WHEN TO STAY, WHEN TO SWITCH: OUTCOME-BASED EVALUATION OF ADAPTIVE LLM INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive LLM inference chooses a model or reasoning budget for each request. Evaluating these choices after an input change requires knowing whether the chosen computation still produces a good answer at an appropriate cost. A route can stay unchanged while its outcome becomes inadequate; a switch can be necessary for a good outcome. We introduce PO-CTRLBENCH, a paired evaluation framework that measures every available action's quality and cost through repeated execution. These outcome profiles distinguish three requirements: keep an adequate action, allow a change among adequate actions, or adapt when no action remains adequate at both inputs. We show why routes alone cannot identify these requirements, separate outcome failures from selection inconsistency, and give finite-sample certification rules. Our controller, PORC, learns the profiles and selects actions using calibrated conservative utility scores. On two-action reasoning tasks, route consistency and endpoint risk rank five methods in exactly opposite orders. Compared with Outcome-greedy, PORC reduces the primary endpoint-risk sum from 12.1% to 2.2% and improves utility by 0.064 (95% source-cluster confidence interval $[0.049, 0.079]$). Under source shift, changing the decision scores on the same predictor reduces full paired risk from 23.89% to 12.50% and raises utility by 0.040. Paired outcome evaluation thus makes the distinction between stable computation and appropriate computation measurable.

1 INTRODUCTION

An LLM system can answer a request with a small model, a stronger model, or a longer reasoning process. Extra computation can improve an answer, but its value varies across requests. Model routing and test-time compute allocation address this quality–cost tradeoff by choosing how each request is processed (Chen et al., 2024; Ong et al., 2025; Snell et al., 2025; Zhang et al., 2025). Once a controller makes these choices, a further question arises: *when the input changes, does its choice remain appropriate?*

Consider a controller that switches from SHORT to THINK after a prompt edit. If SHORT still gives the better quality–cost tradeoff, the switch lowers utility. If the edit makes THINK necessary, the same switch is the correct response. Likewise, an unchanged route may continue to work well or may keep choosing an action that has become inadequate. A route records *what was selected*; judging that selection requires the outcomes of the available alternatives. This distinction matters for both routine edits, such as formatting or added information, and deliberate attempts to redirect a router (Kassem et al., 2026; Shafran et al., 2025).

We address this problem with PO-CTRLBENCH. For an original prompt and its edited version, we repeatedly execute every registered action and record answer quality and cost. These measurements form an *outcome profile*. An action is acceptable when its expected utility is close enough to the best available action. Comparing the two profiles and their acceptable-action sets then answers three concrete questions: should the controller retain an adequate action, may it change between adequate actions, or must it adapt because the acceptable sets no longer overlap?

The framework separates two evaluation goals. *Outcome adequacy* checks whether the selected action is acceptable at each input. *Selection consistency* checks whether the choices satisfy the paired

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.

Same switch: SHORT → THINK

Utilities (S, T)	Original	Edited
A: keep SHORT	$(1, 0)$	$(1, 0)$
B: switch to THINK	$(1, 0)$	$(0, 1)$

Best → worst in the primary study

Route consistency	Endpoint risk
Native	PORC
Fixed threshold	Outcome-greedy
Cost-minimizing	Cost-minimizing
Outcome-greedy	Fixed threshold
PORC	Native

Figure 1: Identical route changes can require opposite judgments. Left: illustrative outcome worlds; larger utility is better. Right: route consistency and primary endpoint risk reverse the order of the same five methods. Table 3 gives the risk values.

stability requirement. For equivalent profiles, a switch between two acceptable actions can violate consistency without losing useful outcome quality. We make that contribution explicit in the loss. This yields an exact decomposition and a bound that recovers information about endpoint failures even when an evaluation reports only full paired risk and route-change rate.

We also introduce PORC, a potential-outcome robust controller that uses the same representation for decisions. It predicts each action's quality and cost, adjusts utility scores using residuals from held-out calibration sources, and executes the chosen action. Full action execution is an offline measurement step; an inference request needs one profile prediction and one selected action. Comparing point and conservative scores on the same predictor makes the effect of this decision rule directly observable.

The paper makes three contributions. First, we define paired stability and adaptation requirements from all-action outcomes and show why route observations alone cannot identify them without assumptions linking text to outcomes. Second, we separate endpoint inadequacy from selection consistency and give finite-sample rules for certifying relations and action adequacy. Third, we demonstrate the practical value of this evaluation: the primary study reverses five methods' route-based ranking, and PORC's calibrated scores improve both full paired risk and utility on a fixed predictor under source shift. The main text presents the definitions, decision rule, and evidence; the appendix supplies proofs and execution details.

2 What should remain stable after an input change?

2.1 Measure the outcomes of each available action

Let $Z = (X, X')$ be an original–edited prompt pair and $\mathcal{A} = \{1, \ldots, K\}$ the registered actions. An action specifies a model, reasoning mode, or compute choice together with its decoder and forcing rule. Executing action a on x produces quality $Q_a(x) \in [0, 1]$ and normalized cost $C_a(x) \in [0, 1]$. The outcome profile collects their means:

$$P(x) = \{(\mathbb{E}Q_a(x), \mathbb{E}C_a(x)) : a \in \mathcal{A}\}, \qquad U_a(x) = \mathbb{E}Q_a(x) - \lambda \mathbb{E}C_a(x). \tag{1}$$

Here $\lambda \geq 0$ sets the cost preference, and expectations average generation randomness under the fixed execution mechanism. These are potential outcomes of action interventions (Rubin, 1974; Imbens & Rubin, 2015): forcing each action lets us measure its response even when the native controller would select another action.

Define action regret and the acceptable-action set by

$$\mathrm{Reg}_a(x) = \max_{b \in \mathcal{A}} U_b(x) - U_a(x), \qquad \mathcal{A}_\gamma(x) = \{a : \mathrm{Reg}_a(x) \leq \gamma\}. \tag{2}$$

The threshold $\gamma \geq 0$ allows near-best actions. For example, utilities $(0.80, 0.78)$ make both actions acceptable at $\gamma = 0.03$, while $(0.80, 0.70)$ makes only the first acceptable. Keeping a set of acceptable actions avoids treating an immaterial utility difference as a mandatory route change.

Table 1: Three outcome relations and their paired requirements. Every requirement includes acceptable choices at both inputs. Unresolved (U) means that finite data do not certify a relation.

Relation	Outcome condition	Required behavior
Equivalent (E)	Acceptable sets overlap; $d_P \leq 1$	Keep the same acceptable action.
Preserving (P)	Acceptable sets overlap; $d_P > 1$	Stay acceptable; changing action is allowed.
Critical (C)	Acceptable sets are disjoint	Change action and remain acceptable.

2.2 Separate stable profiles from necessary adaptation

We compare quality and cost separately, using tolerances $\epsilon_Q, \epsilon_C > 0$:

$$d_P(x, x') = \max_{a \in \mathcal{A}} \max \left\{ \frac{|\mathbb{E}Q_a(x) - \mathbb{E}Q_a(x')|}{\epsilon_Q}, \frac{|\mathbb{E}C_a(x) - \mathbb{E}C_a(x')|}{\epsilon_C} \right\}. \tag{3}$$

The profiles agree within tolerance when $d_P \leq 1$. Table 1 combines this comparison with acceptable-set overlap. We check disjointness first: even a small profile change can require adaptation when actions are close to the acceptance boundary. Transformation names describe how prompts were made; the measured profiles determine their relation.

For an equivalent pair, overlap makes a stable, acceptable choice possible. Requiring the same choice is a selection-consistency criterion in addition to outcome adequacy. For a preserving pair, the profile has changed enough to allow a different choice while a common acceptable action remains available. For a critical pair, no single action can be acceptable at both endpoints, so an adequate controller must adapt.

Theorem 1 (Routes alone do not identify the requirement)**.** *Let $K \geq 2$ and $0 \leq \gamma < 1$. Without assumptions linking text to action outcomes, the prompts, their declared semantic relation, and the selected routes cannot determine the correct stability or adaptation requirement for every compatible outcome profile.*

To see why, fix an observed SHORT→THINK switch. Utilities $(1, 0)$ at both inputs require SHORT at both. Utilities $(1, 0)$ followed by $(0, 1)$ require a switch. Both worlds have the same route observations and opposite requirements (Figure 1). The missing information is each action's outcome. A useful sufficient condition goes in the other direction: if every action's utility changes by at most η and the original winner leads by more than 2η, it remains the unique winner. The factor 2 accounts for the winner losing η while a competitor gains η.

2.3 Measure endpoint failures and consistency separately

For a policy π, let b_i^0 and b_i^1 indicate regret above γ at the original and edited inputs. Write $B_i = b_i^0 \vee b_i^1$ for a failure at either endpoint and $F_i = \mathbf{1}\{\pi(x_i) \neq \pi(x_i')\}$ for a route change. Logical or is denoted by $\vee$. The full paired losses are

$$\ell_E = B_i \vee F_i, \qquad \ell_P = B_i, \qquad \ell_C = B_i \vee (1 - F_i). \tag{4}$$

These losses implement Table 1. In particular, ℓ_P allows a route change, while ℓ_E records a consistency failure even if both chosen actions are acceptable.

A worked example. Table 2 shows how the same SHORT→THINK decision is evaluated in all three relations. For this illustration, costs are zero, so the listed utilities are also mean qualities. We use $\gamma = 0.05$, $\epsilon_Q = 0.15$, and $\epsilon_C = 0.05$. In the equivalent row, both actions are acceptable before and after the edit. Switching keeps regret within the threshold but violates the requirement to make a consistent selection. In the preserving row, THINK becomes the only acceptable edited-input action, while both actions were acceptable originally. The switch satisfies the full requirement, and keeping THINK at both inputs would also have been acceptable. In the critical row, the acceptable action changes from SHORT alone to THINK alone, so a switch is necessary.

The example clarifies why the outcome relation and the realized policy behavior are separate objects. A preserving label says that a common acceptable action exists; it does not say that the controller

Table 2: Illustrative outcomes for one observed SHORT→THINK switch. Costs are zero; the two entries in each utility vector correspond to SHORT and THINK. B records endpoint inadequacy, and ℓ_L includes the paired selection requirement. These are constructed examples.

Relation	$U(x)$	$U(x')$	Shared acceptable actions	B	ℓ_L
E	$(0.80, 0.78)$	$(0.81, 0.79)$	SHORT, THINK	0	1
P	$(0.80, 0.78)$	$(0.60, 0.90)$	THINK	0	0
C	$(0.80, 0.40)$	$(0.40, 0.80)$	None	0	0

selected that action originally. A critical label says that every adequate pair of choices must differ. An equivalent label makes stable adequacy feasible, while the consistency part of the loss measures whether the controller actually chooses it. Reporting B alongside the full loss keeps these interpretations visible.

Proposition 1 (What full paired risk counts). *Set $s_i(E) = F_i$, $s_i(P) = 0$, and $s_i(C) = 1 - F_i$. For a full label L_i,*

$$\ell_{L_i} = b_i^1 + (1 - b_i^1)b_i^0 + (1 - b_i^1)(1 - b_i^0)s_i(L_i). \tag{5}$$

If labels and adequacy use the same profile, λ, and γ, the selection-only term can occur only on equivalent pairs. Consequently,

$$R_{\text{full}} = R_B + S_E, \qquad \max\{0, R_{\text{full}} - f\} \le R_B \le R_{\text{full}}, \tag{6}$$

where R_B averages B_i, f averages F_i, and S_E averages $(1 - B_i)\mathbf{1}\{L_i = E\}F_i$.

The three terms in Equation 5 count, without overlap, edited-input failure, additional original-input failure, and consistency failure when both endpoints are adequate. On a critical pair, adequate choices already belong to disjoint sets and therefore differ. This removes a redundant route penalty on C and yields Equation 6.

For empirical evaluation, bars denote rollout means and hats denote computed labels or risks. We replace true utilities by $\overline{U}_a = \overline{Q}_a - \lambda \overline{C}_a$, then compute both labels and losses from that same profile. The empirical full risk is $\widehat{R}_{\text{full}} = N^{-1} \sum_i \overline{\ell}_{\widehat{L}_i}$. Equation 6 applies to these empirical quantities as well. Its bounds isolate endpoint outcomes from consistency; they are algebraic bounds on the evaluated records.

3 How much evidence is needed for a pair?

Repeated execution gives an estimate of a profile. We use simultaneous intervals to decide when those measurements support a relation. For each action, a pair supplies four endpoint means (quality and cost at each input) and two signed differences, for $6K$ intervals in total. We compute differences within each paired seed; endpoints may be dependent within a seed, while seeds are independent. Quality and cost lie in $[0, 1]$, and signed differences lie in $[-1, 1]$.

Let $[Q_a^-, Q_a^+]$ and $[C_a^-, C_a^+]$ be endpoint intervals. They induce $U_a^- = Q_a^- - \lambda C_a^+$ and $U_a^+ = Q_a^+ - \lambda C_a^-$. Define inner and outer acceptable sets:

$$\begin{aligned} \mathcal{A}_\gamma^-(z) &= \{a : \max_b U_b^+(z) - U_a^-(z) \le \gamma\}, \\ \mathcal{A}_\gamma^+(z) &= \{a : \max_b U_b^-(z) - U_a^+(z) \le \gamma\}. \end{aligned} \tag{7}$$

On simultaneous coverage, $\mathcal{A}_\gamma^- \subseteq \mathcal{A}_\gamma \subseteq \mathcal{A}_\gamma^+$. Thus overlapping inner sets certify a common acceptable action, and disjoint outer sets certify that no common acceptable action exists. Difference intervals wholly inside their tolerances certify profile equivalence; any difference interval wholly outside its tolerance certifies profile change. We issue C for certified disjointness, E for certified overlap and equivalence, P for certified overlap and change, and U otherwise.

Theorem 2 (Finite-sample relation and action certificates). *For a fixed pair, suppose the $6K$ intervals each have error probability at most $\alpha/(6K)$. With probability at least $1 - \alpha$, every relation issued by the above rule is correct. On the same event, every action's regret lies between*

$$g_a^- = \max\{0, \max_{b \ne a}(U_b^- - U_a^+)\}, \qquad g_a^+ = \max\{0, \max_{b \ne a}(U_b^+ - U_a^-)\}. \tag{8}$$

Therefore $g_a^+ \le \gamma$ certifies an acceptable action, and $g_a^- > \gamma$ certifies an inadequate action.

A union bound covers all interval inputs. Set containment proves the relation claims, and maximizing bounded utility differences proves the regret claims. Inserting the resulting binary failure bounds into Equation 4 also bounds the full pair loss. Allocating error across pairs gives a study-wide guarantee; allocating it across a finite, predeclared set of rollout depths permits stopping at the first certifying depth. Appendix A gives the full proof, a Bentkus bounded-mean reference inversion (Bentkus, 2004), and the loss bounds.

From an action certificate to a risk bound. At endpoint z_i^j, define $b_i^{j,-} = \mathbf{1}\{g^-_{\pi(z_i^j)} > \gamma\}$ and $b_i^{j,+} = \mathbf{1}\{g^+_{\pi(z_i^j)} > \gamma\}$. For a certified full relation L_i, set

$$\ell_i^- = b_i^{0,-} \vee b_i^{1,-} \vee s_i(L_i), \qquad \ell_i^+ = b_i^{0,+} \vee b_i^{1,+} \vee s_i(L_i). \tag{9}$$

The lower loss counts failures that the intervals establish; the upper loss also counts failures the intervals cannot rule out. On simultaneous coverage, $\ell_i^- \le \ell_{L_i} \le \ell_i^+$. For N fixed pairs, let $\mathcal{I}$ index certified full relations and $N_U = N - |\mathcal{I}|$. With joint coverage across pairs, the risk computed from true utilities satisfies

$$\frac{1}{N}\sum_{i\in\mathcal{I}} \ell_i^- \le R_{\text{full}} \le \frac{1}{N}\left(\sum_{i\in\mathcal{I}} \ell_i^+ + N_U\right). \tag{10}$$

Thus measurement uncertainty propagates from action outcomes to relations, endpoint adequacy, and full paired risk. Unresolved relations contribute zero to the lower bound and one to the upper bound; resolved relations retain the uncertainty in their selected actions.

The primary endpoint protocol. The primary experiment uses an $E_{\text{prof}}/C/U$ certificate rule. It first certifies C when different actions are uniquely acceptable at the two endpoints; otherwise it certifies E_{prof} when profiles agree within tolerance, and returns U in the remaining cases. Full E adds acceptable-set overlap. To preserve the meaning of the primary experiment, we report its two registered endpoint components:

$$\widehat{\text{PPFR}} = \frac{1}{N}\sum_i \mathbf{1}\{\widehat{L}_i = E_{\text{prof}}\}\bar{b}_i^1, \qquad \widehat{\text{CRMR}} = \frac{1}{N}\sum_i \mathbf{1}\{\widehat{L}_i = C\}\bar{b}_i^1. \tag{11}$$

Here $\bar{b}_i^1$ uses evaluation regret threshold γ_{eval}; primary labels use γ_{cert}. Both rates use *all* N pairs as denominator. Their sum measures edited-input failures on resolved primary pairs; source-shift and three-action experiments evaluate the full losses in Equation 4.

If N_U primary records are unresolved, their unspecified binary failures give the completion range

$$[R_L, R_L + N_U/N], \qquad R_L = \widehat{\text{PPFR}} + \widehat{\text{CRMR}}. \tag{12}$$

The endpoints assign zero or one failure to every unresolved record (Manski, 2003). This range holds with the observed labels and failures fixed. Source-cluster confidence intervals instead quantify variation across sampled sources, while simultaneous outcome intervals quantify measurement uncertainty.

4 PORC: CHOOSE COMPUTATION FROM PREDICTED OUTCOMES

Learn the profile. PORC predicts $\widehat{P}_\theta(x)$ from the input and checkpoint identifier. Forced-execution means supervise every action's quality and cost. The two-action predictor combines word and character TF–IDF features, a construction-fitted SVD representation, and a shared 128-unit GELU hidden layer with four outputs $(\widehat{Q}_S, \widehat{C}_S, \widehat{Q}_T, \widehat{C}_T)$. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{profile}} + 0.5\mathcal{L}_{\text{eq}} + 0.5\mathcal{L}_{\text{crit}}. \tag{13}$$

The first term fits quality and cost by mean squared error. The pair terms penalize prediction drift on equivalent training pairs and enforce the target action margins on critical training pairs; the critical hinge margin is 0.03. Training relations come from construction profiles. Appendix B specifies feature fitting and optimization.

The data flow follows the information available when a decision is made. Construction sources provide all-action responses for learning. Tuning sources select the fitted predictor, and separate calibration sources determine its score offsets. Both the predictor and the decision rule are then fixed. On a held-out prompt, the policy records its choice before fresh forced executions measure the outcomes of all available actions. Those executions supply the evaluation profile, its relation to the paired prompt, and the loss of the recorded choice. Keeping all versions of a source in one split prevents an edited question from serving as training evidence for its own evaluation.

This protocol also gives different controllers a common outcome reference. Once the held-out all-action profile is measured, each policy's recorded choice can be evaluated against the same alternatives. A comparison can therefore attribute a changed loss to different choices while holding the measured action consequences fixed. The all-action execution budget is paid during offline construction and evaluation; it is distinct from the generated-token cost of the selected inference action reported in the result tables.

Calibrate at the source level. Each calibration source contributes the maximum absolute prediction residual over its original and edited prompts, separately for each profile coordinate. For n_{cal} sources, the radius is the residual at rank $\lceil (n_{\mathrm{cal}}+1)(1-\beta) \rceil$. We use $n_{\mathrm{cal}} = 15$ and $\beta = 0.10$, so each radius is the largest of 15 source residuals. Under exchangeable calibration and test sources, this construction covers a new source's coordinate residual with probability at least $1-\beta$ (Vovk et al., 2005; Lei et al., 2018). The calibration target is the rollout mean. The decision uses separate radii r_a^Q, r_a^C for quality and cost.

Choose by conservative utility. PORC selects

$$\pi_{\mathrm{LCB}}(x) \in \arg\max_a \{\widehat{U}_a(x) - e_a\}, \qquad \widehat{U}_a = \widehat{Q}_a - \lambda \widehat{C}_a, \quad e_a = r_a^Q + \lambda r_a^C. \tag{14}$$

This rule lowers predicted quality and raises predicted cost before comparing actions. PORC-point instead maximizes $\widehat{U}_a$ using the *same predictor*. Both break two-action ties in favor of SHORT. At deployment, a request passes through the predictor and executes the selected action; evaluation profiles are never inputs to this decision.

For two actions, the rule has a simple interpretation. Set $d(x) = \widehat{U}_T(x) - \widehat{U}_S(x)$ and $\tau = e_T - e_S$. PORC-point chooses THINK when $d(x) > 0$, while PORC-LCB chooses THINK when $d(x) > \tau$. A positive τ changes THINK to SHORT on $0 < d(x) \le \tau$; a negative τ changes SHORT to THINK on $\tau < d(x) \le 0$. All other choices stay the same. The radii are action-specific offsets within each calibration group. Equal offsets leave every decision unchanged.

Proposition 2 (Regret under a joint prediction-error bound)**.** *If $|\widehat{U}_a(x) - U_a(x)| \le e_a$ for every action, the action $\widehat{a}$ chosen by Equation 14 satisfies*

$$\mathrm{Reg}_{\widehat{a}}(x) \le 2e_{a^*}, \qquad a^* \in \arg\max_a U_a(x). \tag{15}$$

The conservative score of the selected action is at least that of a^*, whose true utility is at most $2e_{a^*}$ above its score. This proves the claim. Its condition concerns simultaneous errors against true utilities. With rollout-mean measurement bounds ν_a, the corresponding result is $\mathrm{Reg}_{\widehat{a}} \le 2e_{a^*} + \nu_{a^*} + \nu_{\widehat{a}}$ (Appendix A.4). Under source shift, we assess the fixed decision rule empirically.

5 Experiments

We evaluate three questions: does the outcome-based criterion change the ranking of controllers, does profile-based decision making improve their choices, and do the three paired requirements remain useful when actions select different models?

5.1 Setup and comparison design

Two-action reasoning. The primary study uses AdaptThink-1.5B checkpoints with penalty parameters $\rho \in \{0.05, 0.075, 0.10\}$ (Zhang et al., 2025). Actions are SHORT and THINK. SHORT inserts `</think>` immediately; THINK suppresses that token and EOS at the first decoding step.

Table 3: Primary endpoint risks and quality–cost outcomes at $\lambda = 0.25$. PPFR and CRMR use all 720 records as denominator. The five routing methods appear from highest to lowest route consistency; the constant policies show fixed-action tradeoffs.

Method	PPFR↓	CRMR↓	$Q \uparrow$	$C \downarrow$	$U \uparrow$
Native	9.2%	14.0%	0.766	0.656	0.602
Fixed threshold	6.7%	10.2%	0.753	0.580	0.608
Cost-minimizing	5.8%	9.3%	0.740	0.504	0.614
Outcome-greedy	4.5%	7.6%	0.776	0.620	0.621
PORC	0.7%	1.5%	0.812	0.508	0.685
Always SHORT	11.8%	16.5%	0.722	0.118	0.693
Always THINK	0.8%	1.9%	0.798	0.975	0.554

GSM8K and MATH-500 (Cobbe et al., 2021; Hendrycks et al., 2021) each provide 60 construction, 30 validation, and 30 held-out source questions. Half the validation sources tune the predictor and half calibrate it. Construction, tuning, calibration, and evaluation are source-disjoint.

The six edit families are candidate verification, helpful interval, modulo-7 edit, Markdown wrapper, polite wrapper, and whitespace delimiter. They supply candidate relations; outcomes determine the issued labels. Quality is exact answer match after gold-independent extraction. Cost is generated tokens, including the forced prefix, capped at and divided by 8192. Decoding uses temperature 0.6, top-p 0.95, one beam, and a 16384-token context cap.

We use four construction rollouts per action and 128 fresh certification rollouts per action and prompt. The 240 pairs per checkpoint give 720 pair–checkpoint records and 368640 forced executions. The primary protocol fixes $\lambda = 0.25$, $(\epsilon_Q, \epsilon_C) = (0.15, 0.05)$, $\gamma_{\text{cert}} = 0.03$, $\gamma_{\text{eval}} = 0.05$, and pair error budget $\alpha_{\text{pair}} = 0.05$. It issues 396 E_{prof}, 264 C, and 60 U labels, resolving 91.67% of records. The empirical risk gates are PPFR$\leq 5\%$ and CRMR$\leq 8\%$.

What each comparison changes. Native uses the checkpoint's own route. Fixed threshold applies a fixed cutoff to its route score. Cost-minimizing learns from forced costs. Outcome-greedy learns forced quality and cost and selects the largest predicted utility. PORC adds pair supervision and calibrated conservative scores. The primary PORC–Outcome-greedy comparison therefore evaluates complete training and decision pipelines. PORC-point versus PORC-LCB fixes the profile predictor and tie rule, changing the action-specific score offsets in Equation 14. Always SHORT and Always THINK provide transparent fixed-action references.

Statistical units. All edits, checkpoints, and executions from one source form one cluster. Primary confidence intervals use 10000 paired bootstrap resamples of the 60 held-out source clusters (Efron, 1979), preserving all 720 nested records and holding outcome estimates and labels fixed. All three training seeds are retained. The original Holm family contains two primary risks and utility for PORC against four baselines, totaling 12 contrasts (Holm, 1979). These intervals describe source variation in the empirical evaluation.

5.2 Outcome evaluation reverses the route-based ranking

Descending route consistency ranks Native, Fixed threshold, Cost-minimizing, Outcome-greedy, and PORC. Ranking the same five methods by lower primary endpoint risk gives exactly the reverse order (Figure 1, Table 3); adjacent source-cluster contrasts have Holm-adjusted $p < 0.01$. A controller can therefore look more stable while making less adequate choices after an edit. The outcome criterion makes this disagreement visible.

PORC lowers PPFR from 4.5% to 0.7% and CRMR from 7.6% to 1.5% relative to Outcome-greedy. The paired differences are -3.8 percentage points (95% source-cluster interval $[-5.1, -2.5]$) and -6.1 points ($[-7.9, -4.4]$). Utility rises by 0.064 ($[0.049, 0.079]$). All 12 primary baseline contrasts have Holm-adjusted $p < 0.001$. Both datasets show lower component risks: PORC's PPFR/CRMR are 0.5%/1.2% on GSM8K and 0.9%/1.8% on MATH-500.

Table 4: Full paired risk and endpoint-only bounds under source shift. Risk, route changes, and bounds are percentages. Endpoint bounds follow Equation 6 on the same records; utility uses $\lambda = 0.25$. PORC variants share predictions and tie-breaking.

Method	Full risk	Route changes	Endpoint bounds	Utility
Native	43.06	28.61	$[14.45, 43.06]$	0.570
PORC-point	23.89	15.28	$[8.61, 23.89]$	0.665
PORC-LCB	12.50	7.50	$[5.00, 12.50]$	0.705

Unresolved pairs preserve the comparison. The primary risk sum is 2.2% for PORC and 12.1% for Outcome-greedy. With the observed labels and failures fixed, the $60/720$ unresolved mass gives completion ranges $[2.2\%, 10.5\%]$ and $[12.1\%, 20.4\%]$, respectively. Even assigning failure to every unresolved PORC record and success to every unresolved Outcome-greedy record preserves the ordering. This directly accounts for unresolved cases in the reported endpoint comparison.

Quality and cost explain the gain. PORC improves mean quality by 0.036 and reduces normalized token cost by 0.112 relative to Outcome-greedy. At $\lambda = 0.25$, their utility contributions are 0.036 and 0.028, totaling 0.064. Holding the selected actions fixed, the difference at another cost weight is $0.036 + 0.112\widetilde{\lambda}$, positive for every $\widetilde{\lambda} \geq 0$.

The constant policies explain why utility and paired risks are reported together. Always SHORT has utility 0.693 and exceeds both risk gates; Always THINK meets the gates with cost 0.975 and utility 0.554. PORC meets the same gates with cost 0.508 and utility 0.685. The comparison identifies a useful allocation: low endpoint risk with substantially less generation than always thinking.

5.3 The same predictor improves through its decision rule

The post-hoc source-shift study evaluates 240 sources and 720 nested pair–checkpoint records. It uses full point labels and losses from the same rollout profiles at $\lambda = 0.25$ and a common adequacy threshold. Both endpoints and the relation-specific selection requirement contribute to $\widehat{R}_{\text{full}}$.

Changing point scores to conservative scores lowers full risk from 23.89% to 12.50%, a reduction of 11.39 percentage points, and raises utility from 0.665 to 0.705 (Table 4). Since the predictor, profiles, and tie rule are shared, this comparison isolates the empirical effect of the score adjustment. The threshold interpretation in Section 4 identifies exactly which inputs can change actions.

The corresponding marginal 95% source-cluster risk intervals are $[20.42\%, 27.36\%]$ for PORC-point and $[9.86\%, 15.42\%]$ for PORC-LCB. Combining them gives a difference interval of $[-17.50, -5.00]$ percentage points with at least 90% joint coverage under the original interval assumptions. The comparison also remains negative if at most 5% of the shared evaluation records are replaced while policies, rules, and weights remain fixed (Appendix A.5).

Here the utility gain comes from spending more computation where it improves quality. Quality increases from 0.740 to 0.812, and cost from 0.300 to 0.428. The quality gain 0.072 exceeds the added cost penalty 0.032, giving 0.040. For these fixed actions, repricing gives $0.072 - 0.128\widetilde{\lambda}$, which remains positive for $\widetilde{\lambda} < 0.5625$. Thus an effective adjustment can reduce or increase computation depending on its outcome value.

Endpoint gains persist after removing consistency. Full risk includes the selection term on equivalent pairs. Equation 6 lets us assess outcome failures alone. PORC-LCB's endpoint-risk upper bound is 12.50%, below Native's lower bound of 14.45%. At reported precision, its empirical two-endpoint risk is therefore at least 1.95 percentage points lower, independently of the consistency contribution. The PORC-point range overlaps the PORC-LCB range; the fixed-predictor result establishes lower *full* risk and higher utility.

Table 5: Three-action evaluation at $\lambda = 0.25$. The reported $E/P/C$ contributions use full point-label losses and the full evaluation denominator. All actions have quality and token-length measurements.

Method	E risk	P risk	C risk	Q	C	U
Always strong	4.2%	27.5%	4.1%	0.328	0.332	0.2450
PORC-point	4.4%	30.1%	4.1%	0.330	0.305	0.2538
PORC-LCB	4.0%	25.2%	3.2%	0.342	0.312	0.2640

5.4 Three actions separate the three requirements

The external study selects Qwen2.5-0.5B, 1.5B, or 7B Instruct on GSM8K, MATH-500, AIME, and AMC questions. Construction, validation, and evaluation contain 48, 32, and 32 source questions. Evaluation forms 64 pairs from 96 prompt profiles, with 128 rollouts per action and prompt, totaling 36864 executions. Cost is generated length capped at and divided by 2048; model size is not part of this token-length measure.

On the same predictor, conservative scores reduce the reported E, P, and C risks by 0.40, 4.90, and 0.90 percentage points and raise utility by 0.0102 (Table 5). The preserving category matters here: it captures changed profiles for which the controller can remain adequate with either a shared action or an appropriate switch. Alongside the always-strong reference, these results show how the framework compares different allocations while retaining the quality and token costs of their choices.

6 Related work

Routing and adaptive reasoning. FrugalGPT, RouteLLM, model-choice predictors, and cascades select models to improve quality–cost tradeoffs (Chen et al., 2024; Ong et al., 2025; Šakota et al., 2024; Aggarwal et al., 2024; Song et al., 2025). Adaptive reasoning allocates thinking effort within a model (Snell et al., 2025; Fang et al., 2025; Zhang et al., 2025; Tu et al., 2025). RouterBench and RouterEval compare routing quality and cost across methods (Hu et al., 2024; Huang et al., 2025). Our evaluation adds a paired question to these outcome objects: which actions remain acceptable after an edit, and what does that imply for the relationship between the two choices?

Behavioral testing and router robustness. Behavioral tests use structured input changes to probe model behavior (Ribeiro et al., 2020). Router-robustness and rerouting studies examine sensitivity to inputs and adversarial changes (Kassem et al., 2026; Shafran et al., 2025; Zhang et al., 2026). PO-CTRLBench determines the evaluation relation through measured action outcomes. Its three relations distinguish profile stability, continued action adequacy, and necessary adaptation. The loss decomposition then states how route consistency contributes beyond endpoint quality–cost failures.

Uncertainty and action evaluation. Potential outcomes specify the response to each action (Rubin, 1974; Imbens & Rubin, 2015); selective prediction and conformal methods provide tools for abstention and calibrated inference (Geifman & El-Yaniv, 2019; Angelopoulos et al., 2024; Uddin & Bauer, 2026). We use these ideas for paired action requirements: simultaneous outcome intervals certify set overlap or separation, and source-level residuals adjust predicted utility. Completion ranges describe the unresolved portion of a fixed empirical evaluation (Manski, 2003).

7 Conclusion

PO-CTRLBench makes stability and adaptation testable through the quality and cost of every available action. Its paired relations specify when to retain an adequate choice, when a change is allowed, and when adaptation is necessary. The loss decomposition separates outcome failures from selection inconsistency, and simultaneous intervals provide finite-sample certificates. Experiments show why the evaluation target matters: route consistency reverses the primary outcome ranking, while PORC's calibrated scores improve full paired risk and utility on the same predictor under source shift. These results support evaluating adaptive computation through its consequences at both inputs and the requirements those consequences establish.

AI USE STATEMENT

We used generative AI tools to assist with grammar checking and correction, as well as checking LaTeX formatting. The authors are responsible for the final text, research claims, and related outputs.

ETHICS STATEMENT

The experiments use benchmark questions and fixed inference controllers. Quality and cost are defined for the registered action sets and tasks; applications should specify these quantities for their own use conditions.

REPRODUCIBILITY STATEMENT

Sections 2–5 specify the evaluation objects, decision rule, study sizes, and main results. Appendix A supplies the proofs and reference interval construction; Appendix B gives execution, training, calibration, and reporting details.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. AutoMix: Automatically mixing language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4164. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ecda225cb187b40ea8edc1f46b03ffda-Abstract-Conference.html.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Vidmantas Bentkus. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004. doi: 10.1214/009117904000000360.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=cSimKw5p6R.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*, 2021.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: LLM learns when to think. *arXiv preprint arXiv:2505.13379*, 2025.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024. URL https://arxiv.org/abs/2403.12031.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. RouterEval: A comprehensive benchmark for routing LLMs to explore model-level scaling up in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3860–3887, 2025. doi: 10.18653/v1/2025.findings-emnlp.208. URL https://aclanthology.org/2025.findings-emnlp.208/.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Aly M. Kassem, Bernhard Schölkopf, and Zhijing Jin. How robust are router-LLMs? analysis of the fragility of LLM routing capabilities. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7496–7507, 2026. doi: 10.18653/v1/2026.eacl-long.351. URL https://aclanthology.org/2026.eacl-long.351/.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, 2003. doi: 10.1007/b97478.

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 with preference data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Marija Šakota, Maxime Peyrard, and Robert West. Fly-swat or cannon? cost-effective language model choice via meta-modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 606–615, 2024. doi: 10.1145/3616855.3635825.

Avital Shafran, Roei Schuster, Thomas Ristenpart, and Vitaly Shmatikov. Rerouting LLM routers. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=U6C7odo5SX.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. IRT-Router: Effective and interpretable multi-LLM routing via item response theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15629–15644, 2025. doi: 10.18653/v1/2025.acl-long.761. URL https://aclanthology.org/2025.acl-long.761/.

Songjun Tu, Jiahao Lin, Qichao Zhang, Xiangyu Tian, Linjing Li, Xiangyuan Lan, and Dongbin Zhao. Learning when to think: Shaping adaptive reasoning in R1-style models via multi-stage RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0510. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/16371a9d5fed65d6d78ca3a7fa6e598c-Abstract-Conference.html.

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, 2026. doi: 10.18653/v1/2026.acl-srw.70. URL https://aclanthology.org/2026.acl-srw.70/.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. Adaptthink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730, 2025. doi: 10.18653/v1/2025.emnlp-main.184. URL https://aclanthology.org/2025.emnlp-main.184/.

Wenhui Zhang, Huiyu Xu, Zhibo Wang, Zhichao Li, Zeqing He, Xuelin Wei, and Kui Ren. Rerouteguard: Understanding and mitigating adversarial risks for LLM routing. *arXiv preprint arXiv:2601.21380*, 2026. URL https://arxiv.org/abs/2601.21380.

A PROOFS AND FINITE-SAMPLE CALCULATIONS

A.1 IDENTIFICATION AND PAIRED LOSS

Proof of Theorem 1. Fix two actions S, T, the prompt pair, its declared semantic relation, and an observed switch $S \to T$. In world A, both prompts have utilities $(U_S, U_T) = (1, 0)$. For $0 \le \gamma < 1$, S is the only acceptable action at both endpoints. In world B, original utilities are $(1, 0)$ and edited utilities are $(0, 1)$; the acceptable sets are disjoint and the switch is optimal. Zero costs and binary qualities realize both profiles. The observable prompts and routes coincide, so a rule based on these observations cannot give the correct requirement in both worlds. A randomized rule cannot be correct with probability one in both either. Assigning utility zero to additional actions extends the construction to any $K \ge 2$. □

A sufficient margin for stability. Suppose $|U_a(x) - U_a(x')| \le \eta$ for every action and the unique winner a^* at x leads each competitor by more than 2η. For $b \ne a^*$,

$$U_{a^*}(x') - U_b(x') \ge U_{a^*}(x) - U_b(x) - 2\eta > 0.$$

Thus a^* remains uniquely optimal. At a margin of 2η, decreasing the winner by η and increasing a competitor by η can create a tie, so the factor 2 is tight.

Proof of Proposition 1. The three terms in Equation 5 are mutually exclusive. They partition the logical or $b_i^1 \vee b_i^0 \vee s_i(L_i)$ by whether the edited endpoint fails, the original endpoint alone fails, or both endpoints are adequate and the selection requirement fails. This proves the identity.

For E, $\ell_E = B_i + (1 - B_i)F_i$. For P, $\ell_P = B_i$. For C, disjoint acceptable sets imply $F_i = 1$ whenever $B_i = 0$; if $B_i = 1$, loss is already one. Therefore

$$\ell_{L_i} = B_i + (1 - B_i)\mathbf{1}\{L_i = E\}F_i.$$

Averaging gives $R_{\text{full}} = R_B + S_E$. Since $0 \le S_E \le f$, Equation 6 follows. The same argument applies to empirical profiles when labels and adequacy share the profile, cost weight, and threshold. Common nonnegative normalized weights also preserve the result. □

For any two policies evaluated on the same records, Equation 6 yields

$$\widehat{R}_B(\pi_0) - \widehat{R}_B(\pi_1) \ge \max\{0, \widehat{R}_{\text{full}}(\pi_0) - \widehat{f}(\pi_0)\} - \widehat{R}_{\text{full}}(\pi_1). \tag{16}$$

A positive right side establishes an empirical endpoint-risk advantage using aggregate statistics.

A.2 REFERENCE INTERVALS FOR BOUNDED MEANS

The following inversion supplies intervals with the coverage required by Theorem 2. It uses the bounded-variable form of Bentkus (2004, Theorem 1.2). Let $W_1, \ldots, W_m$ be independent observations in $[0, 1]$ with common mean θ, and let $S_m = \sum_s W_s$. For integer $j \in \{0, \ldots, m\}$, define

$$T_m(j; \theta) = \sum_{k=j}^{m} \binom{m}{k} \theta^k (1 - \theta)^{m-k}. \tag{17}$$

At a noninteger $t \in (0, m)$, write $j = \lfloor t \rfloor$ and $w = t - j$, and interpolate in log space:

$$T_m^\circ(t; \theta) = T_m(j; \theta)^{1-w} T_m(j + 1; \theta)^w.$$

At integers use T_m; take $T_m^\circ = 1$ for $t \le 0$ and zero for $t > m$. For $t > 0$, the $\theta = 0$ boundary is zero; for $t \le m$, the $\theta = 1$ boundary is one. The tail bound is

$$\Pr_\theta(S_m \ge t) \le q_m(t; \theta) := \min\{1, \mathrm{e}\, T_m^\circ(t; \theta)\}. \tag{18}$$

Here e is Euler's number. For observed sum s and total two-sided error δ, set

$$L_m(s; \delta) = \inf\{\theta : q_m(s; \theta) \ge \delta/2\}, \qquad U_m(s; \delta) = 1 - L_m(m - s; \delta). \tag{19}$$

This allocates $\delta/2$ to each tail. At $s = 0$, $L_m = 0$; at $s = m$, $U_m = 1$.

Lemma 1 (Coverage of the reference inversion)**.** *Under the above conditions,* $\Pr_\theta\{L_m(S_m;\delta) \le \theta \le U_m(S_m;\delta)\} \ge 1-\delta$.

Proof. For fixed θ, $q_m(t;\theta)$ is nonincreasing in t. The rejection event $q_m(S_m;\theta) < \delta/2$ is an upper-tail region and has probability at most $\delta/2$ by Equation 18, including a limit at an open boundary. If $L_m(S_m;\delta) > \theta$, this rejection event occurs. Applying the same argument to $1 - W_s$ bounds the probability that the upper endpoint is below θ by $\delta/2$. A union bound proves coverage. □

Computation and rescaling. Because $q_m(s;\theta)$ is nondecreasing in θ, bisection computes each lower endpoint. Start from $[l,h] = [0,1]$; at midpoint c, set $l = c$ if $q_m(s;c) < \delta/2$ and $h = c$ otherwise. Return the left endpoint when the bracket reaches the chosen numerical precision. The resulting lower endpoint is rounded outward; applying the same routine to $m - s$ and subtracting from one rounds the upper endpoint outward. If a numerical comparison is uncertain, keep the current outer bracket. Evaluate binomial tails and their interpolation in log space.

For responses on $[a,b]$, normalize by $(V-a)/(b-a)$ and map the interval back by $u \mapsto a+(b-a)u$. Signed paired differences use $[a,b] = [-1,1]$, giving the map $u \mapsto 2u-1$. Computing each difference within its seed preserves dependence between the two endpoint outcomes. For all-zero normalized observations, the closed-form interval is

$$\left[0,\ 1-\left(\frac{\delta}{2\mathrm{e}}\right)^{1/m}\right],$$

which also checks the boundary and tail allocation.

A.3 Relation, regret, and full-loss certificates

Proof of Theorem 2. Let $\mathcal{G}$ be simultaneous coverage of the $4K$ endpoint means and $2K$ signed differences. A union bound gives $\Pr(\mathcal{G}) \ge 1-\alpha$. On $\mathcal{G}$, the quality–cost interval map covers each utility. For every action,

$$\max_b U_b^- - U_a^+ \le \max_b U_b - U_a \le \max_b U_b^+ - U_a^-.$$

This establishes the inner and outer set inclusions in Equation 7. Overlapping inner sets certify a common acceptable action; disjoint outer sets certify disjoint true sets. A difference interval contained in its closed tolerance interval proves equivalence for that coordinate. A difference interval disjoint from its tolerance interval proves change. Combining these predicates in the stated order proves the relation claims.

For regret, exclude the self-difference, which is exactly zero. Each competitor difference $U_b - U_a$ lies in $[U_b^- - U_a^+, U_b^+ - U_a^-]$. Taking the maximum with zero proves Equation 8 simultaneously for every action, including the policy-selected action. □

The primary rule certifies a uniquely acceptable action a at endpoint z when

$$U_a^-(z) > \max_{b\neq a} U_b^+(z) + \gamma_{\mathrm{cert}}. \tag{20}$$

The strict inequality is required because regret equal to γ_{cert} is acceptable. Different certified endpoint actions imply C. Otherwise, containment of all quality- and cost-difference intervals within tolerance implies E_{prof}; remaining cases are U. This rule and the full rule share interval inputs and differ in the overlap requirement for E.

Full-loss and risk bounds. The regret bounds imply $b_i^{j,-} \le b_i^j \le b_i^{j,+}$. Logical or is monotone in each binary argument, proving Equation 9. On joint coverage across pairs, summing these bounds over certified relations and using $0 \le \ell_{L_i} \le 1$ for unresolved relations proves Equation 10. In contrast, Equation 12 fixes observed primary labels and failures and completes unspecified binary records. Its endpoints are attained by assigning all unresolved indicators zero or one.

Error allocation and stopping. For pair i, look t, and interval coordinate d, choose error budgets with $\sum_{i,t,d} \delta_{i,t,d} \leq \alpha_{\text{cert}}$. The resulting coverage holds jointly without requiring independence between pairs or looks. The primary single-pair allocation is $0.05/(6K)$ per two-sided interval, or $0.05/12$ for two actions; each reference tail receives half. For J predeclared looks, equal allocation is $\alpha/(6KJ)$ per interval. Stopping at the first certifying look then selects a conclusion valid on the joint coverage event. The registered finite look set is $\{32, 64, 128, 256\}$.

A.4 Calibration and conservative decisions

For source g and profile coordinate d, use the residual

$$R_{g,d} = \max_{v} |\widehat{P}_d(x_{g,v}) - \overline{P}_d(x_{g,v})|,$$

where v indexes the original and paired candidate prompts. Within a prespecified source group, sort the n_{cal} residuals. With $k = \lceil (n_{\text{cal}} + 1)(1 - \beta) \rceil$, set $r_d = R_{(k),d}$ when $k \leq n_{\text{cal}}$ and $r_d = +\infty$ otherwise. A predictor fitted independently of calibration, together with exchangeable calibration and test source residuals, gives coordinatewise coverage of a new source residual at least $1 - \beta$. At $n_{\text{cal}} = 15$ and $\beta = 0.10$, $k = 15$ and the radii are finite.

Proof of Proposition 2. Let $L_a = \widehat{U}_a - e_a$. Under the joint true-utility error condition, $L_a \leq U_a \leq L_a + 2e_a$. Since the selected action maximizes L_a, $U_{\widehat{a}} \geq L_{\widehat{a}} \geq L_{a^*}$ while $U_{a^*} \leq L_{a^*} + 2e_{a^*}$. Subtraction gives the claim and allows tied optima. □

To include measurement error, suppose all coordinates satisfy $|\widehat{Q}_a - \overline{Q}_a| \leq r_a^Q$, $|\widehat{C}_a - \overline{C}_a| \leq r_a^C$, $|\overline{Q}_a - \mathbb{E}Q_a| \leq \xi_a^Q$, and $|\overline{C}_a - \mathbb{E}C_a| \leq \xi_a^C$. Set $e_a = r_a^Q + \lambda r_a^C$ and $\nu_a = \xi_a^Q + \lambda \xi_a^C$. The triangle inequality gives $|\widehat{U}_a - U_a| \leq e_a + \nu_a$. Keeping the original score $L_a = \widehat{U}_a - e_a$ yields

$$\text{Reg}_{\widehat{a}} \leq 2e_{a^*} + \nu_{a^*} + \nu_{\widehat{a}}. \tag{21}$$

Indeed, $U_{\widehat{a}} \geq L_{\widehat{a}} - \nu_{\widehat{a}} \geq L_{a^*} - \nu_{\widehat{a}}$ and $U_{a^*} \leq L_{a^*} + 2e_{a^*} + \nu_{a^*}$. If coordinate residual failure probabilities are β_d and joint measurement error probability is at most δ_{meas}, the combined event has probability at least $\max\{0, 1 - \sum_d \beta_d - \delta_{\text{meas}}\}$. This states the probability scope of the true-utility conclusion explicitly.

An independent utility interval can also certify the error premise directly. If $[U_a^-, U_a^+] \subseteq [\widehat{U}_a - e_a, \widehat{U}_a + e_a]$ for every action, simultaneous interval coverage implies the condition of Proposition 2.

A.5 Sensitivity of the fixed-predictor comparison

Combining marginal confidence intervals. For risks r_1, r_0 with marginal intervals $[l_1, u_1]$ and $[l_0, u_0]$, both intervals covering implies $r_1 - r_0 \in [l_1 - u_0, u_1 - l_0]$. Two intervals with coverage at least 0.95 have joint coverage at least 0.90 by a union bound. The source-shift intervals therefore give $[9.86 - 27.36,\ 15.42 - 20.42] = [-17.50, -5.00]$ percentage points for PORC-LCB minus PORC-point.

Replacing evaluation records. Fix policies, predictions, calibration, evaluation rules, and normalized weights. A pair's difference of two binary losses lies in $[-1, 1]$, so replacing one equally weighted record changes the risk difference by at most $2/N$. Replacing at most k shared labels or complete outcome records therefore changes it by at most $2k/N$. If a replaced prompt is reused in several records, every affected record counts toward k. With $k = 36$ of $N = 720$, the observed -11.39 percentage-point difference remains in $[-21.39, -1.39]$. For nonuniform weights w_i, the bound is $2 \sum_{i \text{ replaced}} w_i$. This calculation holds the evaluated policies fixed.

Repricing fixed actions. For two fixed policies, the utility difference at weight $\widetilde{\lambda}$ is $\Delta Q - \widetilde{\lambda} \Delta C$. Primary PORC versus Outcome-greedy gives $0.036 + 0.112\widetilde{\lambda}$; source-shift PORC-LCB versus PORC-point gives $0.072 - 0.128\widetilde{\lambda}$. These calculations reuse measured quality and cost for the existing choices. Risk labels retain their original weight.

B EXECUTION AND REPRODUCIBILITY DETAILS

B.1 SOURCES, PROMPTS, AND OUTCOME MEASUREMENTS

Table 6 records the units used in the three experiments. Source identities define the split and resampling unit; multiple edits, checkpoints, and rollouts remain nested within a source. In the primary study, each dataset's 30 validation sources split into 15 tuning and 15 calibration sources. The two-action checkpoint penalties are $\rho \in \{0.05, 0.075, 0.10\}$ and are separate from both regret thresholds.

Table 6: Study sizes and roles. A profile contains outcomes for all registered actions on one prompt. Pair–checkpoint records can share a source.

Study stage	Units and execution budget
Primary construction	60 sources per dataset; 4 rollouts per action and prompt
Primary validation	30 sources per dataset, split equally for tuning and calibration
Primary evaluation	60 sources; 240 pairs per checkpoint; 720 records; 128 rollouts per action and prompt; 368640 executions
Shift development	120 sources; 720 pair–checkpoint records
Post-hoc source shift	240 sources; 720 pair–checkpoint records
Three-action construction	48 sources; 144 profiles; 96 pairs; 4 rollouts; 1728 executions
Three-action validation	32 sources; 96 profiles; 64 pairs; 4 rollouts; 1152 executions
Three-action evaluation	32 sources; 96 profiles; 64 pairs; 128 rollouts; 36864 executions

Prompt construction. Primary transformations add a candidate answer for verification, supply a helpful interval, edit a modulo-7 requirement, or change Markdown, politeness, or whitespace. Their intended relation is a construction label; measured outcomes determine $E_{\text{prof}}/C/U$ or full $E/P/C$ relations. External pairs include a wrapper with unchanged answer requirements and a two-question prompt that requires two answers. For example, the hash-selected Coinland question asks for the digit sum of the largest price not representable using 6-, 10-, and 15-cent coins. Its wrapper adds a polite instruction; its two-question variant adds an independent complex-arithmetic problem with explicit A/B answer requirements.

Scoring and execution. The extractor reads the last `FINAL_ANSWER` line with a boxed-answer fallback and fixed conservative normalization before exact comparison to gold. Two-answer prompts use `FINAL_A`/`FINAL_B`, explicit sections, or exactly two boxed answers. Extraction itself does not read the gold answer. Quality is the resulting exact-match indicator. In the primary study, cost counts generated tokens including the forced prefix, clips at 8192, and divides by 8192; a budget violation is $C > 0.5$. SHORT inserts `</think>` before generation; THINK suppresses this token and EOS on the first step. Temperature is 0.6, top-p is 0.95, beam count is one, and the primary context cap is 16384.

External actions select Qwen2.5-0.5B/1.5B/7B Instruct. Cost clips generated length at 2048 and divides by 2048. This normalization measures output tokens and does not convert different model sizes to FLOPs, latency, or monetary cost.

B.2 PREDICTOR FITTING AND CALIBRATION

Word TF–IDF uses 1–2-grams; character TF–IDF uses `char_wb` 3–5-grams. Each block has minimum document frequency 2 and at most 32768 features. TruncatedSVD uses at most 256 dimensions with seed 1729. Feature fitting and standardization use construction data. Checkpoint identifiers are fixed by content hash. The shared predictor has 128 GELU hidden units, dropout 0.1, and sigmoid quality–cost outputs for each action.

The objective combines profile mean squared error, equivalent-pair prediction regularization with weight 0.5, and a critical-switch hinge with weight 0.5 and margin 0.03. Construction profiles determine training relations and target actions. Optimization uses AdamW, learning rate 10^{-3}, weight

decay 10^{-4}, at most 400 full-batch epochs, and early-stopping patience 40. Initialization and dropout seeds are 17, 23, and 41; all are retained in the summaries.

Calibration uses maximum absolute residuals over each source's original, equivalent-candidate, and critical-candidate prompts, separately for every action's quality and cost. The held-out calibration half has 15 sources per dataset, with the finite-sample-corrected 90% quantile described in Appendix A.4. Policy selection and calibration precede held-out outcome generation. Deployment reads the input's predicted profile; forced held-out outcomes enter evaluation.

B.3 Comparison and reporting conventions

Outcome-greedy is a learned predictor of forced quality and cost followed by point-utility maximization. PORC-point uses the PORC predictor without its score offsets; PORC-LCB uses Equation 14 with the same predictions and tie rule. Constant policies always select their named action. These distinguish a comparison of complete learned pipelines, a comparison of decision rules on shared predictions, and the behavior of fixed actions.

Primary PPFR and CRMR are full-sample joint rates on fixed E_{prof} and C protocol labels. Source-shift risk is the average full point-label loss. For three-action evaluation, the full-sample contribution of relation L and the corresponding conditional rate are

$$\widehat{R}_L^{\mathrm{full}} = \frac{1}{N}\sum_i \mathbf{1}\{\widehat{L}_i = L\}\overline{\ell}_L(i), \qquad \widehat{R}_L^{\mathrm{cond}} = \frac{\sum_i \mathbf{1}\{\widehat{L}_i = L\}\overline{\ell}_L(i)}{\sum_i \mathbf{1}\{\widehat{L}_i = L\}}. \tag{22}$$

The conditional rate requires a nonzero denominator. Full-sample contributions sum to $\widehat{R}_{\mathrm{full}}$. Table 5 retains the reported full-sample contributions. Means and rates are displayed at the original reporting precision.

Primary certification uses a pair-level error budget of 0.05 with equal allocation across $6K$ two-sided interval inputs. The reference inversion in Appendix A.2 proves the theoretical coverage statement. Reported primary risks retain the original protocol labels, while full point-label risks are computed from rollout means. This preserves the distinction between recorded empirical comparisons and guarantees conditional on valid interval inputs.

Bootstrap draws resample entire sources, keeping all associated edits, checkpoints, and outcomes together. Primary inference uses 10000 draws over 60 sources and retains the original 12-contrast Holm family. The post-hoc source-shift marginal intervals use 10000 draws over 240 sources. Empirical completion, interval combination, and record replacement have the definitions given above and do not change the resampling unit.