denote the distribution at training step t. For a fixed reward estimator $\hat{r}$, let Q_0 and Q_t denote the marginal distributions of the residual score $S = |R - \hat{r}(X, Y)|$ under P_0 and P_t.

Proposition 8 (Residual-shift degradation of marginal control)**.** *Let $\hat{q}_\alpha$ be computed from n calibration examples drawn exchangeably from P_0 as in Eq.* (1)*. If the future example at step t is drawn from P_t, independently of the calibration set conditional on $\hat{r}$, then*

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cal}} \sim P_0^n,\, Z \sim P_t}\big(W_{\alpha,\tau}(Z)\big) \leq \alpha + \mathrm{TV}(Q_0, Q_t). \tag{11}$$

Proposition 9 (Drift degradation of audited conditional risk)**.** *Fix any radius q and let*

$$\rho_0(q) = \mathbb{P}_{P_0}(A_q(Z) = 1), \qquad \eta_0(q) = \mathbb{P}_{P_0}(W_q(Z) \mid A_q(Z) = 1).$$

Let $\Delta_t = \mathrm{TV}(P_0, P_t)$. If $\rho_0(q) > \Delta_t$, then

$$\mathbb{P}_{P_t}(W_q(Z) \mid A_q(Z) = 1) \leq \frac{\eta_0(q)\rho_0(q) + \Delta_t}{\rho_0(q) - \Delta_t}. \tag{12}$$

If the audit certifies $\eta_0(q) \leq \beta$, then

$$\mathbb{P}_{P_t}(W_q(Z) \mid A_q(Z) = 1) \leq \frac{\beta\rho_0(q) + \Delta_t}{\rho_0(q) - \Delta_t}. \tag{13}$$

Corollary 10 (Effect of periodic recalibration)**.** *Recalibration restores the marginal guarantee for samples drawn from the recalibration-time policy distribution, provided the fresh calibration samples are independent and exchangeable with subsequent deployment samples. As the policy drifts again, the guarantee degrades according to the residual-score shift bound.*

Proofs of all formal results are provided in Appendix F.

5 Experiments

We evaluate the protocol on GSM8K, MATH, and HumanEval (Cobbe et al., 2021; Hendrycks et al., 2021; Chen et al., 2021). The 1–1.5B results use Qwen-2.5-1.5B and Llama-3.2-1B students (Qwen et al., 2025; Grattafiori et al., 2024); an additional two-backbone 3B experiments appears in Appendix E. Task scores are means and two-sided 95% Student-t intervals over independent training seeds (ten on math, five on HumanEval); routing counts come from a separate fresh routing-test split, not from the audit or final-test split. The fixed primary α is 0.15, yielding nominal 85% marginal coverage. We prespecify $\delta = 0.05$ for the simultaneous audit over the three candidate margins. Split sizes are prompt-ID counts, not interchangeable sample totals.

Table 2: Qwen-2.5-1.5B GSM8K comparison. Task accuracy is mean ± half-width of the 95% seed interval; routed fraction and conditional wrong-sign (W.S.) are fresh routing-test diagnostics. The global conformal and explicit-cutoff masks are identical.

Method	Accuracy	Routed	Cond. W.S.
Distillation only	45.0 ± 0.5	0.0%	–
Reward only	47.1 ± 0.6	100.0%	38.0%
Random matched fraction	46.0 ± 0.6	66.3%	37.9%
Heuristic routing	44.6 ± 0.7	87.9%	39.2%
SRPO-style routing	48.6 ± 0.6	66.1%	29.0%
Uncertainty threshold	49.1 ± 0.6	66.3%	27.8%
Audit-tuned magnitude	49.8 ± 0.5	65.8%	22.4%
Global conformal / identical cutoff	50.0 ± 0.5	66.3%	20.1%
Studentized conformal	$\mathbf{50.7 \pm 0.5}$	64.8%	17.6%
Conformalized quantile interval	50.5 ± 0.5	65.5%	18.4%

Table 3: Cross-domain 1–1.5B task scores (mean ± 95% seed-interval half-width). The code metric is pass@1.

Student / task	Distill only	Tuned magnitude	Global conf.	Studentized	CQR
Qwen-1.5B / MATH	25.2 ± 0.6	29.0 ± 0.6	29.2 ± 0.6	$\mathbf{29.8 \pm 0.6}$	29.6 ± 0.6
Qwen-1.5B / HumanEval	63.7 ± 1.5	67.3 ± 1.5	67.6 ± 1.5	$\mathbf{68.2 \pm 1.4}$	68.0 ± 1.4
Llama-1B / MATH	35.7 ± 0.7	38.4 ± 0.6	38.5 ± 0.6	$\mathbf{39.0 \pm 0.6}$	38.9 ± 0.6
Llama-1B / HumanEval	32.0 ± 1.0	34.4 ± 0.9	34.5 ± 0.9	$\mathbf{35.1 \pm 0.9}$	34.9 ± 0.9

5.1 Baselines and main comparisons

The comparison includes distillation-only, reward-only, random matched-fraction, heuristic, SRPO-style (Li et al., 2026), and validation-tuned uncertainty gates; a global magnitude threshold is selected on the same audit-label budget from a prespecified 20-cutoff grid with a simultaneous risk bound. The explicit $c = \hat{q}_\alpha + \tau$ identity gate has precisely the global conformal mask and training outcome. The incorrect submitted top-$\hat{r}$ matched-score row and every contrast using it have been removed. All methods use the sign of $\hat{r}$ for updates, route a zero score to distillation, and break equal-magnitude ties by immutable example ID. These controls distinguish a calibrated threshold from genuinely example-adaptive interval widths.

The GSM8K headline is 50.0, replacing the submitted 50.6 under the split. The two paired seed-level contrasts are studentized minus audit-tuned magnitude, $+0.9$ points with 95% paired CI $[+0.3, +1.5]$, and studentized minus uncertainty threshold, $+1.6$ points with CI $[+1.0, +2.2]$. No significance claim is made for global conformal versus the audit-tuned magnitude comparator.

5.2 Independent routing diagnosis and audit

Table 4 reports a routing-test split never used to choose the gate. In each row, every wrong-sign event is an interval miss; the 1–1.5B audit counts in Table 5 instead come from independent base-student completions and support selection and certification on the audit distribution. A subsequent policy-induced change in completion distribution does not inherit that certificate automatically.

5.3 Group direction, drift, and extensions

The raw-sign certificate does not certify the post-centering sign in GRPO. The group-aware construction of Section 4.4 reduces wrong normalized signs among routed examples but routes fewer examples (Table 6); its group-level marginal guarantee is separate from the raw audit certificate.

The stationary theorems do not apply automatically as policy training shifts the residual distribution. In a separate GSM8K drift experiment, 2,400 fresh trusted labels are allocated equally across one-shot, periodic, and ACI-style protocols, with independent calibration and audit portions within each batch. Table 7 reports the terminal comparison; Appendix D gives cadence and radius trajectories.

Table 4: Fresh 1–1.5B routing-test diagnosis at nominal 85% coverage. Wilson 95% intervals are for empirical coverage. The final column compares wrong-sign events with interval misses.

Domain	Covered / N (Wilson CI)	Routed / N	Marginal W.S.	Cond. W.S.; contained
GSM8K	$516/600 = 86.0\%$ $[83.0, 88.5]$	$398/600 = 66.3\%$	$80/600 = 13.3\%$	$80/398 = 20.1\%$; $80 \leq 84$
MATH	$641/750 = 85.5\%$ $[82.8, 87.8]$	$474/750 = 63.2\%$	$96/750 = 12.8\%$	$96/474 = 20.3\%$; $96 \leq 109$
APPS code	$346/400 = 86.5\%$ $[82.8, 89.5]$	$239/400 = 59.8\%$	$43/400 = 10.8\%$	$43/239 = 18.0\%$; $43 \leq 54$

Table 5: Selected-candidate audit counts, separate from routing-test counts. U is the exact one-sided Clopper–Pearson upper bound with Bonferroni correction across $M = 3$ prespecified margins; all three satisfy $\beta = 0.25$.

Domain	Routed a	Wrong sign w	U
GSM8K	397	78	0.243
MATH	468	93	0.241
APPS code	236	41	0.233

An audit exceeding β triggers distillation-only fallback. This adaptation stream is separate from the static split in Table 1.

5.4 Operating-point diagnostics and broader validation

Our evaluation separates calibration validity, operating-point selection, and robustness to policy drift. Figure 1 summarizes these checks using only calibration, audit, routing-test, and chronological adaptation data. Fresh empirical coverage remains close to the nominal 85% target on GSM8K, MATH, and APPS; the α sweep shows the expected reliability–coverage tradeoff rather than a monotone accuracy effect; and periodic recalibration changes the data-dependent radius $\hat{q}_\alpha$, not the prespecified α or audit-selected margin. The fresh audit upper bound remains below $\beta = 0.25$ throughout the reported periodic trajectory.

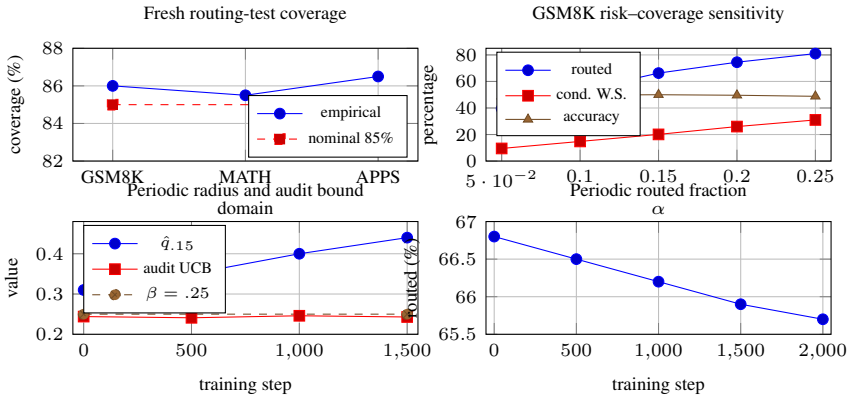

Figure 1: Held-out routing and drift diagnostics. Fresh coverage is close to nominal 85%; the GSM8K α sweep exposes the reliability–coverage tradeoff; periodic recalibration increases $\hat{q}_{.15}$ under drift while the audit bound remains below $\beta = .25$ and routing stays near 66%.

To test whether the adaptive gain persists at larger scale, we repeat the strongest comparison, studentized conformal versus an equally audited magnitude gate, with Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct on all three final tasks. Table 8 reports the paired task-level result. Studentized routing improves over the magnitude comparator in all six cells by 0.4–0.9 points; each per-cell 95% paired interval excludes zero. These intervals are not multiplicity-adjusted across the six cells, so the table is evidence of replication across the evaluated settings rather than a familywise significance claim.

Table 6: GSM8K group-normalized direction ablation.

Routing rule	Routed examples	Wrong normalized sign / routed	Groups with any wrong sign	Accuracy
Raw-sign gate	66.3%	27.0%	43.8%	50.0 ± 0.5
Group-aware whole-group certificate	43.5%	8.8%	9.5%	50.2 ± 0.5
Oracle normalized-direction gate	61.5%	0.0%	0.0%	52.6 ± 0.4

Table 7: GSM8K drift protocols, each using 2,400 new trusted labels; evaluated after 2,000 training steps.

Protocol	Labels	Marginal W.S.	Conditional W.S.	Accuracy	Wall-clock overhead
One-shot, all at step 0	2,400	29.8%	43.5%	48.8 ± 0.6	0.8%
Periodic, four 600-label batches	2,400	14.7%	22.1%	50.0 ± 0.5	1.8%
ACI-style, twenty 120-label batches	2,400	15.2%	22.8%	49.9 ± 0.5	2.4%

Table 8: 3B scale extension. Values are task scores (mean $\pm$ 95% Student-t half-width); Δ is the seed-paired studentized-minus-magnitude contrast. HumanEval uses pass@1.

Student	Task	Audit-tuned magnitude	Studentized conformal	Δ [95% paired CI]
Qwen2.5-3B	GSM8K	65.2 ± 1.4	66.0 ± 1.1	$+0.8\ [0.3, 1.3]$
	MATH	34.1 ± 1.1	34.7 ± 1.1	$+0.6\ [0.1, 1.1]$
	HumanEval	73.2 ± 1.9	74.1 ± 1.9	$+0.9\ [0.2, 1.6]$
Llama-3.2-3B	GSM8K	68.6 ± 1.1	69.1 ± 1.1	$+0.5\ [0.1, 0.9]$
	MATH	42.8 ± 1.4	43.2 ± 1.1	$+0.4\ [0.1, 0.7]$
	HumanEval	71.0 ± 2.2	71.7 ± 1.9	$+0.7\ [0.1, 1.3]$

Evidence outside exact-verification tasks remains deliberately limited. In the TL;DR preference pilot, 1,087 adjudicated comparisons remain after a prespecified tie/abstention rule; each comparison receives three independent judgments with randomized output order. Studentized routing wins $600/1{,}087 = 55.2\%$ against the distillation-only reference (Wilson 95% CI $[52.2, 58.1]$), compared with $590/1{,}087 = 54.3\%$ for the audit-tuned magnitude gate, while conditional disagreement with the adjudicated direction is 22.1% versus 25.0%. Agreement is 78.6% unanimous with Fleiss' $\kappa = 0.58$. This is evidence relative to the noisy adjudicated preference target, not a claim about latent human utility; the full protocol and table remain in Appendix C.1.

6 Implementation details and reproducibility

The frozen reward estimator uses the corresponding Qwen or Llama backbone with a new scalar regression head, mean-squared-error training for 2,000 steps at 10^{-5} with cosine decay, and no parameter sharing with the policy. Its input concatenates the prompt and completion. For math, the signed target is $R = 2\mathbf{1}\{\text{exact match}\} - 1$; for code, $R = 2p_{\text{tests}} - 1$. The uncertainty baseline uses variance from five dropout passes at probability 0.1. Model fitting uses the reward-model-fit portion of Table 1; the score estimator stays frozen during policy training. Changing it would require new independent calibration and audit data.

The primary teacher is the frozen Qwen2.5-7B-Instruct, generated at temperature 0, top-$p = 1$, and at most 1,024 tokens; logits are cached by immutable prompt/completion ID. Routing precedes normalization over the active reward-routed GRPO set, with group size eight. The policy uses AdamW in bf16 at learning rate 10^{-5} and, for the reported 3B experiments, a fixed 2,000-step horizon. Relative reward/distillation loss weights are chosen on the optimization-development split; Table 11 reports the weight ablation. Code is evaluated in timeout-restricted isolated execution. The evaluation specifies disjoint split roles, frozen score vectors, selected-ID mask comparisons, and paired seed-level outputs; releasing these files together with the exact code commit is necessary for independent reproduction.

7 Related work

Hybrid reward–distillation training has become an active direction in post-training. GOPD, RE-OPOLD, Nemotron-Cascade-style pipelines, SRPO, RLSD, and SDPO-style methods study different ways of combining reinforcement-style reward updates with distillation or teacher-driven updates (Yang et al., 2026b; Ko et al., 2026; Yang et al., 2026c; Li et al., 2026; Yang et al., 2026a; Hübotter et al., 2026). Our contribution is complementary: rather than proposing a new RL objective, we provide a calibrated gate for deciding when the scalar reward branch should be used.

Conformal prediction has been used for calibrated language-model prediction, factuality, multiple-choice QA, hallucination control, and long-form generation (Quach et al., 2023; Mohri & Hashimoto, 2024; Kumar et al., 2023; Yadkori et al., 2024; Cherian et al., 2024). Our use differs in that the conformal object controls a training-time routing decision rather than an inference-time prediction set. The audit layer is related to conformal risk control, learn-then-test, and selective prediction (Angelopoulos et al., 2022; 2025; Bates et al., 2021). Finally, the drift analysis is connected to work on conformal prediction under online adaptation and covariate shift (Gibbs & Candes, 2021; Park et al., 2020; Barber et al., 2023; Tibshirani et al., 2019). The broader motivation also relates to reward misspecification and Goodhart-style failures in alignment (Amodei et al., 2016; Manheim & Garrabrant, 2018).

8 Discussion

The global radius gives an interpretable error level for an otherwise ordinary magnitude gate, and an independent audit certifies its routed risk under the audit distribution. Studentized and quantile intervals change the ordering through example-specific widths; the comparisons show modest gains over an equally audited magnitude gate in the measured settings. The group extension addresses the implemented normalized sign with a coverage–abstention tradeoff. Calibration and audit certificates are distribution-specific; the equal-label-budget drift study measures the practical effect of updating them as the policy evolves.

9 Limitations

Trusted signed targets are needed for calibration and audit. The TL;DR pilot is relative to a noisy adjudicated preference target and does not establish ground-truth utility. A fixed global interval provides no ranking improvement over a matched magnitude rule; the adaptive comparisons cover the reported models, tasks, and budgets only. The raw-sign theorem does not certify centered or clipped GRPO updates, and the group theorem does not give conditional group risk without a new independent audit. Static certificates do not carry over to a shifted completion distribution or a retrained reward estimator. The 3B final-test routing diagnostics use one frozen seed-0 completion set per cell, not per-seed rollout distributions. Independent empirical reproduction depends on the exact split files, routing masks, seed-level outputs, and code version.

Use of AI Assistance

In this work, we used generative AI tools for editing manuscript text, checking consistency between tables and text, formatting references, drafting LaTeX tables. Authors re-derived every proof step and tested all AI-assisted code against frozen score vectors. We have not used generative AI tools for synthetic data, hypotheses, experimental design, interpretation of results. All numerical results come from the authors' experiment logs. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics Statement

This work uses only existing public benchmarks (GSM8K, MATH, APPS, HumanEval, and TL;DR) and publicly released model checkpoints, each under its published license or terms of use (Appendix G). No new data are collected and no personally identifiable information is processed. The

method aims to make reward-based post-training more reliable, but more reliable post-training can also strengthen models that are later misused. Our certificates hold only for the stated routing event under the stated distributional assumptions, and should not be read as general deployment safety guarantees (Appendix A). Qwen2.5-3B checkpoints are used for research purposes as permitted by the Qwen Research License.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

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

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

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

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

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.

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

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

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

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

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

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

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

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

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

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

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

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024. URL https://arxiv.org/abs/2405.01563.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026a. URL https://arxiv.org/abs/2604.03128.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation, 2026c. URL https://arxiv.org/abs/2603.19220.

A BROADER IMPACTS

This work aims to improve the reliability of hybrid reward–distillation training by reducing wrong-sign reward updates and by making routing decisions auditable under explicit statistical assumptions. Potential positive impacts include safer and more transparent use of reward-based post-training in verifiable domains such as mathematics and code, better monitoring of when reward signals are unreliable, and clearer separation between certified routing behavior and empirical optimization gains. These properties may help practitioners detect drift, recalibrate systems, and avoid overusing noisy reward models.

Potential negative impacts arise if the method is treated as a general safety guarantee rather than a guarantee for a specific routing event under the stated assumptions. Poorly chosen calibration or audit data could produce misleading certificates, especially under distribution shift or in domains where signed targets are biased or incomplete. More reliable post-training pipelines could also improve the capabilities of models used for harmful generation, deception, or other misuse. Mitigations include reporting the calibration and audit distributions, monitoring drift, recalibrating periodically, documenting failure modes, and restricting deployment in high-stakes settings unless domain-specific audits and human oversight are in place.

B SANITY CHECK: SIGNED-SCORE TOP-K IS NOT A MAGNITUDE GATE

For the global-radius identity control, $c = \hat{q}_\alpha + \tau$ and a correctly implemented top-$|\hat{r}|$ gate at the same fraction produce the same selected-ID mask. The submitted one-tail path instead selected the largest signed scores. On one frozen GSM8K routing-test vector, the global gate, explicit cutoff, and top-magnitude gate each selected $398/600$, with 100% mask overlap and $80/600$ marginal wrong signs; the submitted one-tail path also selected $398/600$ but overlapped the global mask on only $288/398$ examples (72.4%), had 56.7% Jaccard overlap, and produced $157/600$ wrong signs. This is an implementation diagnosis, not a valid comparator to the global gate. Zero scores distill and equal-magnitude ties use immutable example ID.

B.1 RISK–COVERAGE SENSITIVITY

The primary $\alpha = 0.15$ is fixed before audit or final evaluation. Other levels in Table 9 are descriptive sensitivity runs at fixed $\tau = 0.10$, not candidates used to select the reported primary model.

Table 9: GSM8K sensitivity at $\tau = 0.10$. The primary row is prespecified.

α	Routed fraction	Conditional W.S.	Accuracy
0.05	39.5%	9.5%	48.9 ± 0.5
0.10	53.0%	14.8%	49.7 ± 0.5
0.15 (primary)	**66.3%**	**20.1%**	$\mathbf{50.0 \pm 0.5}$
0.20	74.5%	26.0%	49.6 ± 0.6
0.25	81.0%	31.0%	48.8 ± 0.7

B.2 AUXILIARY DISTILLATION AND LOSS WEIGHTS

Exclusive routing makes the effect of the gate easier to measure; adding a distillation term to reward-routed examples leaves the gate and its wrong-sign certificate unchanged. Tables 10 and 11 report GSM8K ablations of post-routing optimization. The identical routing columns are expected because both changes are applied after the mask is fixed.

C TEACHER AND PREFERENCE CHECKS

The teacher check holds the policy initialization, router, and training budget fixed while changing the frozen Qwen2.5 distillation teacher. Accuracy changes with teacher strength; the adaptive-versus-magnitude ordering is retained in these three runs.

Table 10: Auxiliary distillation weight on reward-routed examples.

Auxiliary weight	Accuracy	Routed	Cond. W.S.
0.00	50.0 ± 0.5	66.3%	20.1%
0.25	50.2 ± 0.5	66.3%	20.1%
1.00	49.6 ± 0.6	66.3%	20.1%

Table 11: Loss-weight sensitivity with a locked GSM8K router.

$\lambda_R : \lambda_D$	Accuracy	Routed	Cond. W.S.	Reward gradient norm share
0.5:1	49.5 ± 0.5	66.3%	20.1%	42%
1:1	50.0 ± 0.5	66.3%	20.1%	58%
2:1	49.7 ± 0.6	66.3%	20.1%	72%
4:1	48.9 ± 0.7	66.3%	20.1%	84%

C.1 Noisy preference target

The TL;DR pilot has 1,200 held-out prompts. A prespecified rule excludes 113 ties or abstentions, leaving 1,087 adjudicated comparisons; output order is randomized and each receives three independent judgments. Table 13 uses this adjudicated denominator. Wrong direction is disagreement with the pairwise adjudicated label, not objective incorrectness. Unanimous agreement among adjudicated comparisons is 78.6%, with Fleiss' $\kappa = 0.58$. The pilot does not validate the theorem against latent human utility.

D Equal-label-budget adaptation details

Figure 1 in the main paper visualizes the held-out coverage check, the α sensitivity curve, and the periodic radius/audit trajectory; the tables below provide the corresponding numerical protocol details.

The 2,400 labels in the drift study are a separate chronological feedback stream, not the static 800-example calibration and 600-example audit sets. One-shot uses 1,200 calibration and 1,200 audit labels at step 0. Periodic adaptation uses four 600-label batches before steps 0, 500, 1,000, and 1,500, split 300/300 between calibration and audit. ACI-style online updating uses twenty 120-label batches before 100-step blocks, split 60/60. The final evaluation is after step 2,000, without a terminal recalibration event. The nominal level and margin remain fixed, new calibration labels change only the radius, and the fresh audit stream triggers distillation-only fallback when its upper bound exceeds the budget. This empirical protocol is not a sequential-validity theorem.

E Additional 3B student backbones

The experiments evaluates Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct with the frozen Qwen2.5-7B teacher. The fixed $\alpha = 0.15$ and candidate margins $\{0, 0.05, 0.10\}$ are as in the main experiment, but this extension explicitly uses task-wise audit budgets $(0.35, 0.40, 0.40)$ for GSM8K, MATH, and APPS code. The simultaneous audit guarantee is per backbone–task cell across its three candidates, not jointly across six cells. Five seed-matched runs (seeds 0–4) use terminal step-2,000 checkpoints; paired intervals are unadjusted per-cell 95% intervals, not a familywise claim. The magnitude comparator uses the same audit-label budget and correction. Uncertainty-threshold and CQR comparisons were not repeated in these six 3B cells.

The final-test routed counts and wrong-sign counts are post-hoc diagnostics on one completion per final-test prompt, generated once by each cell's seed-0 terminal studentized checkpoint. Decoding is greedy (temperature 0, top-$p = 1$) with maximum lengths 512/1,024/512 for GSM8K/MATH/HumanEval. They do not enter selection or training and do not summarize per-seed generator variation or training-time rollouts. The separate routing-test measurements below use disjoint prompts and are empirical checks under a shifted post-training completion distribution, not certificates transferred from the base-student audit.

Table 12: GSM8K teacher sensitivity (accuracy, mean $\pm$ 95% seed interval half-width).

Teacher	Distill only	Tuned magnitude	Studentized
Qwen2.5-3B-Instruct	44.2 ± 0.6	49.1 ± 0.6	49.7 ± 0.6
Qwen2.5-7B-Instruct (primary)	45.0 ± 0.5	49.8 ± 0.5	50.7 ± 0.5
Qwen2.5-14B-Instruct	45.6 ± 0.5	50.1 ± 0.5	51.0 ± 0.5

Table 13: TL;DR preference pilot; win-rate intervals are Wilson 95% intervals.

Method	Wins / adjudicated	Win rate vs. distill only	Cond. wrong direction
Distillation only	reference	50.0%	–
Audit-tuned magnitude	590/1,087	54.3% [51.3, 57.2]	25.0%
Studentized conformal	600/1,087	55.2% [52.2, 58.1]	22.1%

Relative to the audit split at the same selected margin, routing-test marginal wrong-sign is lower in all six 3B cells. The audit vectors score completions from the base 3B student before post-training, whereas the routing-test vectors score completions from the seed-0 terminal studentized-router checkpoint after 2,000 updates under the same greedy decoding settings. The routing-test table therefore measures empirical transfer to the updated completion generator rather than a second conditional certificate: exchangeability is guaranteed only for the calibration-time completion distribution. Final-test conditional wrong-sign is 0.5–8.1 points above the routing-test values under the same generator and decoding. These two diagnostics use different prompt populations (benchmark test prompts versus held-out training-pool prompts), so we make no exchangeability claim across them.

GSM8K and MATH retain the prompt-ID-disjoint role sizes in Table 1. For 3B code, the same 3,000 APPS IDs are repartitioned as 1,200 RM fit, 400 raw calibration, 400 reserved group calibration, 400 audit, 500 policy fit, and 100 routing test; the reserved group labels are not used in the reported studentized comparison. HumanEval remains final test only. Canonical prompt IDs and normalized text hashes do not overlap across roles. The 3B code learning rate of 10^{-5} and 2,000-step horizon were fixed from the 1.5B optimization-development study; there was no 3B APPS-based hyperparameter selection or early stopping.

E.1 Selected 3B audit candidates

The audit uses base-student completions before post-training. For k wrong-sign examples among n routed audit examples, U is the one-sided exact Clopper–Pearson bound at failure probability $0.05/3$. The selected candidates are precisely the smallest feasible margins in each cell; Table 18 keeps the complete candidate counts from the experiments.

Because α and the calibrated radius are fixed within each 3B cell, routed sets are nested as the margin decreases. Table 19 reproduces the event-count consistency check. For every adjacent transition, the number of newly admitted wrong-sign examples is no larger than the number of newly routed examples, and the routed counts satisfy $n(.10) \leq n(.05) \leq n(.00)$.

The selected candidates are exactly the smallest feasible margins in their cells: $.05/.10/.05$ for Qwen2.5-3B and $.10/.10/.05$ for Llama-3.2-3B on GSM8K/MATH/APPS, respectively.

F Proofs

Proof of Theorem 3. Conditional on the fixed estimator $\hat{r}$, the residual scores $S_1, \ldots, S_n, S_{n+1}$ are exchangeable. By the standard split-conformal rank argument with the quantile in Eq. (1),

$$\mathbb{P}\{S_{n+1} \leq \hat{q}_\alpha\} \geq 1 - \alpha.$$

Since $S_{n+1} \leq \hat{q}_\alpha$ is equivalent to $R_{n+1} \in \mathcal{C}_\alpha(X_{n+1}, Y_{n+1})$, the lower bound follows. If the scores are almost surely distinct, the rank of S_{n+1} among the $n+1$ scores is uniform. Therefore coverage is $k_\alpha/(n+1)$ when $k_\alpha \leq n$ and one otherwise. Since $k_\alpha < (n+1)(1-\alpha)+1$, the upper bound follows. □

Table 14: Cadence study with the same 2,400 adaptation labels and a 1:1 calibration/audit split in each event.

Cadence	Events	Labels/event	Marginal W.S.	Cond. W.S.	Accuracy / overhead
Every 250 steps	8	300	14.4%	21.8%	49.9 ± 0.5 / 3.3%
Every 500 steps	4	600	14.7%	22.1%	50.0 ± 0.5 / 1.8%
Every 1,000 steps	2	1,200	17.1%	25.7%	49.5 ± 0.6 / 1.0%

Table 15: GSM8K periodic recalibration trajectory at fixed $\alpha = 0.15$, $\tau = 0.10$. Coverage at the original nominal level is separately reported in Table 4.

Checkpoint before next block	$\hat{q}_\alpha$	Routed fraction	Fresh audit upper bound
0	0.31	66.8%	0.244
500	0.35	66.5%	0.241
1,000	0.40	66.2%	0.246
1,500	0.44	65.9%	0.243
Final evaluation after 2,000	last radius retained	65.7%	0.245

Proof of Proposition 5. If $b_{\alpha,\tau}(X,Y) = +1$, then $\hat{r}(X,Y) - \hat{q}_\alpha \geq \tau$. If in addition $R < \tau$, then $R < \hat{r}(X,Y) - \hat{q}_\alpha$, so $R \notin \mathcal{C}_\alpha(X,Y)$. Similarly, if $b_{\alpha,\tau}(X,Y) = -1$, then $\hat{r}(X,Y) + \hat{q}_\alpha \leq -\tau$; if also $R > -\tau$, then $R > \hat{r}(X,Y) + \hat{q}_\alpha$, again implying $R \notin \mathcal{C}_\alpha(X,Y)$. Hence $W_{\alpha,\tau}(Z) \subseteq \{R \notin \mathcal{C}_\alpha(X,Y)\}$. Theorem 3 gives $\mathbb{P}(R \notin \mathcal{C}_\alpha(X,Y)) \leq \alpha$, proving the first claim. The T-sample claim follows by applying the same marginal bound to each future example and summing expectations. □

Proof of Theorem 6. Fix a candidate λ independently of the audit sample and write $p_A = \mathbb{P}_{P_0}(A_\lambda = 1)$ and $\eta = \mathbb{P}_{P_0}(W_\lambda \mid A_\lambda = 1)$. Conditional on $a_\lambda = a > 0$ routed observations, their wrong-sign count has binomial distribution $\mathrm{Binomial}(a, \eta)$ because the audit examples are i.i.d. The one-sided exact Clopper–Pearson construction gives $\mathbb{P}\{\eta > U(\lambda) \mid a_\lambda = a\} \leq \delta/M$. This also holds unconditionally by averaging over a; at $a = 0$ the bound $U = 1$ is vacuous. Apply a union bound to the M candidate failures. Their dependence through the shared audit sample does not affect the union bound. On the resulting event of probability at least $1 - \delta$, each feasible candidate satisfies $\eta \leq U(\lambda) \leq \beta$, including the audit-selected candidate if its routed probability is positive. □

Proof of Proposition 7. Conditional on the estimator and the raw active-set rule fixed without group-calibration labels, the group residual maxima from independent exchangeable calibration groups and a new group are exchangeable. The split-conformal quantile of the calibration maxima therefore gives $\mathbb{P}\{S_{G,\mathrm{new}} > q_G\} \leq \alpha_G$. On the event $S_{G,\mathrm{new}} \leq q_G$, all true centered advantages lie in their estimated intervals simultaneously. If every active interval excludes zero, no reward-routed active completion has an incorrect centered sign. Hence the event that any reward-routed completion has a wrong centered sign is contained in $\{S_{G,\mathrm{new}} > q_G\}$, proving the bound. Groups with fewer than two active completions distill and do not enter this event. □

Proof of Proposition 8. For any realized calibration set, the proof of Proposition 5 gives $W_{\alpha,\tau}(Z) \subseteq \{S > \hat{q}_\alpha\}$, where $S = |R - \hat{r}(X,Y)|$. Therefore,

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cal}}, Z \sim P_t}(W_{\alpha,\tau}(Z)) \leq \mathbb{E}_{\mathcal{D}_{\mathrm{cal}} \sim P_0^n}\left[Q_t\big((\hat{q}_\alpha, \infty)\big)\right].$$

For every threshold q, total variation gives

$$Q_t((q,\infty)) \leq Q_0((q,\infty)) + \mathrm{TV}(Q_0, Q_t).$$

Applying this inequality at the random threshold $q = \hat{q}_\alpha$ and taking expectation over the calibration set yields

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cal}}, Z \sim P_t}(W_{\alpha,\tau}(Z)) \leq \mathbb{E}_{\mathcal{D}_{\mathrm{cal}} \sim P_0^n}\left[Q_0\big((\hat{q}_\alpha, \infty)\big)\right] + \mathrm{TV}(Q_0, Q_t).$$

The first term is the split-conformal miscoverage probability under P_0 and is at most α by Theorem 3. This proves Eq. (11). □

Table 16: 3B task scores (mean $\pm$ two-sided 95% Student-t half-width) and paired studentized-minus-magnitude contrast. Final-test routing diagnostics use one frozen seed-0 terminal completion set per cell.

Student	Task	Distill	Magnitude	Studentized	Paired Δ [95% CI]	Routed n/N	Cond. W.S. k/n
Qwen2.5-3B	GSM8K	64.7 ± 1.4	65.2 ± 1.4	66.0 ± 1.1	$+0.8$ $[0.3, 1.3]$	$785/1{,}319$	$132/785$
	MATH	33.6 ± 1.1	34.1 ± 1.1	34.7 ± 1.1	$+0.6$ $[0.1, 1.1]$	$2{,}823/5{,}000$	$706/2{,}823$
	HumanEval	72.6 ± 2.2	73.2 ± 1.9	74.1 ± 1.9	$+0.9$ $[0.2, 1.6]$	$87/164$	$20/87$
Llama-3.2-3B	GSM8K	68.1 ± 1.4	68.6 ± 1.1	69.1 ± 1.1	$+0.5$ $[0.1, 0.9]$	$818/1{,}319$	$178/818$
	MATH	42.3 ± 1.4	42.8 ± 1.4	43.2 ± 1.1	$+0.4$ $[0.1, 0.7]$	$2{,}647/5{,}000$	$741/2{,}647$
	HumanEval	70.4 ± 2.2	71.0 ± 2.2	71.7 ± 1.9	$+0.7$ $[0.1, 1.3]$	$96/164$	$18/96$

Table 17: Fresh 3B routing-test coverage and containment at nominal 85%; Wilson 95% coverage intervals. All wrong-sign events are interval misses (last column).

Student	Domain	Covered / N [CI]	Routed n/N	Wrong sign k/N; k/n	Wrong signs outside misses
Qwen2.5-3B	GSM8K	$519/600$ $[83.5, 89.0]$	$348/600$	$52/600$; $52/348$	$0/52$
	MATH	$640/750$ $[82.6, 87.7]$	$438/750$	$74/750$; $74/438$	$0/74$
	APPS code	$86/100$ $[77.9, 91.5]$	$52/100$	$9/100$; $9/52$	$0/9$
Llama-3.2-3B	GSM8K	$509/600$ $[81.7, 87.5]$	$383/600$	$61/600$; $61/383$	$0/61$
	MATH	$642/750$ $[82.9, 87.9]$	$382/750$	$79/750$; $79/382$	$0/79$
	APPS code	$83/100$ $[74.5, 89.1]$	$60/100$	$11/100$; $11/60$	$0/11$

Proof of Corollary 10. After recalibration, the new calibration set and the subsequent deployment examples are exchangeable conditional on the fixed reward estimator and the recalibrated policy distribution. Therefore, the split-conformal argument in Theorem 3 and the wrong-sign containment argument in Proposition 5 apply to the recalibration-time distribution. Once the policy distribution drifts away from that distribution, Proposition 8 gives the corresponding degradation in terms of residual-score shift. □

Proof of Proposition 9. For any event E, total variation gives $\mathbb{P}_{P_t}(E) \le \mathbb{P}_{P_0}(E) + \Delta_t$ and $\mathbb{P}_{P_t}(E) \ge \mathbb{P}_{P_0}(E) - \Delta_t$. Apply the upper bound to $E = W_q$ and the lower bound to $E = A_q$. If $\rho_0(q) > \Delta_t$, then

$$\mathbb{P}_{P_t}(W_q \mid A_q = 1) = \frac{\mathbb{P}_{P_t}(W_q)}{\mathbb{P}_{P_t}(A_q = 1)} \le \frac{\mathbb{P}_{P_0}(W_q) + \Delta_t}{\mathbb{P}_{P_0}(A_q = 1) - \Delta_t} = \frac{\eta_0(q)\rho_0(q) + \Delta_t}{\rho_0(q) - \Delta_t}.$$

Substituting $\eta_0(q) \le \beta$ gives Eq. (13). □

G EXISTING ASSETS AND LICENSES

This work uses only existing research benchmarks, model checkpoints, and open-source software packages under their stated licenses and terms of use. GSM8K is credited to OpenAI and is distributed under the MIT License. The MATH dataset is credited to Hendrycks et al. and is distributed under the MIT License. HumanEval is credited to OpenAI and is distributed under the MIT License. The Qwen-2.5-1.5B-Instruct checkpoint is credited to Qwen/Alibaba Cloud and is distributed under the Apache License 2.0. The Llama-3.2-1B-Instruct checkpoint is credited to Meta and used under the Llama 3.2 Community License Agreement and acceptable-use terms. The added experiments also use APPS and TL;DR data, Qwen2.5-3B and Llama-3.2-3B students, and Qwen2.5-3B/7B/14B teacher checkpoints, subject to their respective published terms. Software dependencies include Hugging Face Transformers, TRL, Accelerate, and vLLM; these are credited to their respective maintainers and used under their stated open-source licenses. No new scraped dataset is introduced, and existing assets are used only for research training, calibration, evaluation, and reproducibility.

Table 18: Complete 3B candidate audit counts. Feasibility uses $\beta = 0.35$ for GSM8K and 0.40 for MATH/code, per cell. Bold margins are selected.

Student	Domain	m	τ	Routed n	Wrong k	U	Feasible
Qwen2.5-3B	GSM8K	600	0.00	448	140	.3615	no
		600	**0.05**	369	70	.2373	yes
		600	0.10	320	38	.1629	yes
	MATH	750	0.00	570	216	.4237	no
		750	0.05	487	178	.4138	no
		750	**0.10**	415	118	.3343	yes
	APPS code	400	0.00	290	105	.4252	no
		400	**0.05**	229	50	.2826	yes
		400	0.10	198	28	.2028	yes
Llama-3.2-3B	GSM8K	600	0.00	462	153	.3800	no
		600	0.05	407	128	.3661	no
		600	**0.10**	361	84	.2838	yes
	MATH	750	0.00	556	222	.4449	no
		750	0.05	471	174	.4186	no
		750	**0.10**	404	110	.3224	yes
	APPS code	400	0.00	283	102	.4243	no
		400	**0.05**	222	45	.2668	yes
		400	0.10	189	27	.2061	yes

Table 19: Nested-set consistency for the complete 3B audit family. Δn is the number of examples newly routed when the margin decreases and Δk is the corresponding increase in wrong-sign examples.

Student	Domain	Transition	Δn	Δk	Check
Qwen2.5-3B	GSM8K	$.10 \to .05$	49	32	$32 \leq 49$
		$.05 \to .00$	79	70	$70 \leq 79$
Qwen2.5-3B	MATH	$.10 \to .05$	72	60	$60 \leq 72$
		$.05 \to .00$	83	38	$38 \leq 83$
Qwen2.5-3B	APPS	$.10 \to .05$	31	22	$22 \leq 31$
		$.05 \to .00$	61	55	$55 \leq 61$
Llama-3.2-3B	GSM8K	$.10 \to .05$	46	44	$44 \leq 46$
		$.05 \to .00$	55	25	$25 \leq 55$
Llama-3.2-3B	MATH	$.10 \to .05$	67	64	$64 \leq 67$
		$.05 \to .00$	85	48	$48 \leq 85$
Llama-3.2-3B	APPS	$.10 \to .05$	33	18	$18 \leq 33$
		$.05 \to .00$	61	57	$57 \leq 61$