PIVOT: RISK-CONTROLLED ROUTING OF VERIFIER FEEDBACK FOR RELIABLE POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training relies on verifiers with different reliability, coverage, and cost. Existing routers principally estimate label accuracy, although learning depends on the update induced by feedback. We introduce PIVOT, a sequential training-time router that treats verifier feedback as a policy intervention. It predicts reliability-adjusted directional alignment before acquisition, maps heterogeneous outputs to preference graphs, and gates the resulting optimizer-native update. Across Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct in four domains under a common cost cap, PIVOT reaches 68.4 trusted utility. It improves on common-pool BayesianRouter-H by 1.6 points and uncertainty routing by 2.3 points, with worst-slice utility of 61.1. At the pre-specified $\alpha = 0.05$ budget, held-out gold-graph inconsistency (GG-risk) is 0.047 at 82% coverage. Setup cost amortizes after 37.6K prompts relative to BayesianRouter-H and 67.6K relative to uncertainty routing. Fully stale reuse after policy shift raises GG-risk from 0.048 to 0.082, delimiting the phase-wise guarantee.

1 INTRODUCTION

Verification has become part of the learning algorithm. Modern post-training draws rewards from unit tests, exact-answer checkers, learned reward models, rubric-guided judges, and retrieval-backed critics. Each source occupies a different reliability–coverage–cost regime. Programmatic checks provide strong supervision when correctness is executable; open-ended tasks require learned or generative judgments whose errors can be amplified by optimization. The central decision is which feedback to acquire and admit into the current policy update.

Prior work addresses complementary parts of this decision. Calibration and conservative ensembles improve reward robustness, conformal selection controls evaluator-side risk, Uncertainty Router forwards uncertain pairs to a stronger judge, and BayesianRouter adapts reward-model selection to policy drift (Leng et al., 2025; Coste et al., 2024; Gui et al., 2024; Overman & Bayati, 2025; Badshah et al., 2026; Xu et al., 2025b; Wu & Lu, 2026). These methods estimate label confidence, evaluator competence, or output risk. The downstream update induced in the current policy remains outside their routing criterion.

Endpoint correctness and update usefulness answer different questions. An exact-outcome graph can favor a lucky guess over an incomplete response containing valid reasoning, even when imitating the latter would produce the better update. A partial gold graph can also leave two incorrect responses incomparable, making an additional process-informative edge unverified rather than disproved (Figure 1). PIVOT keeps conservative graph consistency for admission and estimates update usefulness separately. Controlled evidence likewise shows that similarly accurate reward models can induce substantially different optimized policies (Wen et al., 2025). Correctness alone may therefore misrank the interventions that govern learning.

We formulate verifier routing as a sequential training decision. A pre-query router selects one source using information available before acquisition. A deterministic adapter converts its native output to a directed preference graph. The post-query gate then combines calibrated strict graph consistency with optimizer-native update alignment to accept, escalate, or skip. An independent holdout certificate controls consequence-weighted GG-risk among admitted feedback subject to minimum

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.

weighted coverage. Routing and gating consequently remain complementary: the former determines which evidence is purchased, while the latter controls whether the observed evidence enters training.

PIVOT contributes a policy-conditional verifier score evaluated in the trainable LoRA space, a common graph interface for heterogeneous feedback, and a certified sequential acquisition policy. Matched-cost experiments across policy models and domains isolate pre-query routing, reliability, directional alignment, approximation error, and realized policy improvement.

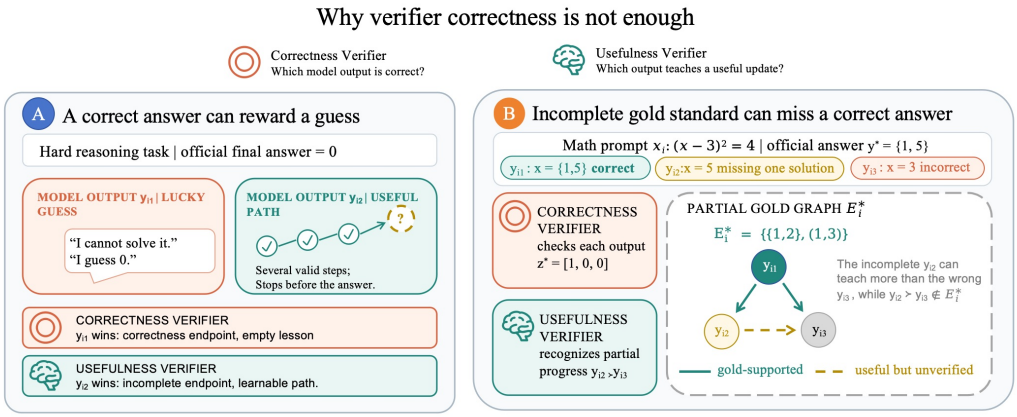


Figure 1: Two cases separate strict gold-graph consistency from useful supervision. An endpoint checker can prefer a correct guess to incomplete yet useful reasoning (A), while a partial gold graph can reject an unverified edge between outcome-tied responses (B). The non-contradiction label C^{nc} is used only for mechanism sensitivity; certification retains strict consistency C. Online, PIVOT queries one verifier, gates its adapted graph by calibrated consistency and update alignment, and permits one Qwen2.5-72B-Instruct repair. Disjoint folds fit, audit, and certify the sequential policy.

2 RELATED WORK

Feedback for post-training. RLHF, DPO, and GRPO convert feedback into policy updates, while heterogeneous verifiers broaden the supervision pool (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024; Cobbe et al., 2021; Guan et al., 2024; Zhao et al., 2026; Yuan et al., 2026). Process reward models order intermediate reasoning through human labels, rollouts, or tree preferences, although harder reasoning remains challenging (Lightman et al., 2024; Wang et al., 2024a; He et al., 2024; Zheng et al., 2024). Transformations, ensembles, calibration, robust objectives, rubrics, and co-evolving supervisors improve feedback quality or coverage (Wang et al., 2024b; Coste et al., 2024; Leng et al., 2025; Lambert et al., 2025; Malik et al., 2026; Wang et al., 2026c; Viswanathan et al., 2025; Gunjal et al., 2026; Zhang et al., 2026; Hong et al., 2025). Optimizer-aware influence and retrospective update evaluation connect data selection to policy change (Xia et al., 2024; Wang et al., 2026a;b). PIVOT applies this perspective prospectively to verifier acquisition and admission.

Routing and risk control. Inference routers allocate generators, judges, or test-time compute (Ong et al., 2025; Tsiourvas et al., 2025; Qian et al., 2025; Taberner-Miller, 2026; Dughmi et al., 2026; Saliencro et al., 2026). BayesianRouter and Uncertainty Router instead select training-time evaluators (Wu & Lu, 2026; Xu et al., 2025b). Conformal methods control output- or evaluator-side error and extend to adaptive streams and distribution shift (Badshah et al., 2026; Gui et al., 2024; Overman & Bayati, 2025; Xu et al., 2025a; Uddin & Bauer, 2026; Chen et al., 2026; Khosravi & Huo, 2026; Liu et al., 2026). PIVOT adds an optimizer-aware acquisition criterion and a frozen-phase certificate for accepted training feedback. The certificate is intentionally limited to conditional gold-graph inconsistency within a renewed phase.

3 PROBLEM FORMULATION

At phase t, frozen policy π_{θ_t} samples a response group $Y_i = \{y_{ij}\}_{j=1}^{K}$ for prompt x_i. Domain d provides three first-acquisition sources $V_d^{(0)}$ and one escalation-only source e_d; their four outputs form the full matrix on setup folds. The generative escalation source is `Qwen/Qwen2.5-72B-Instruct`, whereas first-stage generative judging uses `google/gemma-2-27b-it`; each uses its own frozen role prompt and inference call. We use *independent* only for this implementation separation, without assuming statistically independent errors. Each queried source emits a native object—test outcomes, scalar scores, pairwise labels, rubric vectors, or evidence critiques. A deterministic, versioned adapter maps it to

$$F_{i,v} = (E_{i,v}, B_{i,v}), \qquad E_{i,v} \subseteq [K] \times [K], \tag{1}$$

where $(j,k) \in E_{i,v}$ means $y_{ij} \succ y_{ik}$ and $B_{i,v}$ records responses vetoed for a critical violation. Exact outcomes induce edges directly; scalar and frozen weighted-rubric totals require a difference of at least the source's validation-scale margin γ_v; pairwise judges require position-swapped agreement and confidence at least γ_v; retrieval critics require a frozen evidence-support margin. Ties add no edge, and an empty graph, cycle, non-unique maximum, or vetoed winner forces abstention. Every margin, rubric weight, prompt, and veto is selected on route validation and fixed before certificate labels are accessed; the supplement gives the full mapping and sensitivity analysis.

A disjoint gold audit produces a partial graph $E_i^\star$ and critical set $B_i^\star$. Audit pairs are recorded as supported, contradicted, or underdetermined; only supported directions enter $E_i^\star$. The certificate conservatively counts a promoted direction absent from this partial graph as failure rather than assuming that an unverified edge is correct. We call the resulting label *strict gold-graph consistency*; it is a one-sided, auditable event rather than a complete semantic judgment of every response pair:

$$C_{i,v} = \mathbb{I}[E_{i,v} \neq \emptyset,\ E_{i,v} \subseteq E_i^\star,\ \mathrm{Win}(E_{i,v}) \cap B_i^\star = \emptyset]. \tag{2}$$

Thus, $C_{i,v} = 1$ only if every promoted direction is supported by the gold graph and no gold-critical response is promoted. The label leaves the quality ordering of underdetermined pairs unresolved. Supplementary Section B defines a non-contradiction sensitivity C^{nc} that does not penalize such edges; this relaxed label is excluded from certification because absence of a contradiction is weaker than positive support. For each domain, a source-disjoint direction fold $A_{d,t}$ turns official or executable gold labels into the differentiable surrogate

$$\begin{aligned} J_{d,t}^\star(\psi) &= -\frac{1}{|A_{d,t}|} \sum_{i \in A_{d,t}} L_{\mathrm{aud}}(\psi; F_i^\star), \\ L_{\mathrm{aud}} &= -\frac{1}{|E_i^\star|} \sum_{(j,k) \in E_i^\star} \log \sigma\big(\ell_\psi(y_{ij}|x_i) - \ell_\psi(y_{ik}|x_i)\big), \end{aligned} \tag{3}$$

where ℓ_ψ is length-normalized sequence log-probability and groups with no supported gold edge do not contribute a direction. Gold edges use official answer equivalence for math, hidden tests for code, executable constraints for instruction following, and official answers plus a frozen evidence reference for science. This surrogate constructs only the audit direction; mechanism and terminal claims use disjoint official or executable task utilities. The ideal, unavailable routing target is

$$\Delta_{i,v,t}^\star = J_{d(i),t}^\star(\psi_t + d_{i,v,t}) - J_{d(i),t}^\star(\psi_t), \tag{4}$$

where $d_{i,v,t}$ is the virtual optimizer step induced by $F_{i,v}$. Both the target and rollout distribution depend on the current policy.

4 PIVOT

4.1 CANONICAL FEEDBACK-TO-UPDATE INTERFACE

For a valid graph, each response receives its normalized net number of wins over the other $K-1$ responses. GRPO standardizes these scores within the group before reference correction and clipping; DPO uses each directed edge as a chosen–rejected pair. The same adapter therefore defines both losses. SKIP contributes no loss, and escalation returns a new $F_{i,v}$ through the same adapter. Supplementary Section B gives the exact score.

The primary experiment updates rank-16 LoRA parameters ψ only. With optimizer state frozen at the start of routing, let M_t be the optimizer's preconditioner and define

$$d_{i,v,t} = -\eta M_t \nabla_\psi \mathcal{L}_{\mathrm{pol}}(\psi_t; x_i, Y_i, F_{i,v}). \tag{5}$$

This virtual step includes the actual group normalization, reference term, and clipping configuration; it is not a response-local gradient. The direction fold gives $g^{\mathrm{aud}}_{d,t} = \nabla_\psi J^\star_{d,t}(\psi_t)$ from Equation 3; it never fits or certifies the router. A fixed sparse JL map S_t then yields

$$\tau_{i,v,t} = \cos(S_t g^{\mathrm{aud}}_{d(i),t},\ S_t d_{i,v,t}). \tag{6}$$

We distinguish the unsketched first-order inner product $\langle g, d\rangle$, cosine direction, update norm, JL-estimated direction, and realized minibatch step. PCA-compressed gradients retain useful OOD-detection signal in vision models (Wu et al., 2024); that result motivates compression while leaving LoRA/JL fidelity unvalidated in the present setting.

4.2 Reliability and Sequential Acquisition

A cross-fitted logistic consistency head followed by held-out isotonic calibration estimates $q_{i,v,t} = \Pr(C_{i,v} = 1 \mid \phi^+_{i,v,t})$ from post-query graph margins, uncertainty, veto, and verifier-identity features. Its calibration-bin lower estimate q is transformed as

$$r^{(\beta)}_{i,v,t} = \big[q_{i,v,t} - \beta(1 - q_{i,v,t})\big]_+, \qquad \beta \geq 1. \tag{7}$$

$\beta = 1$ recovers $[2q-1]_+$ under a binary symmetric direction-flip model. For heterogeneous feedback it remains a precision-sensitive ranking margin, not an unbiased correction; β, false acceptance, and false rejection are varied separately.

Crucially, q, F, and τ are unavailable before purchasing v. A cross-fitted histogram gradient-boosted regressor over shared pre-query features and verifier identity therefore predicts the reliability-adjusted directional score

$$m_{i,v,t} = \mathbb{E}\big[r^{(\beta)}_{i,v,t}\tau_{i,v,t} \mid \phi^-_{i,v,t}\big], \qquad b^{(0)}_i = \arg\max_{v \in V^{(0)}_{d(i)}} \big\{\hat{m}_{i,v,t} - \lambda_c \bar{c}_v\big\}, \tag{8}$$

where $\bar{c}_v \in [0,1]$ is expected incremental cost normalized by the most expensive first-acquisition source. The pre-query feature set excludes every current verifier output and every gradient derived from one. The score m is not an estimator of $\Delta^\star$: it omits update magnitude, curvature, minibatch interaction, and clipping after aggregation. We therefore compare it with norm-aware first-order and directly learned one-step-improvement predictors.

The complete sequential candidate is indexed abstractly by

$$\lambda = (\lambda_c, \gamma_q, \gamma_\tau, e) \in \Lambda, \tag{9}$$

where Λ is fixed before certificate labels are accessed. The four coordinates do not form a four-threshold gate: λ_c affects only pre-query acquisition cost, γ_q and γ_τ are post-query admission thresholds, and e permits at most one call to e_d after initial rejection. Supplementary Section C gives the Cartesian grid and the phase-wise selected values. For a queried source,

$$G^\lambda_{i,v} = \mathbb{I}\big[\mathrm{valid}(F_{i,v}),\ q_{i,v,t} \geq \gamma_q,\ r^{(\beta)}_{i,v,t} > 0,\ \tau_{i,v,t} \geq \gamma_\tau\big], \tag{10}$$

where validity means a nonempty acyclic graph, a unique maximum, and no vetoed winner; all conditions are conjunctive and cost acts only in pre-query acquisition. If $G^\lambda_{i,b^{(0)}_i} = 1$, the initial feedback is admitted. Otherwise $e = 1$ acquires e_d, re-adapts its output, recomputes q, r, d, τ, and applies the same gate; $e = 0$ or a failed second gate skips the group. Thus $v^{\mathrm{fin}}_i \in \{b_i, e_d, \perp\}$ and $A^\lambda_i = \mathbb{I}[v^{\mathrm{fin}}_i \neq \perp]$. Route-training uses all four matrix outputs; policy-update prompts query only the realized sequence.

The router and gate therefore answer different questions. The router compares unqueried sources and determines the initial verifier, update direction, and acquisition cost. Once an output has been purchased, the gate can only accept or reject that graph; it cannot replace a poor initial choice with another first-stage verifier. The Qwen-72B path is a capped repair after rejection rather than a default decision rule.

4.3 CONDITIONAL GOLD-GRAPH RISK CERTIFICATE

Candidate a^λ is the complete policy in Equations 8–10. Let C_i^λ be Equation 2 for its final feedback. Consequence weights are fixed from task metadata before rollouts: $\rho_i \in \{1, 0.5, 0.25\}$ for outcome-changing, substantive, and formal/format violations, respectively. Uniform weights are a sensitivity analysis. We use $\kappa = 0.50$ as the primary consequence-weighted coverage floor and define

$$N_t(\lambda) = \mathbb{E}[\rho A^\lambda (1 - C^\lambda)], \qquad D_t(\lambda) = \mathbb{E}[\rho A^\lambda], \qquad R_t^{\mathrm{sel}}(\lambda) = \frac{N_t(\lambda)}{D_t(\lambda)}. \tag{11}$$

We call R_t^{sel} selective gold-graph inconsistency (GG-risk). It cannot be reduced merely by skipping, so we additionally require $D_t \geq \kappa$ and report raw coverage $\mathbb{E}[A]$. GG-risk is not downstream harm probability: gold-graph omissions, shared audit–verifier errors, and failures outside the declared graph-consistency event are not covered.

The certificate fold contains the full verifier matrix, retry outcomes, and gold graph, so it deterministically replays each candidate's initial choice, gate-dependent escalation probability, and terminal graph. Thus every fixed sequential candidate is observable, with all counterfactual calls and labels included in setup cost. With empirical means $\hat{N}, \hat{D}$ on n_t exchangeable frozen-policy groups, set

$$\epsilon_t = \sqrt{\frac{\log(2|\Lambda|/\delta_t)}{2n_t}}, \qquad U_t^{\mathrm{sel}}(\lambda) = \frac{\hat{N}_t(\lambda) + \epsilon_t}{\hat{D}_t(\lambda) - \epsilon_t}, \qquad D_t^L = \hat{D}_t - \epsilon_t. \tag{12}$$

Candidates with a nonpositive denominator are rejected. We select the highest route-validation utility among candidates satisfying $U_t^{\mathrm{sel}} \leq \alpha$ and $D_t^L \geq \kappa$; ties prefer lower cost and then larger D_t^L, followed by a deterministic candidate-ID rule detailed in the supplement. If none is feasible, the phase skips policy updating and leaves Λ unchanged.

Proposition 1 (phase-wise gold-graph risk certificate)**.** *If certificate and subsequent routed groups are exchangeable draws from the same frozen-policy prompt distribution, verifier randomness follows the same sampling procedure, and finite Λ is fixed before certificate labels are accessed, then with probability at least $1 - \delta_t$, every retained candidate satisfies $R_t^{\mathrm{sel}} \leq \alpha$ and $D_t \geq \kappa$. Across phases, choosing $\sum_t \delta_t \leq \delta_{\mathrm{all}}$ gives simultaneous probability at least $1 - \delta_{\mathrm{all}}$. Supplementary Section A proves the result and states its shift boundary.*

4.4 POLICY UPDATE AND COST MODEL

For an admitted GRPO group, PIVOT multiplies each graph-derived advantage by the final verifier's $r\tau$ factor, masks skipped groups, and clips the result at $\pm A_{\max}$. DPO analogously weights each final admitted edge loss by $r\tau$. No quantity from a rejected initial verifier weights escalated feedback. We report the realized aggregate optimizer step $\psi_{t+1} - \psi_t$ against a fresh audit direction, because the virtual displacement in Equation 5 is only a pre-update surrogate.

The primary scalar cost is amortized dollar-equivalent incremental cost under a rate card fixed before outcome inspection. It includes verifier acquisition, full-matrix route-training and certificate queries, gold-label processing, audit gradients, consistency fitting, gradient extraction, JL projection, routing, escalation, and retries. Shared rollouts and policy optimization are matched and reported separately. We also report raw calls, input/output tokens, GPU-hours, wall time, peak memory, setup share, and the break-even number of routed prompts; FLOPs, dollars, and latency are never directly summed without an explicit conversion rate.

The supplement provides complete phase-wise pseudocode and details the fold construction and certificate implementation.

5 ANALYSIS

Directional evidence. Under smoothness in the trainable LoRA coordinates, the standard first-order bound depends on the optimizer-native inner product between the trusted audit gradient and the candidate update, together with a quadratic update-size penalty. This motivates measuring direction in the actual trainable space. The rectified cosine $r\tau$ remains a directional score: it omits update

Table 1: Terminal evaluation and routed feedback. Each domain provides three first-acquisition sources and one escalation-only source; the four outputs form a setup matrix whose acquisition cost is included. Terminal prompts are source- and semantic-cluster-disjoint from every training and selection role.

Domain	Terminal task / metric	Initial verifier pool	Escalation source
Math	MATH-500 / exact accuracy	exact / Skywork-8B / Gemma-27B pairwise	Qwen-72B pairwise
Code	LiveCodeBench / pass@1	tests / Skywork-8B / static analysis	Qwen-72B execution
Instruction	IFEval / strict satisfaction	constraints / Skywork-8B / Gemma-27B rubric	Qwen-72B rubric
Science	GPQA Diamond / exact accuracy	rule subset / Skywork-8B / Gemma-27B evidence	Qwen-72B evidence

magnitude and therefore is not an estimator of the bound. A Johnson–Lindenstrauss projection preserves the relevant geometry only up to finite-sample distortion. We consequently compare cosine, norm-aware inner product, learned one-step targets, unsketched LoRA signals, projected signals, and realized optimizer steps (Supplementary Sections A.2–A.3).

Reliability and alignment. A binary direction-flip model scales expected signed alignment by $2q - 1$, which motivates the symmetric reliability transform. Real verifier errors may instead be asymmetric, omit rubric dimensions, or alter update magnitude. Our experiments therefore compare symmetric and asymmetric transforms with consistency-only, alignment-only, additive, gated, and direct-improvement alternatives. The gate and router play distinct roles: routing chooses which evidence to acquire before seeing its output, whereas the gate controls whether the selected evidence enters training after adaptation. Matched-weight factorials isolate the router's contribution under a fixed admission rule.

Certificate boundary. The certificate controls the declared GG-inconsistency event on each renewed frozen-policy phase. Its scope is narrower than downstream safety or semantic completeness of the gold graph. We therefore report unconditional erroneous-admission mass, allocate confidence across phases, account for setup cost, and repeat certification after every policy update. Supplementary Section A.3 states the assumptions and failure boundaries.

6 Experimental Design

The experiments test verifier drift, independent predictive validity of the directional score, cost-capped policy utility, compliance of observed GG-risk with its budget, and failure under policy or evidence shift.

6.1 Domains and Feedback Sources

Terminal evaluation uses MATH-500, post-cutoff LiveCodeBench, executable IFEval constraints, and GPQA Diamond (Lightman et al., 2024; Jain et al., 2025; Zhou et al., 2023; Rein et al., 2024). These benchmarks are held out from policy updates, routing, calibration, audit gradients, certification, hyperparameter selection, and checkpoint selection. Source- and semantic-cluster-disjoint pools supply policy updates and audits: NuminaMath plus GSM-style reasoning data for math, APPS train plus CodeContests for code, disjoint executable-constraint pools for instruction following, and MMLU-Pro science subsets. Supplementary Section G details construction and decontamination. RewardBench 2 is restricted to verifier analysis and excluded from downstream policy evidence (Malik et al., 2026). Quantitative gold labels come from official answers, executable tests, or frozen public evidence; no newly collected human annotation enters the reported evaluation.

The shared learned verifier is `Skywork/Skywork-Reward-Llama-3.1-8B-v0.2`. First-stage generative judging uses `google/gemma-2-27b-it` for math pairwise comparison, instruction rubrics, and retrieval-conditioned science assessment. `Qwen/Qwen2.5-72B-Instruct` is escalation-only in every domain (Skywork Team, 2024; Google, 2024; Qwen Team, 2024a). Code uses deterministic static analysis as its third initial source (Table 1). Each model uses a frozen role-specific prompt and a separate inference call.

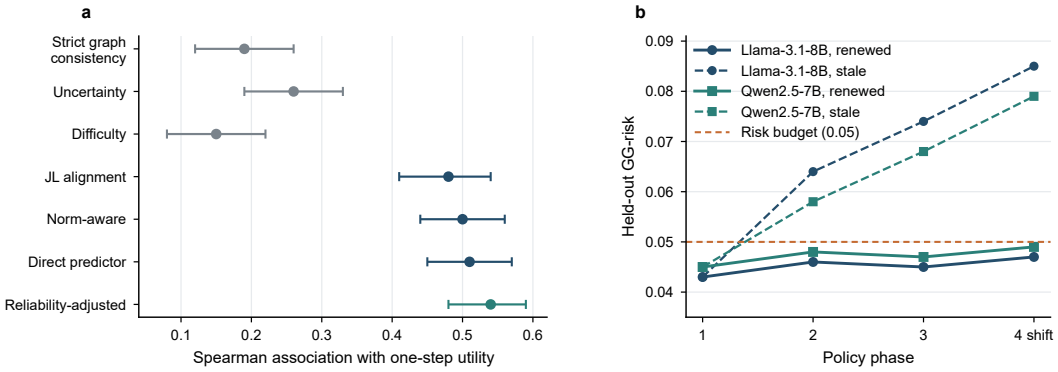


Figure 2: Mechanism and renewal diagnostics. (a) Macro-average Spearman association with matched one-step utility; whiskers are 95% checkpoint-blocked bootstrap intervals. (b) Held-out GG-risk under full renewal or fully stale reuse. Exact seed, domain, and phase values appear in the supplement.

6.2 Policies, Optimizers, and Matched Budgets

The policy models are `meta-llama/Llama-3.1-8B-Instruct` and `Qwen/Qwen2.5-7B-Instruct`, each with a separate rank-16 LoRA adapter (Meta Llama, 2024; Qwen Team, 2024b). GRPO is primary and DPO tests optimizer transfer. Four frozen-policy phases use separate optimizer, router, calibrator, and certificate state. Across both models, the study contains 128K update groups, four-output setup matrices, and five seeds per model. All methods receive matched pools, tokens, update budgets, and total incremental cost. Supplementary Section G specifies sampling, optimization, phase sizes, and setup accounting.

6.3 Baselines

BayesianRouter is compared with PIVOT-RM on the same reward-model pool, labels, trials, tokens, and cost cap (Wu & Lu, 2026). BayesianRouter-H shares PIVOT's heterogeneous pool, while Uncertainty Router receives the same fast/strong pair and call quota. Additional controls cover conformal admission, feedback weighting, value-of-information routing, aggregation, and curricula (Xu et al., 2025a; Chen et al., 2026; Uddin & Bauer, 2026; Saliencro et al., 2026). Supplementary Tables S10–S12 specify every pool and budget.

6.4 Metrics

Trusted utility is the equal-weight macro-average of terminal exact accuracy, pass@1, or strict constraint satisfaction over eight policy-model–domain cells; "worst" is their minimum. Verifier evaluation reports strict graph consistency, false acceptance and rejection, Brier score, ECE, and GG-risk–coverage curves. Routing reports the certificate bound, weighted-coverage lower bound, raw coverage, unconditional erroneous-admission mass, action mix, and raw and amortized costs. Disjoint direction, route, mechanism, certificate, and terminal folds separate every fitted signal from its evaluation target.

7 Results

7.1 Directional Scores Predict Independently Held-Out Utility

Strict gold-graph consistency had Spearman association 0.19 with matched one-step utility, with domain values from 0.15 to 0.24. Predictive uncertainty reached 0.26, compared with 0.48 for τ and 0.54 [0.48, 0.59] for $r\tau$ (Figure 2a). Domain-level $r\tau$ associations were 0.58, 0.55, 0.50, and 0.52 for math, code, instruction, and science. Relaxing unsupported but uncontradicted edges raised consistency to 0.27 [0.20, 0.34]. Paired checkpoint-blocked gaps remained 0.21 [0.13, 0.29] for τ and 0.27 [0.18, 0.35] for $r\tau$. Gold-graph incompleteness explains part of the strict-label gap, while

Table 2: Cost-accounted outcomes at 128K updates. "Worst" is the minimum policy-model–domain utility; cost includes setup amortization.

Method	Avg.↑	Worst↑	GG-risk↓	Coverage↑	Cost↓
Uncertainty Router	66.1	58.5	0.067	0.88	0.58
BayesianRouter (native RM pool)	66.3	58.8	0.065	0.87	0.50
PIVOT-RM (native RM pool)	67.7	60.0	0.050	0.83	0.53
BayesianRouter-H (common pool)	66.8	59.2	0.063	0.86	0.55
SCRC-I admission gate	66.2	58.8	0.043	0.68	0.60
VI-style feedback router	67.2	60.0	0.057	0.84	0.54
PIVOT	68.4	61.1	0.047	0.82	0.52

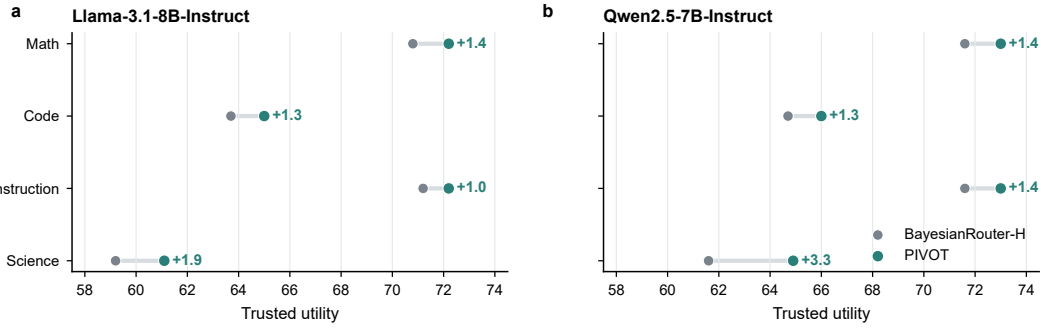


Figure 3: Trusted utility by policy model and domain. Connectors show paired gains over BayesianRouter-H; Supplementary Table S13 reports exact values.

update-aware signals retain the larger association. Norm-aware and direct predictors reached 0.50 and 0.51 with overlapping intervals. JL-256 preserved 93% of unsketched signs.

Relative to native feedback interfaces, graph adaptation reduced utility by 0.2, 0.3, and 0.2 points for exact, scalar, and pairwise feedback. Losses rose to 0.8 and 0.9 points for rubric and retrieval feedback. Adapter abstention ranged from 8.1% for exact feedback to 25.7% for retrieval feedback, identifying open-ended graph construction as a remaining bottleneck (Supplementary Table S23).

7.2 Cost-Capped Policy Improvement Under a Common Budget

At 128K updates, PIVOT-RM reached 67.7 utility versus 66.3 for BayesianRouter on the same reward-model pool. Under the common heterogeneous pool, PIVOT reached 68.4 utility and 61.1 worst-slice utility (Table 2). Gains were 1.6 [1.2, 2.0] over BayesianRouter-H, 1.2 over the VI-style router, and 2.3 [1.8, 2.7] over Uncertainty Router. The worst-slice gain over BayesianRouter-H was 1.9 [0.5, 3.3]. Setup-inclusive cost was 0.520 per prompt, below 0.550 and 0.580 for these comparators. Confidence intervals use a seed-first paired hierarchical bootstrap, treating each seed within a policy model as an independent training replicate.

The gain appeared for both policy models in every domain (Figure 3); model means were 67.6 for Llama-3.1-8B-Instruct and 69.2 for Qwen2.5-7B-Instruct. Science remained the weakest slice, with utilities of 61.1 and 64.9 and coverage of 0.77 and 0.78. Under DPO, PIVOT and BayesianRouter-H obtained 67.0/0.049 and 65.9/0.064 utility/GG-risk; under GRPO, the corresponding values were 68.4/0.047 and 66.8/0.063. Supplementary Tables S12–S18 report resources, domain cells, seeds, and optimizer transfer.

7.3 Gold-Graph Risk Control Requires Renewal

At $\alpha = 0.05$, the certificate bound was 0.049 and held-out GG-risk was 0.047 [0.041, 0.054] at 82% coverage. Targets of 0.03 and 0.10 yielded risk/coverage of 0.026/0.65 and 0.091/0.93, respectively. Fully stale reuse raised GG-risk to 0.082, while phase renewal restored it to 0.048 (Figure 2b). Component-wise reuse separated the failure: a stale directional predictor reduced utility from 68.4 to

67.5, whereas a stale consistency model raised GG-risk to 0.071. New science sources and conflicting retrieval snapshots remained infeasible under the original coverage requirement. Supplementary Tables S19–S22 report the candidate grid, phase values, and component-wise staleness.

7.4 Routing, Weighting, and Setup Amortization

Matched-weight factorials attributed 1.1–1.2 utility points to pre-query routing. Removing the reliability margin or certificate raised GG-risk to 0.079 and 0.084. Disabling escalation reduced utility from 68.4 to 67.8, coverage from 0.82 to 0.77, and cost from 0.52 to 0.47; only 5–17% of domain-level outcomes were admitted after escalation. Removing the coverage floor reduced GG-risk to 0.031 while coverage fell to 0.49. Setup amortized after 37.6K prompts versus BayesianRouter-H and 67.6K versus uncertainty routing (Supplementary Tables S16, S30, and S37).

8 Discussion and Conclusion

PIVOT treats verifier choice as a policy-conditional training decision. Across two policy models and four domains, it reached 68.4 trusted utility at GG-risk 0.047 and 82% coverage. Matched-cost gains were 1.6 points over BayesianRouter-H and 2.3 over Uncertainty Router. Pre-query routing retained a 1.1–1.2-point advantage under matched weighting, while the gate controlled admission risk and escalation repaired a limited subset.

Relaxing incomplete gold edges raised the consistency–utility association from 0.19 to 0.27, yet τ and $r\tau$ remained higher by 0.21 and 0.27. Update direction therefore captures variation beyond graph consistency. The certificate applies to the declared gold-graph event within renewed frozen phases. Shared verifier errors, open-world evidence shifts, and short deployments remain the principal observed boundaries.

Generative AI Use Disclosure

The authors used generative AI tools to edit and condense the manuscript, organize literature during drafting, revise scientific figures, and provide feedback on experimental design and result interpretation. The authors checked all citations, numerical results, figures, methods, and conclusions, and take responsibility for the final content and artifacts.

References

Sher Badshah, Ali Emami, and Hassan Sajjad. Scope: Selective conformal optimized pairwise llm judging, 2026. URL https://arxiv.org/abs/2602.13110. Accepted at the 43rd International Conference on Machine Learning (ICML 2026).

Tiejin Chen, Xiaoou Liu, Vishnu Nandam, Kuan-Ru Liou, and Hua Wei. Conformal feedback alignment: Quantifying answer-level reliability for robust llm alignment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3561–3572, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.184. URL https://aclanthology.org/2026.findings-eacl.184/.

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. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, pp. 50905–50931, 2024.

Shaddin Dughmi, Mahdi Haghifam, and Yusuf Hakan Kalayci. Adaptive generate-rank-verify: Inference-time search with costly verification, 2026. URL https://arxiv.org/abs/2605.17609.

Google. Gemma-2-27B-IT model card. Hugging Face model card, 2024. URL https://huggingface.co/google/gemma-2-27b-it.

Xinyan Guan, Yanjiang Liu, Xinyu Lu, Boxi Cao, Ben He, Xianpei Han, Le Sun, Jie Lou, Bowen Yu, Yaojie Lu, and Hongyu Lin. Search, verify and feedback: Towards next generation post-training paradigm of foundation models via verifier engineering, 2024. URL https://arxiv.org/abs/2411.11504.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Advances in Neural Information Processing Systems*, volume 37, pp. 73884–73919. Curran Associates, Inc., 2024. doi: 10.52202/079017-2350.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, pp. 127924–127945, 2026.

Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. Advancing process verification for large language models via tree-based preference learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2086–2099, 2024. doi: 10.18653/v1/2024.emnlp-main.125.

Haitao Hong, Yuchen Yan, Xingyu Wu, Guiyang Hou, Wenqi Zhang, Weiming Lu, Yongliang Shen, and Jun Xiao. Cooper: Co-optimizing policy and reward models in reinforcement learning for large language models, 2025. URL https://arxiv.org/abs/2508.05613.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, pp. 58791–58831, 2025.

Hamed Khosravi and Xiaoming Huo. Conformal selective acting: Anytime-valid risk control for rlvr-trained llms, 2026. URL https://arxiv.org/abs/2605.20270.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.96.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. In *International Conference on Learning Representations*, pp. 16484–16517, 2025.

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 *International Conference on Learning Representations*, pp. 39578–39601, 2024.

Jiamiao Liu, Dewen Qiao, Yu Zhang, and Xuetao Chen. Certify or refuse: A cross-model map for selective risk control with coverage floors under covariate shift. *arXiv preprint arXiv:2608.10893*, 2026. doi: 10.48550/arXiv.2608.10893. URL https://arxiv.org/abs/2608.10893.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, pp. 144839–144866, 2026.

Meta Llama. Llama-3.1-8B-Instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, pp. 34433–34448, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70305–70337. Curran Associates, Inc., 2025. doi: 10.52202/085713-2364.

Cheng Qian, Zuxin Liu, Shirley Kokane, Akshara Prabhakar, Jielin Qiu, Haolin Chen, Zhiwei Liu, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, Caiming Xiong, and Huan Wang. xrouter: Training cost-aware llms orchestration system via reinforcement learning. *arXiv preprint arXiv:2510.08439*, 2025. doi: 10.48550/arXiv.2510.08439. URL https://arxiv.org/abs/2510.08439.

Qwen Team. Qwen2.5-72B-Instruct model card. Hugging Face model card, 2024a. URL https://huggingface.co/Qwen/Qwen2.5-72B-Instruct.

Qwen Team. Qwen2.5-7B-Instruct model card. Hugging Face model card, 2024b. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Tom Saliencro, Rohan Desai, Priya Nair, Maya Lindqvist, and Daniel Whitmore. Uncertainty is not enough: Value-of-information routing for mixtures of lora experts. *arXiv preprint arXiv:2608.02528*, 2026. doi: 10.48550/arXiv.2608.02528. URL https://arxiv.org/abs/2608.02528.

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. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Skywork Team. Skywork-Reward-Llama-3.1-8B-v0.2 model card. Hugging Face model card, 2024. URL https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B-v0.2.

Annette Taberner-Miller. Paretobandit: Budget-paced adaptive routing for non-stationary llm serving. *arXiv preprint arXiv:2604.00136*, 2026. doi: 10.48550/arXiv.2604.00136. URL https://arxiv.org/abs/2604.00136.

Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Causal llm routing: End-to-end regret minimization from observational data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41861–41886. Curran Associates, Inc., 2025. doi: 10.52202/085713-1252. URL https://doi.org/10.52202/085713-1252.

Iqtedar Uddin and André Bauer. Conformal llm routing with distribution-free safety guarantees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 791–799, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-srw.70. URL https://aclanthology.org/2026.acl-srw.70/.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114728–114754. Curran Associates, Inc., 2025. doi: 10.52202/085713-3824.

Fangxin Wang, Peyman Baghershahi, Langzhou He, Henry Peng Zou, Sourav Medya, and Philip S. Yu. Filter-then-weight: Online data selection and reweighting for llm fine-tuning. *arXiv preprint arXiv:2604.00001*, 2026a. doi: 10.48550/arXiv.2604.00001. URL https://arxiv.org/abs/2604.00001.

Huaiyang Wang, Xiaojie Li, Xiaohan Wang, Zhixia Zhang, Xiaodong Lu, Zixuan Huang, Jiajun Chai, Guojun Yin, Deqing Wang, Haoyi Zhou, Yaodong Yang, Jianxin Li, and Yikun Ban. Policy improvement reinforcement learning. *arXiv preprint arXiv:2604.00860*, 2026b. doi: 10.48550/arXiv.2604.00860. URL https://arxiv.org/abs/2604.00860.

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*, pp. 9426–9439, 2024a. doi: 10.18653/v1/2024.acl-long.510.

Yikai Wang, Shang Liu, and Jose Blanchet. Wasserstein distributionally robust regret optimization for reinforcement learning from human feedback, 2026c. URL https://arxiv.org/abs/2605.00155.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alexander Nicholas D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51161–51176. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/wang24ay.html.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations*, pp. 58769–58790, 2025.

Xinle Wu and Yao Lu. Reward model routing in alignment. In *International Conference on Learning Representations*, pp. 88128–88144, 2026.

Yingwen Wu, Tao Li, Xinwen Cheng, Jie Yang, and Xiaolin Huang. Low-dimensional gradient helps out-of-distribution detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11378–11391, 2024. doi: 10.1109/TPAMI.2024.3459988.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025a. doi: 10.48550/arXiv.2512.12844. URL https://arxiv.org/abs/2512.12844.

Zhenghao Xu, Qin Lu, Qingru Zhang, Liang Qiu, Ilgee Hong, Changlong Yu, Wenlin Yao, Yao Liu, Homing Jiang, Lihong Li, Hyokun Yun, and Tuo Zhao. Ask a strong llm judge when your reward model is uncertain. In *Advances in Neural Information Processing Systems*, volume 38, pp. 74639–74664, 2025b. doi: 10.52202/085713-2506.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable process rewards for agentic reasoning, 2026. URL https://arxiv.org/abs/2605.10325. AAMAS 2026 Workshop on Strategic Engineering, poster.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. In *International Conference on Learning Representations*, pp. 133430–133457, 2026.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. Genprm: Scaling test-time compute of process reward models via generative reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):34932–34940, 2026. doi: 10.1609/aaai.v40i41.40797.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*, 2024. doi: 10.48550/arXiv.2412.06559.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

SUPPLEMENTARY MATERIAL FOR PIVOT: RISK-CONTROLLED ROUTING OF VERIFIER FEEDBACK FOR RELIABLE POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

This supplement provides the formal certificate, canonical feedback adapter, sequential acquisition procedure, audit construction, gradient diagnostics, cost analysis, baseline implementations, and statistical analyses supporting the main paper. Sections A–C develop the guarantee, feedback representation, and acquisition procedure; Sections D–G describe audit, optimization, cost, and implementation details; and Sections H–J report extended results, statistical analyses, failure cases, and safeguards.

A NOTATION AND FORMAL GUARANTEES

Table S1: Core notation. A candidate a^λ denotes the complete sequential policy, including acquisition, post-query gating, escalation, and skipping.

Symbol	Meaning
$E_{i,v}, B_{i,v}$	directed preference edges and critical-veto set returned by the canonical adapter
$C_{i,v}$	strict gold-consistency event for the accepted feedback graph
A_i^λ	indicator that a^λ ultimately admits feedback for policy optimization
ρ_i	pre-specified consequence weight in $[0, 1]$
N_t, D_t	weighted erroneous-admission mass and weighted admission mass
$R_t^{\mathrm{sel}} = N_t/D_t$	consequence-weighted conditional gold-graph inconsistency among admitted feedback (GG-risk)
κ	minimum lower bound required for weighted admission mass
$d_{i,v,t}$	optimizer-native virtual LoRA displacement induced by verifier v
$g_{d,t}^{\mathrm{aud}}$	trusted-objective gradient from the disjoint audit-direction fold for stratum d
$m_{i,v,t}$	pre-query reliability-adjusted directional-alignment score

A.1 CONDITIONAL GG-RISK CERTIFICATE

Throughout the supplement, GG-risk means conditional inconsistency of an admitted feedback graph with the prespecified gold graph. It is a training-feedback validity measure, not a probability of downstream physical, social, or clinical harm; consequence weights only prioritize declared strata.

For a fixed candidate λ in frozen phase t, define

$$X_i^\lambda = \rho_i A_i^\lambda (1 - C_i^\lambda), \qquad Y_i^\lambda = \rho_i A_i^\lambda, \tag{S1}$$

so $X_i^\lambda, Y_i^\lambda \in [0, 1]$. Their population means are $N_t(\lambda)$ and $D_t(\lambda)$. On a certificate split of size n_t, write the corresponding sample means as $\hat{N}_t$ and $\hat{D}_t$ and set

$$\epsilon_t = \sqrt{\frac{\log(2|\Lambda|/\delta_t)}{2n_t}}, \qquad N_t^U = \hat{N}_t + \epsilon_t, \qquad D_t^L = \hat{D}_t - \epsilon_t. \tag{S2}$$

If $D_t^L > 0$, define $U_t^{\mathrm{sel}} = N_t^U / D_t^L$; otherwise set $U_t^{\mathrm{sel}} = +\infty$.

Proposition S1 (phase-wise GG-risk)**.** *Suppose that (i) certificate groups and the subsequently routed groups are exchangeable draws from the same frozen-policy prompt distribution, (ii) verifier randomness follows the same sampling procedure, and (iii) the finite 24-element action-policy set Λ is fixed before certificate labels are accessed. Then, with probability at least $1-\delta_t$, simultaneously for all $\lambda \in \Lambda$,*

$$N_t(\lambda) \le N_t^U(\lambda), \qquad D_t(\lambda) \ge D_t^L(\lambda). \tag{S3}$$

Consequently, every candidate satisfying $U_t^{\mathrm{sel}}(\lambda) \le \alpha$ and $D_t^L(\lambda) \ge \kappa$ obeys $R_t^{\mathrm{sel}}(\lambda) \le \alpha$ and $D_t(\lambda) \ge \kappa$.

Proof. For a fixed λ, one-sided Hoeffding bounds make each relevant tail probability at most $\exp(-2n_t\epsilon_t^2)$. The choice in Equation S2 makes each probability at most $\delta_t/(2|\Lambda|)$. A union bound over the two events and all candidates gives the simultaneous statement. On that event, $D_t > 0$ for every retained candidate and $N_t/D_t \le N_t^U/D_t^L \le \alpha$; the denominator guarantee follows directly. □

For T adaptive phases, pre-allocating δ_t with $\sum_{t=1}^{T} \delta_t \le \delta_{\mathrm{all}}$ yields a simultaneous failure probability no greater than δ_{all}. We use $\delta_t = \delta_{\mathrm{all}}/T$ in the confirmatory analysis. The result is not anytime-valid: a mid-phase policy change, a changed verifier prompt, or any change to the fixed 24-element action-policy set invalidates the corresponding certificate. Deployment-side anytime control solves a different problem and additionally requires predictable policy and score updates, nested monotone gates, monotone conditional GG-risk, frontier stability, and declared within-epoch drift budgets (Khosravi & Huo, 2026). Its guarantee is relative to the verifier's declared failure event, not harms that the verifier fails to represent.

Why the denominator matters. The earlier unconditional quantity $\mathbb{E}[\rho A(1-C)]$ can be made small by admitting almost nothing. Proposition S1 instead bounds the conditional ratio among weighted admissions and rejects candidates whose weighted admission mass is below κ. We still report unweighted coverage $\mathbb{E}[A]$, unconditional erroneous-graph admission mass N_t, and per-stratum ratios; none is interchangeable with R_t^{sel}.

A.2 LoRA Restriction, JL Approximation, and Local Improvement

Let ψ denote all and only the trainable LoRA parameters in the primary experiment. With optimizer state frozen for scoring, M_t is the declared preconditioner and

$$d_{i,v,t} = -\eta M_t \nabla_\psi \mathcal{L}_{\mathrm{pol}}(\psi_t; x_i, Y_i, F_{i,v}). \tag{S4}$$

If $J_t^\star$ is L-smooth in these same coordinates, then

$$J_t^\star(\psi_t + d) \ge J_t^\star(\psi_t) + \langle \nabla_\psi J_t^\star(\psi_t), d\rangle - \frac{L}{2}\|d\|_2^2. \tag{S5}$$

This inequality concerns the unsketched displacement and its norm. It neither proves that a positive cosine is sufficient at an arbitrary step size nor licenses dropping trainable coordinates that the real optimizer updates.

Proposition S2 (finite-set projection diagnostic)**.** *Let $\mathcal{U}$ contain M nonzero normalized LoRA displacements and audit directions fixed before drawing a sub-Gaussian JL map S. If the sketch dimension is $m = O(\varepsilon^{-2}\log(M/\delta))$, then, with probability at least $1-\delta$, all pairwise cosines are preserved to $O(\varepsilon)$. In particular, the sign of an unsketched cosine is preserved only when its absolute margin exceeds the realized distortion bound.*

Proof sketch. Apply the standard JL norm guarantee to every normalized vector and every pairwise sum and difference. Polarization, $4\langle u, v\rangle = \|u+v\|^2 - \|u-v\|^2$, bounds all inner-product errors; the preserved individual norms then bound cosine distortion. The augmented set has $O(M^2)$ elements, retaining logarithmic dependence on M. □

We measure the empirical maximum cosine distortion on a held-out finite vector set. A projected sign without the required unsketched margin is marked "indeterminate" and is excluded from the theoretical interpretation. Work on PCA-compressed gradients in visual OOD detection supports the plausibility of retaining useful information (Wu et al., 2024); it does not establish the LoRA/JL result required here.

A.3 Reliability Transform and Asymmetric Error

Under the stylized binary direction-flip model $d_v = \xi d_i^\star$ with $\Pr(\xi = 1) = q$,

$$\mathbb{E}\langle g^{\mathrm{aud}}, d_v\rangle = (2q-1)\langle g^{\mathrm{aud}}, d_i^\star\rangle. \tag{S6}$$

This motivates $[2q-1]_+$ only for symmetric binary flips. The main method therefore uses

$$r^{(\beta)}(q) = [q - \beta(1-q)]_+, \qquad \beta \geq 1, \tag{S7}$$

where β is a pre-specified false-acceptance penalty and $\beta = 1$ recovers the symmetric margin. It remains a ranking heuristic for graph-valued feedback rather than an estimate of $\Delta^\star_{i,v,t}$: it omits update magnitude, curvature, and minibatch interactions. The default $\beta = 1.5$ is selected once on the route-validation fold, fixed before certificate labels are accessed, and held constant across domains and phases; Table S28 reports alternatives. Prior controlled experiments by Mitsuhashi et al. (2026) used Qwen2.5-0.5B/1.5B with GRPO on GSM8K and found faster degradation under false-negative verifier noise. We therefore report both error types while treating this narrow empirical asymmetry as neither universal nor a theorem.

B Canonical Feedback Representation

Table S2: Native feedback is converted to a common directed graph. Thresholds, prompts, weights, and veto definitions are fixed before each policy phase.

Native source	Directed-edge construction	Critical-veto set	Abstention trigger
Exact answer / tests	(j,k) if j has a strictly better exact outcome or fixed test score	invalid execution, forbidden action, or declared safety test failure	tie at the top or flaky/indeterminate test
Scalar reward model	(j,k) if $s_j - s_k \geq \gamma_v$	critical head exceeds its frozen threshold; empty if no such head exists	no unique maximum, small margin, or non-finite score
Pairwise judge	edge only when position-swapped queries agree and confidence exceeds γ_v	explicit critical-error label	order disagreement, tie, cycle, or prompt failure
Rubric vector	fixed weighted rubric total with margin γ_v	any pre-declared critical rubric violation	unresolved critical item, cycle, or no unique maximum
Retrieval critic	fixed evidence-support score and pair margin	cited evidence establishes a critical contradiction	unavailable evidence, source conflict, or insufficient support
Escalation	initial judgment plus one independent escalation-judge output under a frozen combination rule	combined critical violations	unresolved model disagreement or budget exhaustion

Table S3 defines the modality-native comparators used to measure information lost by the common graph. They share the same response groups, verifier outputs, admission budget, optimizer, update tokens, and cost accounting. The native branch bypasses only graph construction; it does not receive additional verifier calls or gold labels.

Table S3: Native-interface controls for the canonical adapter. These controls retain information that the directed graph may discard.

Source	Best native training interface	Canonical graph interface
Exact answer / tests	centered executable or test score for GRPO; executed chosen–rejected pair for DPO	strict outcome edges and critical vetoes
Scalar reward model	frozen affine-calibrated scalar magnitude, including ties and small differences	margin-thresholded edges using γ_v
Pairwise judge	native pairwise logistic/DPO term with confidence weighting	agreed order-swapped edges only
Rubric vector	dimension-wise weighted objective retaining per-rubric uncertainty and partial credit	edges from a frozen weighted total plus critical vetoes
Retrieval critic	evidence-support-weighted pairwise term retaining source confidence	support-margin edges plus contradiction vetoes

Gold auditing creates a partial order $E_i^\star$ rather than forcing a total ranking where evidence supports ties. An accepted graph satisfies *strict gold-graph consistency* iff it is nonempty, every promoted

edge belongs to $E_i^\star$, and its winner set excludes the gold critical set $B_i^\star$. This is a positive-support event: an edge absent from the partial gold graph fails the event even when the audit records that pair as underdetermined. We additionally report edge-level precision/recall so a large graph is not made incomparable with a single pair.

For mechanism sensitivity, let E_i^- contain directions explicitly contradicted by the gold audit, and let $M_i^\star$ be the outcome-maximal, non-critical responses under the available official or executable labels. We define

$$C_{i,v}^{\mathrm{nc}} = \mathbb{I}\big[E_{i,v} \neq \emptyset,\ E_{i,v} \cap E_i^- = \emptyset,\ \mathrm{Win}(E_{i,v}) \subseteq M_i^\star,\ \mathrm{Win}(E_{i,v}) \cap B_i^\star = \emptyset\big]. \tag{S8}$$

The non-contradiction label stops penalizing an extra edge between outcome-tied responses when the edge is not opposed by the audit. We use C^{nc} only for mechanism sensitivity; the consistency model, gate, GG-risk, and certificate continue to use the positive-support label C. A process-expanded gold graph would require additional supervision such as human step labels, rollout-derived step values, or tree-derived pairwise preferences (Lightman et al., 2024; Wang et al., 2024; He et al., 2024); current process verifiers do not provide uniformly reliable labels on harder reasoning tasks (Zheng et al., 2024).

For valid $E_{i,v}$, net-win scores are

$$s_{ijv} = \frac{1}{K-1} \sum_{k \neq j} (\mathbb{I}[(j,k) \in E_{i,v}] - \mathbb{I}[(k,j) \in E_{i,v}]). \tag{S9}$$

GRPO standardizes these scores within group, then applies the exact reference term and clipping used in the real loss. DPO uses every directed edge once and normalizes the group contribution by $|E_{i,v}|$. A skipped or invalid graph contributes exactly zero loss. Escalated feedback passes through the same adapter, consistency model, and cost accounting as the initial feedback. Let v_i^{final} denote the initial verifier if its graph is admitted and the escalation verifier otherwise. After escalation, the graph, consistency probability q, reliability margin $r^{(\beta)}$, virtual displacement d, and alignment τ are all recomputed. The realized loss weight is $r^{(\beta)}_{i,v_i^{\mathrm{final}},t}\tau_{i,v_i^{\mathrm{final}},t}$, never the weight of the rejected initial verifier.

Algorithm 1 Feedback-to-loss adapter

Map native verifier output to (E, B) using the frozen row in Table S2.
if $E = \emptyset$, E is cyclic, no unique maximum exists, or a critical winner is vetoed **then**
  return ABSTAIN and one mutually exclusive first-failure code.
end if
Compute group net-win scores s_j.
if optimizer is GRPO **then**
  standardize s_j; instantiate the full clipped, reference-corrected group loss.
else
  instantiate one DPO term per edge and normalize by $|E|$.
end if
return the loss, graph metadata, final-verifier identity, and optimizer-native virtual displacement.

C Sequential Acquisition and Data Independence

C.1 Pre-Query and Post-Query Information Sets

The frozen pre-query schema $\phi_{i,v,t}^-$ contains a 64-dimensional seeded projection of the prompt embedding; log prompt and response length; response-group entropy, answer agreement, and pairwise lexical diversity; domain and consequence stratum; policy-model and checkpoint identifiers; verifier capability flags; lagged per-verifier accuracy, Brier score, ECE, and coverage; and the fixed acquisition rate. Current verifier outputs, confidence, rubric decisions, retrieval results, and gradients derived from those outputs are excluded. The primary $\hat{m}$ model is a histogram gradient-boosted regressor with at most depth three and 200 trees. The post-query $\hat{q}$ model is an ℓ_2-regularized logistic model with domain–verifier interactions, followed by cross-fitted isotonic calibration. The 4,096-item route-training fold is partitioned by semantic cluster into 3,072 fit and 1,024 calibration items;

no item crosses the partition. Model class, feature, and calibration ablations appear in Table S29. The predictor is trained to estimate

$$m_{i,v,t} = \mathbb{E}\big[r^{(\beta)}_{i,v,t}\tau_{i,v,t} \mid \phi^{-}_{i,v,t}\big]. \tag{S10}$$

After one verifier is acquired, ϕ^{+} may include its graph, margins, self-reported uncertainty, vetoes, and virtual-step diagnostics. These features feed the calibrated consistency model and post-query gate. A rejected item permits at most one separately capped escalation, with no additional online queries.

The complete 24-element action-policy set is fixed before certificate labels are accessed:

$$\Lambda = \{0, 0.05, 0.10\} \times \{0.65, 0.75\} \times \{0, 0.10\} \times \{0, 1\}, \qquad |\Lambda| = 24, \tag{S11}$$

where $\lambda = (\lambda_c, \gamma_q, \gamma_\tau, e)$. The cost coefficient $\lambda_c \in \{0, 0.05, 0.10\}$ trades predicted update usefulness against normalized first-query cost. The calibrated-consistency floor is $\gamma_q \in \{0.65, 0.75\}$, and the alignment floor is $\gamma_\tau \in \{0, 0.10\}$. Finally, $e \in \{0, 1\}$ disables or permits one Qwen-72B query after initial rejection. Their Cartesian product gives 24 candidates. Candidate IDs increase first in e, then γ_τ, γ_q, and λ_c; Table S22 reports the selected candidate in every phase. For a queried output, admission requires

$$\mathrm{valid}(E_{i,v}) \wedge q_{i,v,t} \geq \gamma_q \wedge r^{(\beta)}_{i,v,t} > 0 \wedge \tau_{i,v,t} \geq \gamma_\tau. \tag{S12}$$

Only features available before the next query enter acquisition cost. In particular, current confidence, graph statistics, and virtual gradients cannot retroactively influence the cost-adjusted first action.

The four candidate coordinates have distinct roles and should not be read as a four-threshold gate. Only γ_q and γ_τ threshold a purchased output. The cost coefficient λ_c changes the initial source ranking, while e controls whether rejection can trigger one repair query. Structural graph validity and $r^{(\beta)} > 0$ are fixed admission checks. Consequently, the post-query gate cannot compare unqueried first-stage verifiers or undo their acquisition cost; that choice remains the responsibility of the pre-query router.

Algorithm 2 One complete frozen-policy phase

Freeze π_t, optimizer state, all four verifier versions, adapters, feature schemas, rate card, Λ, and split hashes.
Sample all route, certificate, audit, update, and terminal groups before label access; acquire the full four-output matrix only on route/certificate groups.
Build domain gold graphs $F^{\star}_i$; compute $g^{\mathrm{aud}}_{d,t}$ on the independent 128-example audit fold and its four-block jackknife diagnostics.
Fit cross-fitted $\hat{m}$, $\hat{q}$, and their calibration maps on route-training folds; store every fitted artifact and version identifier.
for each $\lambda \in \Lambda$ **do**
  Replay the complete sequential policy on certificate groups, including at most one escalation; compute $N^U_t(\lambda)$ and $D^L_t(\lambda)$.
end for
Retain candidates with $N^U_t / D^L_t \leq \alpha$ and $D^L_t \geq \kappa$; select maximal route-validation utility, breaking ties by lower cost, larger D^L_t, then lexicographically smaller candidate ID.
if no candidate is feasible **then**
  skip the phase; do not update the policy.
end if
for each update group i **do**
  Choose $b^{(0)}_i = \arg\max_{v \in V^{(0)}_d} \{\hat{m}_{i,v,t} - \lambda_c \bar{c}_v\}$ and query it.
  Adapt output and compute (E, B, q, r, d, τ).
  if Eq. (S12) fails and $e = 1$ **then**
    query the domain escalation source once and recompute (E, B, q, r, d, τ) entirely from that output.
  end if
  admit with bounded weight $r\tau$ iff the final output passes the gate; otherwise skip the group.
end for
Optimize the buffered admitted losses to obtain π_{t+1}; evaluate the independent terminal fold; invalidate every phase-t certificate.

C.2 SPLIT ROLES AND FULL-INFORMATION COST

Table S4: Role separation per frozen phase. Split sizes and permitted uses are fixed before confirmatory outcomes are inspected.

Role	Size	Permitted use
Route training	4,096	full verifier matrix; fit pre-query directional score and post-query consistency; choose thresholds
Certificate	16,384	full verifier matrix plus gold graph; use 32,768 for $\alpha = 0.03$; evaluate only pre-specified candidates
Audit gradient	128/stratum	independent trusted objective; rotate folds by phase; never fit or certify router
Mechanism outcome	source-disjoint	independently scored executable/task utility; never forms an audit direction, route label, threshold, or certificate
Policy update	16,384 per policy-model phase	sequentially query one verifier, optionally escalate once, then update or skip
Final test	benchmark-fixed	downstream utility only; evaluated once per frozen seed checkpoint

Every full-matrix verifier query and gold label on route-training and certificate folds is included in setup cost. Full information makes each fixed sequential candidate observable without off-policy estimation. Extending the certificate to logged propensities remains future work. Candidate policies, adapter versions, thresholds, escalation rules, consequence weights, and the global δ schedule are versioned and fixed before certificate labels are accessed.

Within a phase, certificate and update groups are sampled before routing by random partition from the same frozen-policy prompt mixture. Verifier sampling temperature, position swap, retrieval snapshot, and retry policy are identical. Routing affects which feedback is admitted; the source distribution of routed groups remains fixed. A policy update, prompt-mixture change, verifier checkpoint or prompt change, or retrieval refresh starts a new phase and requires a new split.

Within each policy-model phase, π_t, optimizer moments, verifier checkpoints and prompts, adapter margins, retrieval index, router, calibrators, thresholds, audit direction, and JL projection remain fixed while 16,384 update groups are routed. PIVOT does not train the verifier models. Admitted losses are buffered and optimized only after acquisition, producing π_{t+1} and ending the phase. Interleaving a policy update with further routing therefore starts a new phase and requires recertification.

Data provenance and decontamination. Policy, verifier, route, certificate, audit, and final-test items are separated by source and semantic cluster. For each artifact, we retain the URL, license, release ID, checksum, and transformation history. Decontamination uses exact and normalized matching, answer-stripped matching, template clustering, and embedding-neighbor review. Contaminated benchmark items are excluded from confirmatory analysis.

Table S5: Source-disjoint data releases. Each non-test cell is associated with a release ID, checksum, temporal cutoff, semantic-cluster exclusion, and retained count in the run metadata.

Domain	Update / route / certificate	Audit-direction source	Mechanism-outcome source	Terminal test
Math	NuminaMath + GSM-style public pool	source-disjoint GSM8K / SVAMP	ASDiv reasoning subset	fixed MATH-500 release
Code	APPS train + CodeContests	MBPP	HumanEval+ / EvalPlus-style set	frozen LiveCodeBench window
Instruction	executable-constraint pool A	executable-constraint pool B	executable-constraint pool C	executable IFEval
Science	MMLU-Pro science subsets	ARC-Challenge	OpenBookQA / independent science QA	official GPQA Diamond

D TRUSTED AUDIT CONSTRUCTION

Audit strata separate exact outcome, executable behavior, instruction constraint, open rubric, and evidence consistency. For multi-objective prompts, the routing and independent test objectives use pre-specified mixtures; an additional evaluation uses a separately constructed mixture to detect audit overfitting. The gold graph $F_i^\star = (E_i^\star, B_i^\star)$ is built from official answer equivalence in math, hidden tests in code, executable constraints in instruction following, and official answers plus a frozen evidence snapshot in science. For every group with at least one supported gold edge, define

$$J_{d,t}^\star(\psi) = -\frac{1}{|A_{d,t}|} \sum_{i \in A_{d,t}} L_{\mathrm{aud}}(\psi; F_i^\star), \tag{S13}$$

$$L_{\mathrm{aud}}(\psi; F_i^\star) = -\frac{1}{|E_i^\star|} \sum_{(j,k) \in E_i^\star} \log \sigma(\ell_\psi(y_{ij}|x_i) - \ell_\psi(y_{ik}|x_i)), \tag{S14}$$

where ℓ_ψ is length-normalized log probability and groups without a supported edge do not contribute. Each domain–phase uses $n = 128$ audit groups partitioned into four source-balanced blocks of 32; leave-one-block-out jackknife estimates quantify audit-gradient variance. For each stratum d, the audit-direction fold defines the LoRA-space gradient $g_{d,t}^{\mathrm{aud}} = \nabla_\psi J_{d,t}^\star(\psi_t)$; it is never a gradient from the training judge. Optimizer-native gold virtual steps are logged only as a separate preconditioning diagnostic and are not substituted for $g_{d,t}^{\mathrm{aud}}$ in the primary score or the smoothness argument.

Table S6: Gold-graph audit diagnostics at $n = 128$ per domain–phase. GPU-hours are per policy-model phase; across the two policy models and four phases, the domain totals are 64/96/104/120 and the complete audit-gradient budget is 384 GPU-hours.

Domain	$\|g^{\mathrm{aud}}\|_2$	Jackknife var.	Cosine SD	GPU-h/model-phase
Math	0.82	0.17	0.07	8
Code	0.91	0.20	0.09	12
Instruction	0.76	0.24	0.11	13
Science	0.69	0.29	0.13	15

Confirmatory certificate labels come from public official answers, executable tests, or source-backed labels with released construction details. Human-coded open-ended cases are used only for diagnosis and do not enter the certificate, g_t^{aud}, or primary utility. Annotators are blinded to router and verifier identity, and the adjudication rule is fixed before annotation. We report the panel size, labels per item, adjudicator count, qualification criteria, sample size, exclusions, raw agreement, interval estimates, adjudication rate, label latency, and paid hours.

Any human-annotation study reports the applicable institutional review or exemption determination together with recruitment, qualifications, compensation, consent, withdrawal, privacy, data retention, and data-governance procedures.

Audit stability is evaluated at 32/64/128/256 examples per stratum, with label-flip and rubric-mixture perturbations. Folds rotate across phases. A held-out objective that never contributes an audit gradient tests whether high alignment merely memorizes the audit mixture.

The mechanism outcome split is disjoint by source and semantic cluster from the audit-direction, route-training, and certificate folds. Its one-step endpoint is executable task score or a public official task score and never contributes to g_t^{aud}. Table S14 therefore evaluates transfer to an independent endpoint rather than correlating the audit gradient with the objective that generated it.

E GRADIENT AND OPTIMIZER DIAGNOSTICS

The primary policy update is LoRA-only so the routed and realized steps occupy the same trainable space. The default uses rank 16 and a seeded sparse JL map to 256 dimensions. The scoring

displacement includes response-group normalization, reference-policy terms, clipping, frozen optimizer moments, and weight decay if enabled. Norms below 10^{-8} or projected signs within the measured distortion margin are indeterminate and route to escalation or skip.

For each mechanism block, we measure four quantities:

1. ideal full-trainable-space alignment in the representation-only diagnostic;
2. unsketched LoRA-space alignment used to isolate subspace restriction;
3. 64/128/256/512-dimensional JL estimates with realized distortion;
4. alignment of the actual minibatch optimizer step $\psi_{t+1} - \psi_t$ with a fresh audit direction.

The decisive association is with independently evaluated one-step utility, not with another proxy cosine. Per-instance virtual steps are microbatched; the audit vector is computed once per stratum. We report backward passes, GPU-hours, wall time, and memory for every representation. The full-trainable-space row is a representation diagnostic. It is excluded from the evidence obtained with Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct and does not represent a third policy model or a 70B policy-training experiment.

F Cost Analysis

The routing score uses expected incremental dollar-equivalent cost based on a rate card fixed before the runs. API calls use billed prices, and local computation uses measured GPU-hours multiplied by a specified rental rate. The primary experiments use no paid human annotation; any human-coded diagnostic is reported separately and does not alter the cost frontier. This conversion provides a scalar routing cost while raw FLOPs, time, dollars, and annotator-hours remain separately reported resources.

Table S7: Fixed rate card. Monetary rates and provider-specific billing quantities are obtained from invoices and run artifacts rather than inferred from normalized results.

Resource	Raw unit	Unit-rate field	Quantity evidence
Hosted verifier input	input tokens	\$3.00 per million tokens	provider invoice and request log, including retries
Hosted verifier output	output tokens	\$15.00 per million tokens	provider invoice and request log, including retries
Local verifier / policy	accelerator-hours	\$10.00 per accelerator-hour	scheduler log keyed by accelerator type and precision
Audit-gradient computation	accelerator-hours and backward passes	same declared local-compute rate	profiler and scheduler log
Request/sandbox overhead	logged verifier calls	\$0.006 per logged call	request, retry, and sandbox-execution logs
Human gold/audit work	hours	n/a (zero in primary experiments)	public labels and executable checks only
Router fitting	accelerator-hours	\$10.00 per accelerator-hour	training log and versioned model hash

Table S8: Raw incremental resource use at the 128K horizon. Calls include retries; hosted tokens are input/output, GPU-hours are audit/total, and setup share is computed after applying Table S7.

Method	Online calls	Setup calls	Hosted tokens	Audit/total GPU-h	Setup share	Human h
Uncertainty Router	142.6K	32.8K	2.58B / 0.34B	0 / 6,035	1.3%	0
BayesianRouter-H	131.1K	655.4K	2.15B / 0.27B	0 / 5,520	11.4%	0
Pivot	145.2K	655.4K	1.96B / 0.24B	384 / 5,228	14.4%	0

For phase t, incremental cost is

$$C_t^{\text{setup}} = C_t^{\text{route matrix}} + C_t^{\text{certificate matrix}} + C_t^{\text{gold}} + C_t^{\text{audit gradient}} + C_t^{\text{fit}}, \tag{S15}$$

$$C_t^{\text{online}} = c_t^{\text{acquired verifiers}} + c_t^{\text{virtual steps}} + c_t^{\text{JL router}} + c_t^{\text{escalation}} + c_t^{\text{retries}}, \tag{S16}$$

where every lowercase online term is a per-update-prompt rate. At an amortization horizon of H update prompts, the primary cost is

$$\bar{C}_t(H) = \frac{C_t^{\text{setup}}}{H} + C_t^{\text{online}}. \tag{S17}$$

The main comparison fixes the same maximum total incremental cost cap for all methods and reports actual spend. Shared response generation and policy-update compute are matched and itemized separately. We report raw API calls, locally hosted verifier calls, GPU-hours, annotator-hours, generated and update tokens, wall time, peak memory, retries, and escalation. Setup-free and setup-inclusive curves are both shown, together with the break-even horizon at which PIVOT becomes cheaper than a direct baseline. No cost-saving claim is made below that horizon.

G BENCHMARKS, BASELINES, AND IMPLEMENTATION DETAILS

The confirmatory domains are the fixed 500-problem MATH subset released with PRM800K for exact reasoning, a time-segmented LiveCodeBench release for code, executable IFEval constraints for instruction following, and official GPQA Diamond for scientific QA (Lightman et al., 2024; Jain et al., 2025; Zhou et al., 2023; Rein et al., 2024). RewardBench 2 is used only for verifier analysis and is excluded from downstream policy evidence (Malik et al., 2026). The two policy models are `meta-llama/Llama-3.1-8B-Instruct` and `Qwen/Qwen2.5-7B-Instruct`; the latter has 7.61B parameters (Meta Llama, 2024; Qwen Team, 2024b). Each policy model uses a separate rank-16 LoRA adapter while its backbone remains frozen. The models also maintain separate optimizer states, routers, calibrators, and certificates.

Policies, rollouts, and optimization. The primary experiment uses GRPO and evaluates transfer with DPO. Each prompt yields eight responses at temperature 0.8 and top-p 0.95. AdamW uses learning rate 5×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, gradient clipping at 1.0, LoRA $\alpha = 32$, and dropout 0.05. The route-training, certificate, audit-gradient, mechanism-outcome, and policy-update allocations are 4,096, 16,384, 128 per domain, 512 per domain, and 16,384 groups per policy-model phase, respectively. Route training is partitioned by semantic cluster into 3,072 fitting and 1,024 calibration/validation groups. Run metadata retain release IDs, checksums, prompt templates, checkpoint revisions, tokenizer versions, and random seeds. The main experiment trains only the two 7–8B policy adapters; larger checkpoints serve exclusively as frozen verifiers.

Neural verifier checkpoints. Every domain shares the frozen scalar reward model `Skywork/Skywork-Reward-Llama-3.1-8B-v0.2` (Skywork-8B), an 8B sequence-classification reward model based on Llama-3.1-8B-Instruct (Skywork Team, 2024). The frozen mid-cost generative checkpoint is `google/gemma-2-27b-it` (Gemma-27B), used for math pairwise judgments, instruction rubrics, and retrieval-conditioned science assessment (Google, 2024). The escalation-only checkpoint is `Qwen/Qwen2.5-72B-Instruct` (Qwen-72B), with 72.7B total and 70.0B non-embedding parameters (Qwen Team, 2024a). Qwen-72B serves all four domains through separate frozen role prompts and never participates in first-stage selection.

Table S9: Resolved verifier pool. Each domain provides three first-acquisition sources and Qwen-72B as an escalation-only source. Route and certificate folds acquire all four outputs; online acquisition begins from the first three only.

Domain	Rule/test	Shared RM	Third initial source	Matrix total	Escalation-only source
Math	exact/symbolic checker	Skywork-8B	Gemma-27B pairwise	4	Qwen-72B pairwise
Code	sandboxed executable tests	Skywork-8B	deterministic static analysis	4	Qwen-72B execution
Instruction	executable constraints	Skywork-8B	Gemma-27B rubric	4	Qwen-72B rubric
Science	deterministic applicable subset	Skywork-8B	retrieval + Gemma-27B	4	retrieval + Qwen-72B

The code static-analysis source is programmatic and does not call an LLM. For Python tasks it combines `ast.parse`, `py_compile`, syntax and import checks, a frozen Ruff rule set, forbidden-API and unsafe-operation checks, and task-specific structural checks. The instruction checker executes requirements including bullet and sentence counts, required or forbidden strings, prefixes and

suffixes, JSON schemas, capitalization, response-length intervals, and joint formatting constraints. Gemma-27B handles semantic rubric dimensions through a frozen weighted rubric vector with pre-declared critical dimensions and veto thresholds.

The science critic combines a frozen English public-evidence snapshot, BM25 retrieval, a frozen BGE English embedding/reranking checkpoint, a fixed top-k passage set, and Gemma-27B evidence assessment. Escalation reuses the same retrieved documents and replaces only the critic with Qwen-72B, preventing additional evidence access from confounding judge capability. Neither science verifier can access GPQA Diamond terminal labels. For code escalation, Qwen-72B receives the problem, candidate code, execution summary, and public-test diagnostics, while LiveCodeBench hidden-test outcomes remain inaccessible.

All generative judges use temperature 0 and role-specific frozen prompts. Math position-swapped comparisons create an edge only when both orders agree. Position swapping, retrieval snapshots, and retry policies are held fixed across route, certificate, and update folds. The cost of the four-output matrix is included for route and certificate folds; the online policy acquires one initial verifier and may acquire one escalation verifier. Here, *independent* means that Qwen-72B escalation is separate from Gemma-27B first-stage judging through a different checkpoint, frozen role prompt, inference call, and output. It does not assert statistically independent model errors.

Trusted-utility aggregation. For policy model p and domain d, $U_{p,d}$ is 100 times the mean public official or executable score on the terminal test split: exact correctness for math and science, pass@1 for code, and strict prompt-level constraint satisfaction for instruction following. Every prompt has equal weight within a domain. Average trusted utility is the unweighted macro-average of the eight policy-model–domain cells; "worst" is their minimum. No baseline-specific standardization, z-scoring, reward-model score, or verifier confidence enters this metric.

Consequence weights. The deterministic mapping from each public evaluation stratum to ρ_i and the resulting stratum mass is fixed before certificate labels are accessed. The mapping is held constant across domains and phases, preventing retrospective assignment from observed model errors. Incomparable public-label response pairs remain absent from $E_i^\star$. Because strict graph consistency counts every promoted yet unverified edge as an error, the certificate does not reinterpret an omitted edge as acceptable.

Reward Model Routing in Alignment is the direct prior work: it routes reward models during iterative online DPO using offline per-RM reliability estimates and online Bayesian adaptation (Wu & Lu, 2026). PIVOT extends this setting with policy-conditional directional scoring, heterogeneous feedback adaptation, and post-query risk control. Costly generate–rank–verify search allocates verification at inference time (Dughmi et al., 2026); expected-reward routing chooses a serving generator before response production (Hasanaliyev et al., 2025); and TARo uses a learned token-level router to control step-wise reward-model guidance for a frozen language model (Rai et al., 2026). These works motivate pre-query acquisition or define relevant boundaries; we do not treat them as native training-feedback baselines.

Core ablations are: no τ, no r, $r + \tau$, reliability-gated τ, direct learned one-step improvement, $\beta \in \{1, 1.5, 2\}$, no certificate, an unconditional-error-mass certificate, no coverage floor, no abstention, no escalation, no phase renewal, same-actor versus independent verifier, full-information online routing, and full/unsketched/sketched/realized gradients. All reuse rollouts and optimizer states and follow the same cost accounting.

H SUPPLEMENTARY QUANTITATIVE RESULTS

This section provides the quantitative analyses supporting the main text. Tables report point estimates and seed-level values where available. Prompt-level resampling is used only within independent training replicates and does not substitute for training-replicate uncertainty. Certificate bounds, held-out GG-risk, coverage, and cost remain separate because they describe different properties of the method.

Table S10: Baseline implementations. "Adapted" methods retain their native routing signal while feeding the selected feedback through the same graph adapter and policy optimizer.

Method	Matched implementation	Competing explanation
Single RM / static router	same verifier and domain map in every phase	heterogeneity or task labels alone suffice
Conservative ensemble	same verifier pool; mean, WCO, and UWO aggregation	WCO/UWO-style conservative aggregation mitigates reward-model overoptimization (Coste et al., 2024)
Calibration-only (adapted)	calibrate post-query graph consistency; no directional-alignment term	calibrated feedback alone explains the gain (Leng et al., 2025)
Uncertainty Router	official pairwise uncertainty score; same fast RM, strong judge, judge-call cap, and total-cost cap	pairwise uncertainty alone is sufficient for escalation (Xu et al., 2025b)
SCRC-I admission gate (adapted)	common-pool graph-error scores with a conformal threshold and the same coverage floor; no audit-gradient feature	selective risk control alone suffices (Xu et al., 2025a)
CFA weighting (adapted)	same preference and reliability labels; weight GRPO/DPO feedback by conformal answer reliability, without τ or sequential certification	reliability weighting, rather than directional routing, explains the gain (Chen et al., 2026)
Proactive conformal gate (adapted)	same fast/strong pair and quota; learn the cheap-feedback acceptance set on disjoint calibration data and otherwise acquire strong feedback	an inference-side cheap/strong guarantee transfers to training acquisition (Uddin & Bauer, 2026)
VI-style feedback router (adapted)	common pool and pre-query features; maximize predicted residual gold-graph-error reduction per cost, without an audit-gradient signal	value of information alone explains routing (Saliencro et al., 2026)
BayesianRouter (native RM pool)	official offline multi-task RM-reliability initialization and online Thompson selection; same RM subset, pairs, tokens, and cost cap	adaptive RM reliability is sufficient without policy-conditional directional scoring (Wu & Lu, 2026)
BayesianRouter-H (common pool)	strict graph consistency is the Bernoulli observation for every base verifier in Table S9; same response groups, label budget, tuning groups, and cost cap	pool access, rather than reliable directional alignment, explains the gain
Judge-panel router (adapted)	cost-adjusted conditional broad/slice gain and held-out stopping; same judge pool, deployable slice keys, and cost cap	evaluator complementarity and specialization suffice without policy-conditional directional scoring (Zhu et al., 2026)
SCOPE/BPE selector (adapted)	two order-swapped judge calls, preference-token probabilities, ties excluded, native marginal FDR target; both calls included in cost	pairwise selection without optimizer-native directional scoring is sufficient (Badshah et al., 2026)
Conformal Arbitrage (adapted)	primary–guardian routing under the same quota and cost	deployment-time arbitration transfers to training feedback (Overman & Bayati, 2025)
DRRO	same RM, on-policy responses, tokens, update count, and compute cap; replace routing with the DRRO objective	robust regret optimization alone handles reward misspecification (Wang et al., 2026)
E2H curriculum	official easy/hard partition and scheduling rule selected on validation; same admitted count and update budget	scheduled curriculum, not verifier value, drives improvement (Parashar et al., 2026)
Oracle	gold audit chooses the action	non-deployable upper reference

Table S11: Information and resource access across baselines. "Equal" denotes the same maximum resource; unused method-specific supervision does not provide additional verifier calls. Native BayesianRouter is retained for implementation fidelity, while its common-pool counterpart isolates verifier access.

Method	Online verifier pool	Route-matrix groups	Gold route labels	Audit directions	Certificate labels	Tuning / total-cost cap
PIVOT	common pool	equal	equal	available	equal	equal / equal
PIVOT–RM	scalar-RM subset	equal over subset	equal	available	equal	equal / equal
Calibration only	common pool	equal	equal	unavailable by definition	equal	equal / equal
BayesianRouter (native RM pool)	scalar-RM subset	equal over subset	equal	unavailable by definition	none	equal / equal
BayesianRouter-H (common pool)	common pool	equal	equal	unavailable by definition	none	equal / equal
Uncertainty Router	fast RM + strong judge	equal for supported pair	equal	unavailable by definition	none	official grid / equal
SCRC-I admission	common pool	equal	equal error labels	unavailable by definition	equal conformal fold	equal / equal
CFA weighting	common pool	equal	equal reliability labels	unavailable by definition	none; equal calibration	equal / equal
Proactive conformal gate	fast RM + strong judge	equal for supported pair	equal cheap-success labels	unavailable by definition	equal native calibration	official grid / equal
VI-style feedback router	common pool	equal	equal error labels	unavailable by definition	none	equal / equal
Conservative ensemble	common pool	equal	equal	unavailable by definition	none	equal / equal

H.1 COST-CAPPED OUTCOMES AND DOMAIN HETEROGENEITY

Table S12: Extended cost-accounted comparison under a common budget cap. "Worst" is the minimum policy-model–domain slice. Coverage is the raw fraction admitted after the post-query gate, and cost is setup-inclusive at 128K update prompts across both policy models. The oracle is not a deployable comparator.

Method	Utility↑	Worst↑	GG-risk↓	Coverage↑	Cost↓
Single learned RM	63.4	55.2	0.116	1.00	0.37
Static task router	64.5	56.7	0.094	0.94	0.43
WCO/UWO ensemble	65.2	57.3	0.073	0.86	0.66
Calibration only	66.7	58.9	0.058	0.83	0.50
Uncertainty Router	66.1	58.5	0.067	0.88	0.58
CFA weighting	66.9	59.1	0.060	0.89	0.50
Proactive conformal gate	66.4	59.0	0.045	0.72	0.59
SCRC-I admission gate	66.2	58.8	0.043	0.68	0.60
VI-style feedback router	67.2	60.0	0.057	0.84	0.54
BayesianRouter (native RM pool)	66.3	58.8	0.065	0.87	0.50
PIVOT–RM (native RM pool)	67.7	60.0	0.050	0.83	0.53
BayesianRouter-H (common pool)	66.8	59.2	0.063	0.86	0.55
Judge-panel router	67.0	59.7	0.061	0.85	0.56
SCOPE/BPE selector	66.2	58.9	0.046	0.74	0.65
SAFER filter	65.9	58.1	0.062	0.77	0.57
Conformal Arbitrage (adapted)	66.0	58.3	0.044	0.70	0.61
RLCF	66.0	57.9	0.075	0.91	0.49
RaR	66.7	59.0	0.069	0.88	0.54
DRRO	66.5	58.0	0.083	1.00	0.39
E2H	66.6	58.4	0.086	0.96	0.44
PIVOT	68.4	61.1	0.047	0.82	0.52
Oracle router	71.0	65.5	0.023	0.90	0.79

Table S13: Policy-model and domain decomposition. Short model names denote the corresponding Instruct checkpoints. Each model mean is the unweighted average of the four domain entries. "Rule" cells in open-ended tasks elsewhere refer only to the subset on which a deterministic check is applicable.

Policy model	Router	Math	Code	Instruction	Science	Mean
Llama-3.1-8B	Uncertainty	70.2	63.0	70.5	58.5	65.6
	BayesianRouter-H	70.8	63.7	71.2	59.2	66.2
	PIVOT	72.2	65.0	72.2	61.1	67.6
	Oracle	74.5	67.5	74.5	65.5	70.5
Qwen2.5-7B	Uncertainty	71.0	64.0	71.0	60.6	66.7
	BayesianRouter-H	71.6	64.7	71.6	61.6	67.4
	PIVOT	73.0	66.0	73.0	64.9	69.2
	Oracle	75.5	68.5	75.5	66.5	71.5

Router	Minimum pooled-domain mean	Worst policy-model–domain slice
Uncertainty Router	59.6	58.5
BayesianRouter-H	60.4	59.2
PIVOT	63.0	61.1
Oracle	66.0	65.5

Among deployable methods in Table S12, PIVOT attained the highest average utility, 68.4, and the highest worst-slice utility, 61.1. The native RM-pool comparison was 67.7 for PIVOT–RM and 66.3 for BayesianRouter, a 1.4-point difference. The common-pool comparison was 68.4 for PIVOT and 66.8 for BayesianRouter-H, a 1.6-point difference. SCRC-I, SCOPE/BPE, and Conformal Arbitrage achieved lower GG-risk values, from 0.043 to 0.046, while admitting only 0.68 to 0.74 of groups. The oracle row remains an upper reference because it selects actions with gold audit information.

The policy-model–domain decomposition shows that the aggregate improvement was retained for both policy models. PIVOT improved the model mean from 66.2 to 67.6 for Llama-3.1-8B-Instruct and from 67.4 to 69.2 for Qwen2.5-7B-Instruct. Science was the limiting domain, with PIVOT utilities of 61.1 and 64.9 for Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, respectively.

H.2 MECHANISM, DRIFT, SEEDS, AND OPTIMIZER TRANSFER

Table S14: Association with matched one-step trusted improvement. Entries are domain-level Spearman coefficients on the source-disjoint mechanism outcome split, pooled across frozen checkpoints and the two policy models; "mean" is the unweighted macro-average. The last column is a 95% checkpoint-blocked bootstrap interval. No mechanism outcome forms the audit direction or a route-training label.

Signal	Math	Code	Instruction	Science	Mean	95% CI
Strict gold-graph consistency	0.24	0.20	0.15	0.18	0.19	[0.12,0.26]
Non-contradiction consistency C^{nc}	0.29	0.25	0.26	0.28	0.27	[0.20,0.34]
Predictive uncertainty	0.28	0.26	0.24	0.27	0.26	[0.19,0.33]
Sample difficulty	0.18	0.16	0.12	0.14	0.15	[0.08,0.22]
JL update alignment τ	0.52	0.49	0.43	0.46	0.48	[0.41,0.54]
Reliability-weighted norm-aware first order	0.54	0.51	0.46	0.49	0.50	[0.44,0.56]
Direct one-step predictor $\hat{\Delta}$	0.55	0.52	0.47	0.51	0.51	[0.45,0.57]
Reliability-adjusted directional score $r\tau$	0.58	0.55	0.50	0.52	0.54	[0.48,0.59]

Relaxing unsupported but uncontradicted edges increased the consistency–utility association from 0.19 to 0.27. Paired checkpoint-blocked bootstrap differences were 0.21 [0.13, 0.29] for $\tau - C^{\mathrm{nc}}$ and 0.27 [0.18, 0.35] for $r\tau - C^{\mathrm{nc}}$. The intervals for norm-aware first order and the direct predictor overlap. The decomposition below shows that the $r\tau$ association is retained for both policy models and all five seeds, with domain coefficients from 0.50 to 0.58.

The label-transition shares were 81% for $C = 1, C^{\mathrm{nc}} = 1$, 6% for $C = 0, C^{\mathrm{nc}} = 1$, and 13% for $C = 0, C^{\mathrm{nc}} = 0$. Positive one-step utility occurred in 59%, 57%, and 34% of these strata, respectively. The relaxed label therefore recovers useful feedback from the 6% reclassified stratum, while its 0.27 association leaves most of the update-utility variation unresolved.

Table S15: Policy-model and seed decomposition of the three strongest mechanism signals. Short model names denote the corresponding Instruct checkpoints. Each entry is a checkpoint-blocked Spearman coefficient.

Signal	Policy model	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean
Norm-aware	Llama-3.1-8B	0.47	0.50	0.48	0.53	0.49	0.49
	Qwen2.5-7B	0.49	0.52	0.50	0.55	0.51	0.51
Direct $\hat{\Delta}$	Llama-3.1-8B	0.48	0.51	0.49	0.54	0.50	0.50
	Qwen2.5-7B	0.50	0.53	0.51	0.56	0.52	0.52
$r\tau$	Llama-3.1-8B	0.50	0.55	0.52	0.57	0.54	0.54
	Qwen2.5-7B	0.51	0.56	0.54	0.55	0.52	0.54

Table S16: Routing $\times$ realized-loss weighting factorial. Bayesian routing uses the common four-output matrix; PIVOT routing uses the policy-conditional score. Unit and $r\tau$ weighting share identical admitted groups within each router.

Router	Weight	Utility	GG-risk	Coverage	Cost
BayesianRouter-H	unit	66.5	0.050	0.80	0.545
BayesianRouter-H	$r\tau$	67.2	0.048	0.80	0.550
PIVOT	unit	67.6	0.049	0.82	0.515
PIVOT	$r\tau$	68.4	0.047	0.82	0.520

This factorial isolates the pre-query router from the realized-loss weight. With unit weights, replacing BayesianRouter-H by PIVOT routing increased utility from 66.5 to 67.6 at similar GG-risk. Under identical $r\tau$ weighting, the same routing change increased utility from 67.2 to 68.4. The resulting 1.1–1.2-point gap shows that the learned source choice remains consequential after post-query weighting is matched.

Table S17: Verifier strict graph consistency changes with policy state. Each cell pair is phase 1 / phase 3. Rule/test values for instruction and science are computed only on their deterministic-check subsets.

Domain	Verifier	Phase 1	Phase 3
Math	Rule/test	0.993	0.991
	Learned RM	0.824	0.771
	Pairwise judge	0.892	0.861
	Retrieval critic	0.842	0.817
Code	Rule/test	0.982	0.975
	Learned RM	0.806	0.752
	Pairwise judge	0.865	0.831
	Retrieval critic	0.814	0.781
Instruction	Rule subset	0.891	0.858
	Learned RM	0.789	0.721
	Rubric judge	0.844	0.798
	Retrieval critic	0.806	0.764
Science	Rule subset	0.781	0.724
	Learned RM	0.761	0.681
	Pairwise judge	0.829	0.763
	Retrieval critic	0.861	0.792

Table S18: Seed-level results and optimizer transfer. Short model names denote the corresponding Instruct checkpoints. Panel (a) reports all five GRPO seed estimates for the direct comparator. Panel (b) applies the same routing procedure under GRPO and DPO; no-τ uses reliability only.

(a) GRPO average trusted utility by seed

Policy model	Router	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean
Llama-3.1-8B	BayesianRouter-H	65.8	66.4	66.1	66.5	66.3	66.2
Llama-3.1-8B	PIVOT	67.2	67.8	67.4	68.0	67.6	67.6
Qwen2.5-7B	BayesianRouter-H	67.0	67.6	67.3	67.7	67.4	67.4
Qwen2.5-7B	PIVOT	68.7	69.2	69.0	69.6	69.5	69.2

(b) Optimizer transfer

Optimizer	Router	Utility	Worst	GG-risk	Coverage	Cost
GRPO	Uncertainty Router	66.1	58.5	0.067	0.88	0.58
GRPO	BayesianRouter-H	66.8	59.2	0.063	0.86	0.55
GRPO	PIVOT without τ	66.5	58.9	0.049	0.80	0.50
GRPO	PIVOT	68.4	61.1	0.047	0.82	0.52
DPO	Uncertainty Router	64.8	57.2	0.068	0.87	0.57
DPO	BayesianRouter-H	65.9	58.2	0.064	0.85	0.54
DPO	PIVOT without τ	66.1	58.5	0.052	0.79	0.49
DPO	PIVOT	67.0	59.4	0.049	0.81	0.51

PIVOT exceeded BayesianRouter-H for every retained seed for both policy models. The optimizer-transfer panel retained the same ordering: under DPO, PIVOT reached utility 67.0 with GG-risk 0.049, while the reliability-only variant reached 66.1 with GG-risk 0.052.

H.3 FINITE-SAMPLE GG-RISK ACCOUNTING AND PHASE RENEWAL

Table S19: Paired uncertainty for confirmatory estimands. The first three rows use a hierarchical bootstrap that resamples paired seeds and then prompts within domain. The held-out GG-risk interval resamples admitted groups and is reported separately from the finite-grid certificate bound.

Estimand	Estimate	95% lower	95% upper
Utility: PIVOT − BayesianRouter-H	1.6	1.2	2.0
Utility: PIVOT − Uncertainty Router	2.3	1.8	2.7
Worst-slice: PIVOT − BayesianRouter-H	1.9	0.5	3.3
Held-out GG-risk of PIVOT	0.047	0.041	0.054

The action-policy set is fixed at $|\Lambda| = 24$, the global allocation is $\delta_{\text{all}} = 0.05$ across $T = 8$ freshly certified policy-model phases, and the coverage floor is $\kappa = 0.50$. The first three phases of each policy model form the six in-distribution confirmatory phases; phase 4 for each model uses a fresh post-shift certificate. Stale diagnostics consume no confidence allocation and carry no certificate claim.

Table S20: Certificate quantities and held-out GG-risk. Panel (a) gives every quantity entering the Hoeffding certificate. Panel (b) separately reports the held-out numerator and admission denominator. Stale reuse appears only in panel (b) because it has no valid certificate on the shifted distribution.

(a) Certificate fold

Setting	α	n_{cert}	ϵ	$\hat{N}$	$\hat{D}$	U^{sel}	D^L
Tight budget	0.03	32,768	0.011684	0.003106	0.521684	0.029	0.51
Main budget	0.05	16,384	0.016523	0.011897	0.596523	0.049	0.58
Relaxed budget	0.10	16,384	0.016523	0.046177	0.676523	0.095	0.66
Shift, renewed	0.05	16,384	0.016523	0.010917	0.576523	0.049	0.56

(b) Held-out fold

Setting	N	D	GG-risk	Coverage	Utility	Certificate status
Tight budget	0.0143	0.5500	0.026	0.65	66.3	issued
Main budget	0.0301	0.6400	0.047	0.82	68.4	issued
Relaxed budget	0.0655	0.7200	0.091	0.93	69.0	issued
Shift, stale	0.0549	0.6700	0.082	0.86	67.9	invalid after shift
Shift, renewed	0.0298	0.6200	0.048	0.80	68.0	issued

The 0.05 operating point therefore reflected a selected trade-off rather than a minimum-error setting. Tightening the budget reduced held-out GG-risk to 0.026 and coverage to 0.65. Relaxing it raised coverage to 0.93 and utility to 69.0, while held-out GG-risk increased to 0.091.

Table S21: Certificate renewal for each policy model. Short model names denote the corresponding Instruct checkpoints. "Stale GG-risk" evaluates the old gate on the new phase and is excluded from Proposition S1. The first six fresh rows are the in-distribution confirmatory phases; all eight freshly certified rows satisfy the declared operating point.

Policy model	Phase	U^{sel}	Held GG-risk	D^L	Coverage	Stale GG-risk	Utility	Cost
Llama-3.1-8B	1	0.047	0.043	0.58	0.81	0.043	65.9	0.48
Llama-3.1-8B	2	0.049	0.046	0.56	0.80	0.064	66.5	0.50
Llama-3.1-8B	3	0.048	0.045	0.55	0.82	0.074	67.1	0.52
Llama-3.1-8B	4 (shift)	0.049	0.047	0.54	0.78	0.085	67.6	0.56
Qwen2.5-7B	1	0.048	0.045	0.60	0.83	0.045	67.3	0.48
Qwen2.5-7B	2	0.049	0.048	0.58	0.82	0.058	68.0	0.50
Qwen2.5-7B	3	0.049	0.047	0.57	0.84	0.068	68.7	0.52
Qwen2.5-7B	4 (shift)	0.049	0.049	0.58	0.82	0.079	69.2	0.56

Table S22: Candidate selected in each phase. Short model names denote the corresponding Instruct checkpoints. "Feasible" is the number of the 24 pre-specified candidates satisfying both the GG-risk bound and coverage floor before tie-breaking.

Policy model	Phase	ID	Feasible	λ_c	γ_q	γ_τ	e
Llama-3.1-8B	1	04	9	0.00	0.65	0.10	1
Llama-3.1-8B	2	12	7	0.05	0.65	0.10	1
Llama-3.1-8B	3	20	5	0.10	0.65	0.10	1
Llama-3.1-8B	4	24	3	0.10	0.75	0.10	1
Qwen2.5-7B	1	04	10	0.00	0.65	0.10	1
Qwen2.5-7B	2	12	8	0.05	0.65	0.10	1
Qwen2.5-7B	3	20	6	0.10	0.65	0.10	1
Qwen2.5-7B	4	24	4	0.10	0.75	0.10	1

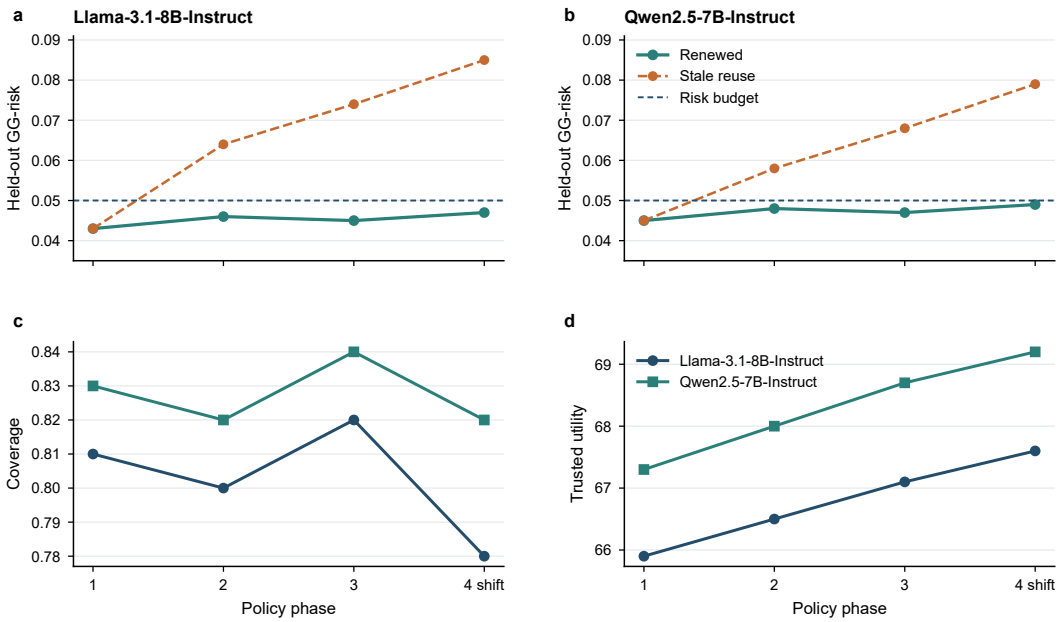


Figure S1: Phase renewal across the two policy models. (a,b) Renewed and fully stale routes are evaluated on the same frozen phase; the dashed line is the nominal GG-risk budget. (c,d) Coverage remains non-trivial while trusted utility rises across renewed phases. Stale reuse is a boundary diagnostic and carries no certificate claim.

Table S23: Verifier graph-consistency and calibration drift. Accuracy, Brier score, and ECE are shown together so confidence recalibration is not conflated with recovery of intrinsic verifier competence. Renewed columns are evaluated after refitting the calibrator, not the verifier.

	Base			Shift, stale			Shift, renewed	
Verifier	Acc.	Brier	ECE	Acc.	Brier	ECE	Brier	ECE
Rule/test	0.987	0.011	0.007	0.974	0.024	0.018	0.022	0.010
Scalar RM	0.803	0.151	0.032	0.724	0.210	0.093	0.194	0.034
Pairwise judge	0.862	0.112	0.025	0.821	0.148	0.065	0.137	0.030
Rubric judge	0.839	0.129	0.033	0.792	0.168	0.075	0.153	0.036
Retrieval critic	0.846	0.126	0.034	0.798	0.166	0.073	0.153	0.037

Renewal preserved the operating point across all eight policy-model phases, with held-out GG-risk from 0.043 to 0.049. The stale gate degraded with policy shift, reaching 0.085 for Llama-3.1-8B-Instruct and 0.079 for Qwen2.5-7B-Instruct during phase 4. Recalibration improved ECE after drift for every verifier, yet it did not restore the verifiers' original accuracy.

H.4 Ablations, Routing Behavior, and Approximation Cost

Table S24: Canonical graph versus modality-native training. Utility and coverage use matched verifier outputs and update budgets. Δ is graph minus native utility; adapter abstention is measured before the PIVOT post-query gate.

Source	Native util.	Graph util.	Δ	Native cov.	Graph cov.	Adapter abst.
Exact answer / tests	70.8	70.6	−0.2	0.91	0.90	0.081
Scalar reward model	66.3	66.0	−0.3	0.88	0.84	0.107
Pairwise judge	67.0	66.8	−0.2	0.83	0.80	0.152
Rubric vector	68.0	67.2	−0.8	0.86	0.74	0.186
Retrieval critic	66.8	65.9	−0.9	0.78	0.65	0.257

Table S25: Adapter outcomes. The first five columns are mutually exclusive first-failure codes; top ties are separated from other non-unique maxima. Adapter pass precedes the post-query gate, final admit follows it, and edge density is relative to $K(K-1)$ directed edges.

Source	Empty/invalid	Cycle	Top tie	Other nonunique	Veto	Adapter pass	Final admit	Edge density
Exact answer / tests	2.8%	0.0%	3.0%	1.1%	1.2%	91.9%	86.2%	0.61
Scalar reward model	0.6%	0.0%	6.2%	2.5%	1.4%	89.3%	78.5%	0.54
Pairwise judge	1.7%	3.9%	5.1%	2.4%	2.1%	84.8%	72.4%	0.42
Rubric vector	1.2%	2.8%	6.8%	3.6%	4.2%	81.4%	69.1%	0.47
Retrieval critic	8.9%	2.3%	5.9%	3.8%	4.8%	74.3%	60.7%	0.36
Escalation panel	1.1%	1.8%	3.4%	2.2%	3.3%	88.2%	76.5%	0.49

Table S26: Adapter pass-through by domain. Open-ended instruction and science groups have lower final admission and sparser gold-supported graphs; this is an intended limitation rather than a uniformly favorable result.

Domain	Adapter pass	Final admit	Edge density	GG-risk
Math	92.0%	87.0%	0.58	0.036
Code	90.0%	85.0%	0.55	0.040
Instruction	82.0%	74.0%	0.44	0.051
Science	78.0%	69.0%	0.39	0.061

The adapter was most reliable for exact and executable feedback. Math and code achieved adapter pass rates of 92.0% and 90.0%, whereas instruction and science achieved 82.0% and 78.0%. Science combined the sparsest accepted graphs, edge density 0.39, with the lowest final admission rate, 69.0%, and GG-risk 0.061.

Table S27: Complete ablation frontier. The four columns are jointly reported; no single one is treated as a sufficient ranking. Full-information routing is a non-deployable diagnostic whose additional calls are included in cost.

Variant	Utility↑	GG-risk↓	Coverage↑	Cost↓
Full PIVOT ($\beta = 1.5$)	68.4	0.047	0.82	0.52
Without τ	66.5	0.049	0.80	0.50
Without reliability margin r	67.0	0.079	0.88	0.49
Direct learned one-step improvement	67.7	0.058	0.84	0.51
Additive $r + \tau$	67.4	0.055	0.85	0.51
Reliability-gated τ	66.9	0.050	0.78	0.50
$\beta = 1$	68.6	0.054	0.85	0.51
$\beta = 2$	68.2	0.044	0.80	0.54
Without certificate	68.8	0.084	0.94	0.47
Unconditional-error-mass certificate	67.3	0.061	0.63	0.47
Without coverage floor	65.9	0.031	0.49	0.48
Without abstention	68.0	0.074	1.00	0.55
Without escalation	67.8	0.052	0.77	0.47
Two escalation opportunities	68.6	0.045	0.84	0.63
Without phase renewal	67.3	0.071	0.86	0.50
Same-actor verifier	67.1	0.078	0.88	0.45
Full-information online routing	69.1	0.043	0.86	0.91
Unsketched LoRA	68.6	0.046	0.82	0.60

The ablation frontier separates the main components. Removing directional alignment reduced utility from 68.4 to 66.5. Removing the reliability margin increased GG-risk from 0.047 to 0.079. A certificate-free router reached utility 68.8, yet GG-risk increased to 0.084. Disabling escalation reduced utility by 0.6 points and coverage by 0.05 while lowering cost by 0.05; allowing two escalations recovered only 0.2 additional utility points at cost 0.63. Together with the 5–17% escalated share in Table S30, this identifies Qwen-72B as a bounded repair mechanism rather than the domi-

nant route. The coverage floor prevented a low-risk, low-coverage solution, which attained GG-risk 0.031 at only 0.49 coverage.

Table S28: Sensitivity to reliability penalty, response-group size, and adapter margin. Margin values multiply each verifier's frozen validation-scale γ_v. The primary setting is $\beta = 1.5$, $K = 8$, and $1.0\gamma_v$.

Axis	Setting	Utility	GG-risk	Coverage	Cost
β	1.0	68.6	0.054	0.85	0.51
	1.5	68.4	0.047	0.82	0.52
	2.0	68.2	0.044	0.80	0.54
K	4	67.6	0.050	0.79	0.47
	8	68.4	0.047	0.82	0.52
	16	68.5	0.046	0.83	0.64
Margin	$0.5\gamma_v$	68.5	0.055	0.87	0.50
	$1.0\gamma_v$	68.4	0.047	0.82	0.52
	$1.5\gamma_v$	67.9	0.043	0.75	0.55

Table S29: Pre-query model, feature, and consistency-calibration ablation. $\hat{m}$ correlation is measured on the independent mechanism outcome split; Brier and ECE evaluate post-query $\hat{q}$.

Variant	$\hat{m}$ corr.	$\hat{q}$ Brier	$\hat{q}$ ECE	Utility	GG-risk	Cost
Full GBT + isotonic	0.51	0.126	0.028	68.4	0.047	0.52
Linear $\hat{m}$ + isotonic	0.42	0.127	0.029	67.5	0.051	0.51
Two-layer MLP $\hat{m}$ + isotonic	0.49	0.128	0.030	68.1	0.049	0.54
Without prompt embedding	0.47	0.127	0.029	68.0	0.049	0.51
Without rollout statistics	0.38	0.130	0.034	67.2	0.055	0.49
Without lagged calibration history	0.43	0.145	0.057	67.4	0.063	0.50
Without capability/cost features	0.48	0.127	0.030	68.2	0.048	0.61
Uncalibrated $\hat{q}$	0.51	0.141	0.071	67.6	0.069	0.50
Temperature-scaled $\hat{q}$	0.51	0.131	0.039	68.0	0.054	0.51

The primary settings balance performance and cost. Increasing K from 8 to 16 changed utility from 68.4 to 68.5, while cost increased from 0.52 to 0.64. Raising β from 1.5 to 2.0 lowered GG-risk from 0.047 to 0.044 and reduced coverage from 0.82 to 0.80. The GBT plus isotonic model combined the lowest reported ECE, 0.028, with the highest utility.

Table S30: Final routing outcomes by domain. The first three columns are feedback admitted from the initial verifier, "escalated" is admitted after the second verifier, and "skip" combines pre-query abstention with final post-query rejection. Rows are mutually exclusive and sum to 100%.

Domain	Rule/test	Learned RM	Judge/critic	Escalated	Final skip
Math	42%	24%	12%	7%	15%
Code	49%	20%	10%	5%	16%
Instruction	16%	29%	23%	14%	18%
Science	9%	26%	25%	17%	23%

Routing behavior tracked these modality differences. Initial rule or test feedback accounted for 42% of math admissions and 49% of code admissions. Instruction and science relied more on learned judges and escalation, with escalation rates of 14% and 17%, respectively.

Table S31: Fresh/stale 2×2 renewal diagnostic. Stale $\hat{q}$ primarily breaks GG-risk control, whereas stale $\hat{m}$ primarily reduces utility; full renewal refits both.

$\hat{m}$	$\hat{q}$	$\hat{m}$ corr.	Utility	GG-risk	Coverage	Cost
fresh	fresh	0.51	68.4	0.047	0.82	0.52
stale	fresh	0.39	67.5	0.049	0.79	0.50
fresh	stale	0.51	67.8	0.071	0.87	0.50
stale	stale	0.31	66.9	0.082	0.86	0.48

Table S32: PIVOT operating point by policy model and domain. Short model names denote the corresponding Instruct checkpoints. Science remains the least covered and most expensive domain, so the aggregate gain is not presented as uniform across task types.

Policy model	Domain	Utility	GG-risk	Coverage	Cost
Llama-3.1-8B	Math	72.2	0.040	0.86	0.47
Llama-3.1-8B	Code	65.0	0.043	0.85	0.48
Llama-3.1-8B	Instruction	72.2	0.051	0.80	0.55
Llama-3.1-8B	Science	61.1	0.056	0.77	0.58
Qwen2.5-7B	Math	73.0	0.039	0.87	0.48
Qwen2.5-7B	Code	66.0	0.042	0.86	0.49
Qwen2.5-7B	Instruction	73.0	0.049	0.81	0.54
Qwen2.5-7B	Science	64.9	0.055	0.78	0.57

Table S33: Audit and gradient sensitivity. Panel (a) changes trusted audit size per stratum. Panel (b) changes JL dimension; its association is the pooled compression diagnostic rather than the domain macro-average in Table S14. Panel (c) reports implementation overhead.

(a) Audit size

n/stratum	In-mixture corr.	Held corr.	Cosine SD	Utility	GG-risk	Rel. cost
32	0.35	0.27	0.19	67.1	0.057	1.05×
64	0.44	0.36	0.14	67.8	0.052	1.08×
128	0.51	0.43	0.10	68.4	0.047	1.12×
256	0.53	0.44	0.08	68.5	0.046	1.20×

(b) JL dimension

Dimension	Pooled corr.	Sign agree	Max distortion	Indeterminate	Route flip	Utility	Rel. cost
64	0.43	82%	0.116	12.4%	9.8%	67.6	1.10×
128	0.48	88%	0.076	8.1%	6.0%	68.1	1.13×
256	0.52	93%	0.049	4.7%	2.9%	68.4	1.18×
512	0.53	96%	0.032	2.8%	1.6%	68.5	1.29×
Unsketched	0.54	100%	0.000	0%	0.0%	68.6	1.31×

(c) Implementation overhead

Representation	Wall time	Peak memory	Utility
No gradient	1.00×	1.00×	66.5
LoRA + JL-256	1.18×	1.11×	68.4
Unsketched LoRA	1.31×	1.24×	68.6
Full trainable space	1.67×	1.52×	68.7

With JL-256, the projected score preserved 93% sign agreement, had maximum measured distortion 0.049, and changed only 2.9% of routes. It reached utility 68.4 at 1.18× relative cost. Unsketched LoRA reached 68.6 utility at 1.31× cost, so the primary projection retained most of the observed signal at lower implementation cost.

H.5 Distribution Shift, Cost, Consequence Weights, and Failures

Table S34: LoRA-rank sensitivity with JL dimension fixed at 256. Rank 16 is the primary setting; larger ranks offer negligible utility change while increasing cost, so the largest subspace is not uniformly preferable.

Rank	$r\tau$ corr.	Utility	GG-risk	Coverage	Cost
4	0.42	67.4	0.052	0.80	0.48
8	0.47	68.0	0.049	0.81	0.50
16	0.51	68.4	0.047	0.82	0.52
32	0.52	68.5	0.046	0.82	0.58
64	0.52	68.4	0.046	0.81	0.69

Table S35: Sensitivity to structured verifier error. Correlated errors shared across learned verifier outputs and shared retrieval bias are harder than symmetric flips and can violate the nominal 0.05 operating point. These rows diagnose a limitation of consistency calibration under shared misspecification.

Perturbation	Utility	GG-risk	Coverage	$\hat{q}$ ECE
Clean	68.4	0.047	0.82	0.028
Symmetric label flips	68.1	0.049	0.81	0.034
Correlated learned-verifier error	66.9	0.069	0.85	0.071
Omitted rubric dimension	67.2	0.061	0.84	0.059
Shared retrieval-source bias	66.7	0.073	0.86	0.078
Position bias	67.6	0.053	0.82	0.043

Table S36: Shift boundary and conservative fallback. Renewed GG-risk is reported only when the coverage floor remains feasible. For open-world evidence shifts that fail the floor, the displayed value is fallback GG-risk and is not a certificate under Proposition S1.

Shift	Stale GG-risk	Renew/fallback GG-risk	Coverage	Utility	Cost	Status
Prompt paraphrase	0.058	0.049	0.79	67.9	0.54	renewed
Policy checkpoint	0.082	0.048	0.80	68.0	0.56	renewed
Domain mixture	0.073	0.050	0.76	67.6	0.57	renewed
Verifier-prompt drift	0.076	0.049	0.75	67.2	0.58	renewed
New science source	0.104	0.056	0.61	65.9	0.60	fallback: infeasible
Conflicting retrieval snapshot	0.118	0.063	0.54	64.8	0.63	fallback: infeasible

Structured shared errors formed the main robustness boundary. Correlated error across learned verifier outputs raised GG-risk to 0.069, and shared retrieval-source bias raised it to 0.073. After renewal, prompt paraphrase, policy-checkpoint, domain-mixture, and verifier-prompt shifts returned to approximately the 0.05 target. New science sources and conflicting retrieval snapshots remained infeasible under the declared coverage floor.

Table S37: Setup-inclusive amortized incremental cost. Panel (a) reports dollar-equivalent cost per update prompt and panel (b) reports the corresponding total implied by the frozen rate card. The fixed specialist has no routing setup and can be preferable for short or narrow deployments despite lower aggregate utility.

(a) Dollar-equivalent cost per update prompt

Update prompts	PIVOT	Uncertainty Router	BayesianRouter-H	Fixed specialist
8K	1.645	0.697	1.488	0.490
16K	1.045	0.635	0.988	0.490
32K	0.745	0.603	0.738	0.490
64K	0.595	0.588	0.613	0.490
128K	0.520	0.580	0.550	0.490

(b) Dollar-equivalent total

Update prompts	PIVOT	Uncertainty Router	BayesianRouter-H	Fixed specialist
8K	13,160	5,578	11,900	3,920
16K	16,720	10,155	15,800	7,840
32K	23,840	19,310	23,600	15,680
64K	38,080	37,621	39,200	31,360
128K	66,560	74,242	70,400	62,720

	Setup	Online	Break-even vs uncertainty	Break-even vs BayesianRouter-H
PIVOT	9,600	0.4450	67,610	37,647
BayesianRouter-H	8,000	0.4875	n/a	n/a
Uncertainty Router	1,000	0.5722	n/a	n/a
Fixed domain specialist	0	0.4900	n/a	n/a

128K component	Setup	Rule/test	Learned RM	Strong judge	Escalation	Gradient/router
Cost per update prompt	0.075	0.021	0.121	0.117	0.096	0.090

Table S38: Full-matrix route-label scaling with verifier-pool size. Pool 4 is the primary setting. Pool 8 yields only a small utility increment at twice the matrix cost. Half-matrix and logged sequential rows are diagnostics and do not inherit Proposition S1 because some counterfactual candidates are unobserved.

Route-label design	Setup calls	Utility	GG-risk	Coverage	Cost	Status
Full matrix, pool 2	327,680	67.6	0.049	0.80	0.46	certified
Full matrix, pool 4	655,360	68.4	0.047	0.82	0.52	certified
Full matrix, pool 8	1,310,720	68.7	0.046	0.83	0.66	certified
Half matrix, pool 4	327,680	68.0	0.051	0.79	0.46	diagnostic
Logged sequential	0	67.8	0.052	0.80	0.44	diagnostic

Table S39: Rate-card sensitivity. Each alternative changes one unit-rate category by $\times 0.5$ or $\times 1.5$ while holding raw usage fixed. Costs are dollar-equivalent per update prompt; crossover columns are update-prompt horizons.

Scenario	PIVOT	BayesianRouter-H	Uncertainty	B/E vs unc.	B/E vs BayesianRouter-H
Frozen reference	0.520	0.550	0.580	67,610	37,647
Hosted API $\times 0.5$	0.452	0.481	0.503	65,900	32,800
Hosted API $\times 1.5$	0.588	0.619	0.657	68,800	41,700
Local GPU $\times 0.5$	0.498	0.542	0.572	56,300	14,200
Local GPU $\times 1.5$	0.542	0.558	0.588	79,800	70,300

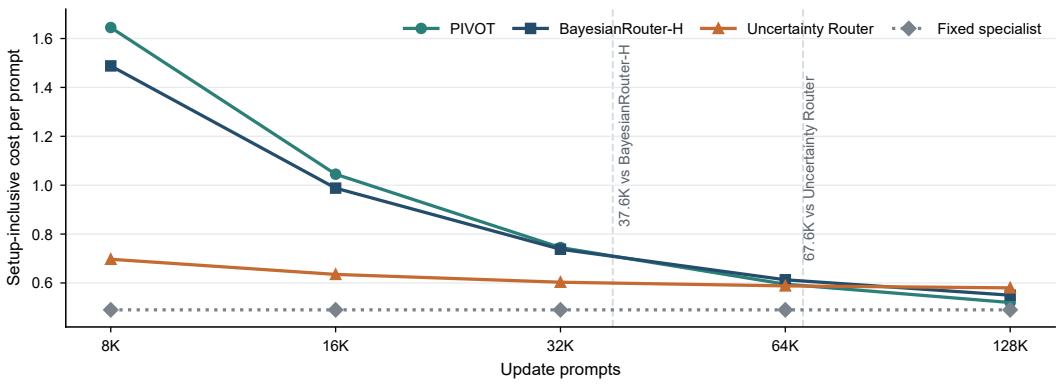


Figure S2: Setup amortization. Every curve includes its own fixed setup and online costs. PIVOT crosses BayesianRouter-H at 37.6K update prompts and Uncertainty Router at 67.6K; the fixed specialist retains lower cost at 128K and lower aggregate utility. Exact values and rate-card sensitivity appear in Table S37.

Table S40: Sensitivity to consequence weights. "High-consequence GG-risk" is gold-graph inconsistency in the prespecified high-consequence stratum. Permuting severity labels tests whether the reduction depends on semantically correct weights rather than generic regularization.

Setting	Utility	Overall GG-risk	High-consequence GG-risk	Coverage	Cost
High weight $\times 0.5$	68.6	0.050	0.064	0.85	0.50
Pre-specified weights	68.4	0.047	0.031	0.82	0.52
High weight $\times 1.5$	68.1	0.044	0.025	0.78	0.55
Uniform weights	68.2	0.049	0.052	0.84	0.50
Permuted severity	67.2	0.058	0.067	0.87	0.49

Table S41: Primary failure taxonomy. Each routed error receives one pre-specified primary category. This diagnostic is excluded from the certificate and primary utility. Counts are reported by domain together with row and column totals.

Category	Math	Code	Instruction	Science	Total
Graph-adapter error	5	5	3	2	15
Verifier-label error	6	7	15	18	46
Incomplete gold graph	1	2	8	11	22
Shared-correlated verifier error	5	6	11	15	37
Audit-objective omission	2	2	6	8	18
Projection/sign error	2	2	2	2	8
Minibatch cancellation	4	3	2	2	11
Policy-shift miscalibration	4	5	5	6	20
Cost-driven under-escalation	3	4	3	5	15
Abstention failure	4	2	1	1	8
Total	36	38	56	70	200

I STATISTICAL ANALYSIS

The independent training unit is one seed for one policy model. We use five seeds for Llama-3.1-8B-Instruct and five for Qwen2.5-7B-Instruct, retaining every seed in the reported comparisons. The primary estimand is the paired difference in final trusted average utility between PIVOT and BayesianRouter-H under the common heterogeneous verifier pool and the same total-cost cap. The source-faithful native RM-pool comparison between PIVOT–RM and BayesianRouter is reported

separately. Two pre-specified secondary estimands are the worst policy-model–domain slice and held-out GG-risk at $\alpha = 0.05$; secondary intervals use Holm adjustment.

For utility, a hierarchical bootstrap resamples seeds first and prompts within domain second, preserving pairing across methods. Per-seed and per-model values accompany the aggregate. For the mechanism study, the unit is a domain–checkpoint–verifier block; Spearman correlations are compared through checkpoint-blocked resampling. We report the held-out GG-risk numerator, denominator, weighted-coverage lower bound, unweighted coverage, and an interval separate from the certificate bound. Multiple GG-risk budgets, ablations, and subgroup analyses are labeled secondary or exploratory. No seed is excluded from the reported comparisons.

J Failure Analysis and Safeguards

This diagnostic does not contribute certificate labels or primary utility. We report the routed-error sample size separately for Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, together with the sampling fraction, blinded labels per item, and adjudicator count. Sampling is stratified by consequence weight, and annotators assign one primary category: graph-adapter error, verifier label error, incomplete gold graph, shared-correlated verifier error, audit-objective omission, projection/sign error, minibatch-cancellation error, policy-shift miscalibration, cost-driven under-escalation, or abstention failure. Unresolved labels follow a pre-specified adjudication rule, and all adjudicated cases are included in the analysis.

The analysis reports category frequencies by domain, policy model, verifier, consequence stratum, and policy phase, together with representative adjudicated cases. Human labels are reported only with the applicable review or exemption status, recruitment and qualification criteria, compensation, consent, privacy, and data-governance procedures. The certificate is interpreted solely for the declared strict graph-consistency event; omissions in the gold graph, consequence weights, or audit objective remain outside its scope.

References

Sher Badshah, Ali Emami, and Hassan Sajjad. Scope: Selective conformal optimized pairwise llm judging, 2026. URL https://arxiv.org/abs/2602.13110. Accepted at the 43rd International Conference on Machine Learning (ICML 2026).

Tiejin Chen, Xiaoou Liu, Vishnu Nandam, Kuan-Ru Liou, and Hua Wei. Conformal feedback alignment: Quantifying answer-level reliability for robust llm alignment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3561–3572, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.184. URL https://aclanthology.org/2026.findings-eacl.184/.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, pp. 50905–50931, 2024.

Shaddin Dughmi, Mahdi Haghifam, and Yusuf Hakan Kalayci. Adaptive generate-rank-verify: Inference-time search with costly verification, 2026. URL https://arxiv.org/abs/2605.17609.

Google. Gemma-2-27B-IT model card. Hugging Face model card, 2024. URL https://huggingface.co/google/gemma-2-27b-it.

Kenan Hasanaliyev, Silas Alberti, Jenny Hamer, Dheeraj Rajagopal, Kevin Robinson, Jasper Snoek, Victor Veitch, and Alexander D'Amour. Expected reward prediction, with applications to model routing. In *2nd Workshop on Models of Human Feedback for AI Alignment at ICML 2025*, 2025. URL https://openreview.net/forum?id=87JyHeA8f6.

Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. Advancing process verification for large language models via tree-based preference learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2086–2099, 2024. doi: 10.18653/v1/2024.emnlp-main.125.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, pp. 58791–58831, 2025.

Hamed Khosravi and Xiaoming Huo. Conformal selective acting: Anytime-valid risk control for rlvr-trained llms, 2026. URL https://arxiv.org/abs/2605.20270.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. In *International Conference on Learning Representations*, pp. 16484–16517, 2025.

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 *International Conference on Learning Representations*, pp. 39578–39601, 2024.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, pp. 144839–144866, 2026.

Meta Llama. Llama-3.1-8B-Instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Ryo Mitsuhashi, Patrick Chen, Isabelle Tseng, Jasin Cekinmez, and Addison J. Wu. Quantifying empirical compute-supervision tradeoffs in rlvr. *arXiv preprint arXiv:2605.25252*, 2026. doi: 10.48550/arXiv.2605.25252. URL https://arxiv.org/abs/2605.25252. Accepted to the 1st Workshop on Combining Theory and Benchmarks (CTB@ICML 2026).

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70305–70337. Curran Associates, Inc., 2025. doi: 10.52202/085713-2364.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. In *International Conference on Learning Representations*, pp. 156944–156972, 2026.

Qwen Team. Qwen2.5-72B-Instruct model card. Hugging Face model card, 2024a. URL https://huggingface.co/Qwen/Qwen2.5-72B-Instruct.

Qwen Team. Qwen2.5-7B-Instruct model card. Hugging Face model card, 2024b. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Arushi Rai, Qiang Zhang, Hanqing Zeng, Yunkai Zhang, Dipesh Tamboli, Xiangjun Fan, Zhuokai Zhao, and Lizhu Zhang. Taro: Token-level adaptive routing for llm test-time alignment. In *Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1004–1025, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.50.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Tom Saliencro, Rohan Desai, Priya Nair, Maya Lindqvist, and Daniel Whitmore. Uncertainty is not enough: Value-of-information routing for mixtures of lora experts. *arXiv preprint arXiv:2608.02528*, 2026. doi: 10.48550/arXiv.2608.02528. URL https://arxiv.org/abs/2608.02528.

Skywork Team. Skywork-Reward-Llama-3.1-8B-v0.2 model card. Hugging Face model card, 2024. URL https://huggingface.co/Skywork/Skywork-Reward-Llama-3.1-8B-v0.2.

Iqtedar Uddin and André Bauer. Conformal llm routing with distribution-free safety guarantees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 791–799, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-srw.70. URL https://aclanthology.org/2026.acl-srw.70/.

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*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510.

Yikai Wang, Shang Liu, and Jose Blanchet. Wasserstein distributionally robust regret optimization for reinforcement learning from human feedback, 2026. URL https://arxiv.org/abs/2605.00155.

Xinle Wu and Yao Lu. Reward model routing in alignment. In *International Conference on Learning Representations*, pp. 88128–88144, 2026.

Yingwen Wu, Tao Li, Xinwen Cheng, Jie Yang, and Xiaolin Huang. Low-dimensional gradient helps out-of-distribution detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11378–11391, 2024. doi: 10.1109/TPAMI.2024.3459988.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025a. doi: 10.48550/arXiv.2512.12844. URL https://arxiv.org/abs/2512.12844.

Zhenghao Xu, Qin Lu, Qingru Zhang, Liang Qiu, Ilgee Hong, Changlong Yu, Wenlin Yao, Yao Liu, Homing Jiang, Lihong Li, Hyokun Yun, and Tuo Zhao. Ask a strong llm judge when your reward model is uncertain. In *Advances in Neural Information Processing Systems*, volume 38, pp. 74639–74664, 2025b. doi: 10.52202/085713-2506.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*, 2024. doi: 10.48550/arXiv.2412.06559.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Bin Zhu, Yi Xie, and Yanghui Rao. Stopping and routing llm judge panels, 2026. URL https://arxiv.org/abs/2608.19802. Accepted at WISE 2026.