WHEN SHOULD LLM AGENTS TALK? MEASURING COORDINATION VALUE AND AUDITING LEARNED ROUTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

After a language-model agent produces an initial answer, the system must decide whether further coordination is worth the inference cost. Even if an initial answer is assessed as incorrect, coordination may not improve it enough to justify its cost. We define the policy-relative Causal Value of Coordination (CVC) as the expected difference in cost-adjusted utility between a coordination action and a feasible alternative. Both start from the same history and follow the same fixed continuation policy. With the incumbent policy as the reference, a performance-difference identity links exact CVC values to end-to-end policy improvement. We estimate CVC through paired counterfactual replay and train ridge and neural predictors to construct candidate routing policies. Frozen candidates are assessed through full-policy comparisons and audits. Our main evaluation uses 2,000 MMLU-Pro questions across two backbones, complemented by supplementary GSM8K and MMLU-Pro experiments. Confidence predicts initial errors, but the prespecified router comparisons do not establish a utility gain. A diagnostic with access to initial correctness yields larger gains, highlighting the gap between error prediction and effective routing. A fixed early-stopping policy reduces voting cost without changing answers; observable-failure recovery yields modest secondary gains. Learned-policy lower bounds remain negative, and no candidate passes the supplementary independent audits. These findings highlight that identifying beneficial coordination, learning effective routing policies, and certifying end-to-end improvements are distinct challenges under the tested conditions.

1 INTRODUCTION

After a language model (LM) agent produces an answer, the system can return it, request self-review, consult a peer, or collect more votes. An additional response may correct a mistake, overturn a correct answer, or leave the result unchanged while adding inference cost. Therefore, the decision depends on more than uncertainty about the initial answer. We ask whether information available at that decision point can identify when coordination is likely to improve the final result enough to justify its cost.

Prior work provides several methods for generating and combining additional responses through debate and voting (Du et al., 2023; Liang et al., 2024). Early-stopping rules reduce unnecessary votes (Liu et al., 2026c), while learned methods choose among available actions or assign credit to earlier decisions (Fan et al., 2026a; Liu et al., 2026b; Chen et al., 2026b). These methods provide the actions and comparison tools for our study. We examine how well a fixed set of agents can use them to choose between coordination and a stated alternative under a declared inference cost.

We call the expected difference in cost-adjusted utility between those choices the policy-relative Causal Value of Coordination (CVC). Both choices begin from the same observed history and follow a specified continuation. When the alternative is the current policy, this quantity is the standard advantage of the current policy (Kakade & Langford, 2002). We estimate the difference by running both choices from shared histories, using those measurements to train candidate routing policies, and evaluating the frozen policies on held-out questions. This lets us examine separately whether an initial error is predictable, whether consultation helps after that error, and whether the resulting

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.

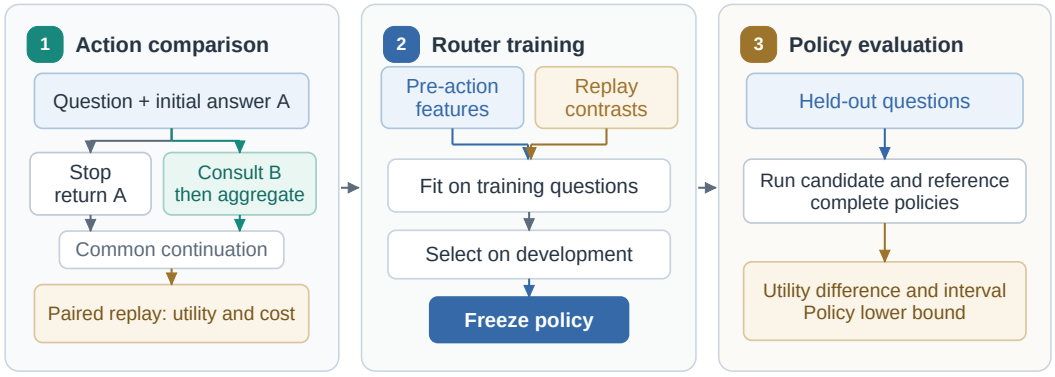


Figure 1: **CVC measurement and full-policy evaluation.** Paired replay compares stopping with peer consultation from the same question and initial answer under a common continuation, yielding utility and cost contrasts. These contrasts and observable features support policy fitting and development selection. On held-out questions, we evaluate frozen candidate and reference policies end to end and report utility differences, uncertainty intervals, and lower bounds on gains.

policy improves overall performance. Figure 1 shows how evidence about one coordination decision is turned into a test of the complete policy.

Our main evaluation uses MMLU-Pro questions for Qwen3-8B and Gemma-3-12B-it, withheld from fitting and selecting coordination policies. Decision rules are trained based on response confidence, observable text features, or frozen embeddings of the question and initial answer. Each rule chooses whether to return the initial answer or consult a peer. We compare them with answering alone and simpler consultation strategies. We also study decisions made after a peer response against a fixed three-vote adaptation of Efficient Majority-then-Stopping (EMS) (Liu et al., 2026c). Supplementary experiments examine alternative reference actions and assess frozen policies on separate held-out questions.

Confidence helps predict initial errors, including among responses that are complete and parsable. Nevertheless, the primary comparisons do not establish a cost-adjusted gain for the selected learned policies. Consultation may repair an error, leave it unresolved, or replace a correct answer, and it incurs extra inference cost; predicting an error therefore does not identify a beneficial action by itself. A diagnostic policy given the initial answer's correctness yields larger gains, illustrating the value of information unavailable at deployment. A rule targeting incomplete responses gives modest secondary gains, while early stopping saves voting calls without changing the full three-vote procedure's answers. The tested increases in training data or replay count do not show a consistent improvement, and all reported main-evaluation lower bounds for learned policies remain negative. These conclusions apply to the policies, models, and budgets evaluated here.

Overall, there are three main contributions in this work:

- We identify the problem of measuring coordination value without a specified alternative and continuation. We formalize the expected cost-adjusted contrast and relate its incumbent-policy form to established policy-improvement results.

- We develop a measurement and evaluation procedure that uses paired replay for supervision, separates development selection from held-out policy evaluation, and reports prediction diagnostics alongside performance contrasts and policy bounds.

- We evaluate confidence, embedding, and text-feature routers and distinguish initial-error prediction from coordination utility through privileged-information, repair, harm, and continuation analyses. EMS-style stopping (Liu et al., 2026c) and observable-failure controls identify savings and recovery opportunities that complement the learned-policy comparisons.

2 Related Work

Coordination and adaptive inference. Multi-agent reasoning combines proposals through debate, critique, revision, or voting (Du et al., 2023; Liang et al., 2024; Chan et al., 2024). Its benefits depend on protocol, confidence, and agent diversity (Marandi, 2026; Zhu et al., 2026; Liu et al., 2026d); self-revision spends additional inference without consulting another agent (Madaan et al., 2023; Shinn et al., 2023). Adaptive consistency and early-stopping self-consistency allocate sampling effort from observed responses (Aggarwal et al., 2023; Li et al., 2024), while FrugalGPT and RouteLLM allocate requests across models (Chen et al., 2023; Ong et al., 2025). These approaches motivate evaluating confidence and cost together. Within multi-agent systems, Intelligent Multi-Agent Debate (iMAD) (Fan et al., 2026a), EMS (Liu et al., 2026c), and Dynamic Auction-based Language Agent (DALA) (Fan et al., 2026b) adapt critique, voting, and communication allocation, respectively. EMS supplies the prior stopping rule used in our fixed and post-peer comparisons. Its structural savings provide a reference for improvements attributed to learned routing.

Value estimation and routing. Cost-aware inference asks whether another computation is worth its expected benefit (Russell & Wefald, 1991; Sabbata et al., 2025). Recent LLM systems use confidence to decide when to escalate an answer (Chuang et al., 2025), combine routing with model cascades (Dekoninck et al., 2025), or learn multi-round model calls and aggregation (Zhang et al., 2026). These methods allocate inference across models; our setting asks whether an agent with an initial answer should consult a peer. TreeCredit is closer to this setting: it compares collaboration operators from a shared prefix and learns from their best sampled continuations (Liu et al., 2026b). CVC likewise compares actions at a common history, but estimates their expected cost-adjusted outcomes under a specified continuation.

Counterfactual credit and policy improvement. Classical multi-agent methods use counterfactual action values for credit assignment and communication selection (Foerster et al., 2018; Ding et al., 2020). Recent LM-agent work trains collaborating models from joint rewards (Liu et al., 2026a), assigns credit to edits of an orchestration program (Chen et al., 2026a), or replays alternative messages from fixed histories to estimate per-decision advantages (Chen et al., 2026b). The fixed-history replay in C3 is closest to our measurement procedure. CVC measures the value of a feasible runtime coordination action relative to a specified alternative, with the agents and continuation fixed and inference cost included. We then evaluate the learned decision rule as a complete policy on held-out questions. With the incumbent policy as reference, CVC is standard action advantage (Kakade & Langford, 2002). Our held-out lower-bound comparisons are related to high-confidence policy improvement (Thomas et al., 2015); they use paired outcomes from complete policies rather than off-policy return estimates.

3 Problem Formulation

3.1 Traces, Objective, Actions, and References

A task $x \sim \mathcal{D}$ is solved by N agents. A coordination policy π maps the history before each decision to a distribution over feasible actions; models, prompts, and other system components remain fixed. A trace is $\tau = (x, h_1, a_1, \ldots, h_{T-1}, a_{T-1}, h_T, y)$, where h_T is terminal, y is the final answer, and T is bounded by a hard budget. Every routing feature must be observable in its history. We score a trace and a policy by

$$J_\lambda(\tau; x) = U(y, x) - \lambda \,\mathrm{Cost}(\tau), \qquad J_\lambda(\pi) = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim\pi}[J_\lambda(\tau; x)]\,, \tag{1}$$

where Cost is a predeclared inference-cost measure and $\lambda \geq 0$ converts cost to utility units. A coordination action transmits information between agents or consults additional agents (Appendix D.1); self-review is a non-communication control. Each comparison names a feasible reference action a^0 at the same history. Fixed references such as Stop support diagnostic comparisons; policy improvement instead compares with the incumbent's action distribution.

The decision time determines which costs enter the comparison. Requesting a new peer response incurs generation cost, whereas revealing a response that has already been generated does not incur that cost again. The common prefix contributes equally to both branches and cancels in their

difference. Specifying the history, action, and reference therefore determines both the information intervention and the resources attributed to it.

3.2 Policy-Relative Causal Value of Coordination

Fix a continuation policy π. At history h and event t, define

$$Q^{\pi}_{\lambda,t}(h,a) = \mathbb{E}[U(Y,x) - \lambda\, C_{t:}(\tau) \mid H_t = h,\ \mathrm{do}(A_t = a),\ A_s \sim \pi(\cdot \mid H_s) \text{ for } s > t], \quad (2)$$

where $C_{t:}$ is the remaining cost. The intervention executes a while every other component behaves as in deployment. Write Y^a and $C^a_{t:}$ for the terminal answer and remaining cost under this intervention and continuation. The utility contrast, cost contrast, and policy-relative CVC are

$$\begin{aligned} \Delta^{\pi}_{U}(h,a;a^0) &= \mathbb{E}^{\pi}\Big[U(Y^a,x) - U(Y^{a^0},x) \mid H_t = h\Big], \\ \Delta^{\pi}_{C}(h,a;a^0) &= \mathbb{E}^{\pi}\Big[C^a_{t:} - C^{a^0}_{t:} \mid H_t = h\Big], \\ \mathrm{CVC}^{\pi}_{\lambda,t}(h,a;a^0) &= Q^{\pi}_{\lambda,t}(h,a) - Q^{\pi}_{\lambda,t}(h,a^0) = \Delta^{\pi}_{U} - \lambda\,\Delta^{\pi}_{C}. \end{aligned} \quad (3)$$

Its sign depends on both the reference and the continuation: an action may help relative to self-review but not voting, or provide information that one continuation uses better than another. An answer change alone therefore does not establish positive value.

The reference and continuation have separate roles. The reference specifies the alternative at the current decision; the continuation specifies how either branch subsequently produces its answer. Holding the latter fixed allows the contrast to isolate the current action within the declared system. Changing the aggregation rule, model, or subsequent decision policy defines a different estimand, even when the message being evaluated is unchanged.

For an incumbent-policy reference, let $V^{\pi}_{\lambda,t}(h) = \mathbb{E}_{A\sim\pi(\cdot|h)}[Q^{\pi}_{\lambda,t}(h,A)]$. The standard advantage function is $A^{\pi}_{\lambda,t}(h,a) = Q^{\pi}_{\lambda,t}(h,a) - V^{\pi}_{\lambda,t}(h)$. We write $\mathrm{CVC}^{\pi}_{\lambda,t}(h,a;\pi) = A^{\pi}_{\lambda,t}(h,a)$ for this policy reference. It equals the action-reference contrast when the incumbent is deterministic; otherwise the reference averages over the incumbent's actions.

3.3 Policy Improvement and Performance Decomposition

The finite-horizon performance-difference lemma (Kakade & Langford, 2002) connects these local advantages to the policy objective (proofs in Appendix B).

Proposition 1 (Policy-relative decomposition)**.** *For any two coordination policies π and π' with bounded horizon,*

$$J_{\lambda}(\pi') - J_{\lambda}(\pi) = \mathbb{E}_{\tau\sim\pi'}\left[\sum_{t<T} A^{\pi}_{\lambda,t}(h_t,a_t)\right]. \quad (4)$$

Corollary 1 (Policy improvement under exact advantages)**.** *Let π' choose, at each history, an action with $A^{\pi}_{\lambda,t}(h,a) > 0$ when one exists, and otherwise act as π. Then $J_{\lambda}(\pi') \geq J_{\lambda}(\pi)$, strictly whenever such a deviation occurs with positive probability under π'.*

The identity is evaluated on histories visited by the new policy. Estimated gates face two gaps: earlier decisions can shift the distribution of later histories, and selecting positive predictions can favor overestimation errors. Neither a positive prediction nor marginal calibration closes these gaps. We therefore evaluate each frozen candidate end-to-end on questions withheld from fitting and selection; Appendix D.3 states the intervention and generalization assumptions.

These interventions are executable operations in a fixed LM system. Replaying both branches at a shared history holds task difficulty and the observed prefix fixed; it does not identify effects for unobserved histories or a different continuation. The formulation applies standard action values and policy-improvement theory to specified communication decisions. Positive value relative to stopping may still fall below an incumbent that uses a stronger action, so empirical comparisons name both their decision time and reference.

4 CVC-GUIDED COORDINATION

To decide whether to consult a peer, we estimate the change in answer utility and remaining cost relative to returning the initial answer. Paired replay compares these actions from the same observed history under a fixed continuation. We fit a decision rule to these estimates using only information available before consultation, then evaluate the resulting policy on held-out questions. Figure 2 summarizes the comparison and decision rule.

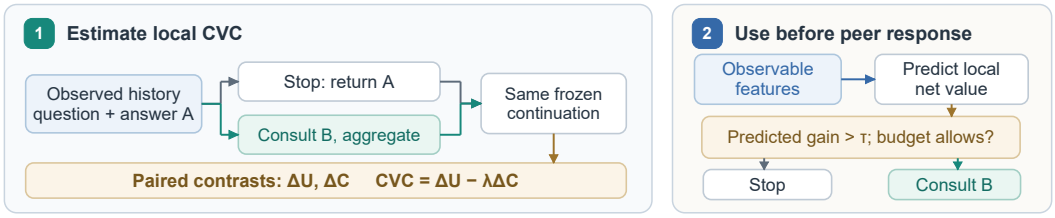


Figure 2: **Local CVC estimation and the pre-peer decision.** Paired replay compares returning the initial answer with consulting a peer from the same observed history; both branches use the same frozen continuation. The utility and remaining-cost contrasts yield CVC. At deployment, a policy predicts cost-adjusted value from pre-action features and consults when its development-selected threshold and budget allow.

Paired counterfactual replay. From a cached history, we execute a and its reference a^0, following the same frozen policy π_k afterwards. For M paired continuations, the targets are

$$\widehat{\Delta}_U = \frac{1}{M}\sum_{m=1}^{M}\left[U(Y_m^a, x) - U(Y_m^{a^0}, x)\right], \qquad \widehat{\Delta}_C = \frac{1}{M}\sum_{m=1}^{M}\left[C_{t:,m}^a - C_{t:,m}^{a^0}\right]. \tag{5}$$

These are unbiased estimates of the expected contrasts when each branch has the correct marginal distribution; their variance depends on the coupling. Pairing holds the question and observed prefix fixed. Repeated continuations reduce label noise at that history but do not increase the number of independent questions. The target $\widehat{\Delta}_U - \lambda\widehat{\Delta}_C$ also permits rescoring at another cost weight without collecting new responses.

Pre-action prediction. The primary task is binary routing after the initial answer: Stop returns that answer, while Coordination requests a peer response and aggregates the two answers. We compare the original 55-dimensional representation with frozen question/answer embeddings and observable confidence features, using linear and nonlinear predictors. Confidence-based initial-failure classifiers provide a separate comparison. Reference correctness and future responses supply supervision, not deployable features. Under matched linear designs, penalties, and bootstrap samples, fitting action values, paired differences, or separate utility and cost heads yields the same value contrast (Appendix E).

Routing decisions. Value routers coordinate when their predicted gain exceeds a development-selected threshold τ. Ensemble-margin variants subtract $\beta(\widehat{\sigma}_U(h) + \lambda\widehat{\sigma}_C(h))$, using the ensemble standard deviations (SDs) of the utility and cost predictions; $\beta \geq 0$ is also selected on development data. The ensemble spread is a heuristic measure of uncertainty and does not provide a conditional-mean guarantee. The reference is retained otherwise, and actions exceeding the remaining budget are excluded. Supplementary four-action and conformal variants, including the distinction between replay-label coverage and value uncertainty, are documented in Appendix E.

After observing the initial and first peer responses, a separate policy chooses return, aggregation, or another peer for voting; every choice pays the shared prefix cost. Its comparator is our fixed three-vote EMS adaptation, EMS3 (Liu et al., 2026c), where 3 denotes the maximum number of responses. EMS3 returns the initial answer when the first two normally completed, valid answers agree; otherwise it collects the third response and applies Vote. Response order is fixed, without EMS's reliability ordering or history updates. Agreement makes further voting redundant, while aggregation remains a distinct action.

Candidate construction and evaluation

1. Collect paired training continuations and fit preprocessing and predictors on training questions.
2. Select regularization, thresholds, and a primary candidate by development utility, breaking ties by lower cost.
3. Freeze candidate parameters, decision rules, and comparison families before inspecting evaluation outcomes.
4. Evaluate complete policies on the evaluation set; report paired performance uncertainty and the declared audit separately.

Held-out full-policy audit. Let $D_i = J_\lambda(\tau_i^{\pi_{\text{cand}}}; x_i) - J_\lambda(\tau_i^{\pi_k}; x_i) \in [-R_\lambda, R_\lambda]$, where $R_\lambda = U_{\max} - U_{\min} + \lambda C_{\max}$ is a conservative radius; the interval width is $2R_\lambda$. For a predeclared family of at most K candidates and error allowance δ, the Hoeffding lower confidence bound (LCB) is

$$\text{LCB}_{\text{audit}} = \overline{D} - R_\lambda \sqrt{\frac{2\log(K/\delta)}{n}}. \tag{6}$$

Proposition 2 (Audited non-degradation)**.** *Suppose at most K candidates are audited, each frozen independently of its audit outcomes, with item-level differences independent and identically distributed (i.i.d.) on $[-R_\lambda, R_\lambda]$ for each comparison. If acceptance requires $\text{LCB}_{\text{audit}} > 0$, then with probability at least $1 - \delta$, every accepted candidate satisfies $J_\lambda(\pi_{\text{cand}}) \geq J_\lambda(\pi_k)$ on the audit distribution.*

This applies Hoeffding's inequality (Hoeffding, 1963) and a union bound (Appendix B). Declined candidates count toward K, and subsequent adaptive updates require previously unused audit data. Questions, rather than continuations or fitting seeds, are the independent units. A declined audit establishes neither benefit nor harm.

The supplementary experiments use their prespecified Hoeffding audits. The main evaluation also reports variance-sensitive empirical-Bernstein bounds (Maurer & Pontil, 2009), with the paired support width stated explicitly. Its performance estimates and audit summaries use the same evaluation set, independently of training and development. Because this set is sampled by category from a finite benchmark pool, the i.i.d. audit formulas are treated as approximations to that design; Appendix A gives the assumptions and error allocation.

5 Experimental Evaluation

We evaluate pre-action routing, initial-response diagnostics, and continuation policies on previously unused MMLU-Pro questions. Supplementary GSM8K and MMLU-Pro experiments provide supporting evidence on reference actions, fitting stability, and independent audits, with their data splits and testing rules reported separately.

Tasks and protocol. The main evaluation set contains 2,000 MMLU-Pro questions (Wang et al., 2024) for each of Qwen3-8B (Yang et al., 2025) and Gemma-3-12B-it. Both backbones use the same questions, sampled with category-proportional quotas after excluding every previously used item.

Policies and statistical comparisons. Stop returns the initial answer A; Coordination returns the aggregation C_0 of A and an independent peer B_0. Vote uses A, B_0, B_1; its three-vote early-stopping adaptation, EMS3, omits B_1 when the first two normally completed, valid answers agree (Liu et al., 2026c). Self-review revises A without a peer. The primary router selects between Stop and Coordination using observable features; a separate learned policy chooses Return-A, aggregation, or voting after observing $A + B_0$. Its comparator is EMS3, with the same prefix cost charged to both policies. A secondary failure trigger coordinates only when A is incomplete or has no valid parsed answer. Random routing uses the selected router's development coordination frequency, rather than matching evaluation cost.

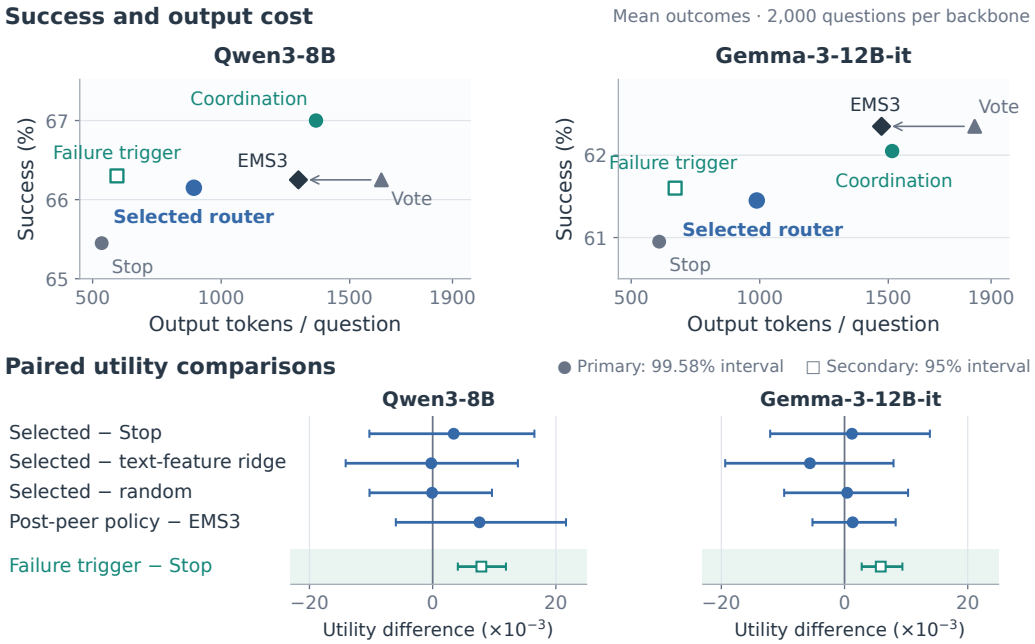


Figure 3: **Policy performance and inference cost.** Each backbone uses the same 2,000 questions at $\lambda = .01$. Top: Mean success and complete-path output cost; arrows connect Vote to its answer-preserving three-vote early-stopping variant (EMS3). Bottom: Paired utility contrasts with primary 99.58% Bonferroni confidence intervals (CIs; filled) and secondary pointwise 95% intervals (open). "Selected" denotes the selected router. Random routing matches development frequency; post-peer policies pay for $A + B_0$. Appendix Table 2 gives numerical summaries.

5.1 Learned Routing Gains Are Inconclusive

Figure 3 (bottom) shows inconclusive primary comparisons on both backbones. The selected embedding MLP has mean gains over Stop of .0034 and .0012; comparisons with text-feature and random routing likewise include zero. All twelve Holm-adjusted p-values equal one, with a minimum unadjusted value of .108. These intervals describe uncertainty in the performance of the frozen selected policies, without establishing equal policy performance or an absence of useful information in richer features.

Figure 3 (top) relates success to inference cost. The selected routers coordinate on 45.4% and 41.7% of questions, gaining 0.70 and 0.50 success points while adding 359 and 380 output tokens per question. Secondary alternatives show limited positive evidence: Qwen embedding ridge and Gemma text-feature ridge have gains over Stop of .0098 and .0068, with unadjusted intervals just above zero. These results do not establish a consistent advantage of either predictor family.

The failure trigger has positive secondary gains in Figure 3 (bottom). It acts on 50 Qwen and 49 Gemma questions, yielding gains over Stop of .0079 and .0059. All but one affected response is incomplete; the remaining Qwen response has no valid parse. The benefit therefore concerns recovery under the 2,048-token cap. Comparisons with the selected learned routers remain inconclusive.

EMS3 preserves Vote's response under the frozen voting rules, avoiding the third response on 72.6% and 72.5% of questions. The savings of 323 and 362 output tokens per question produce utility gains of .0032 and .0036 with identical success outcomes. This structural result concerns savings relative to Vote; improvement over Stop also depends on answer changes.

5.2 Initial Errors Do Not Determine Coordination Value

The privileged diagnostic receives initial correctness and learns on training questions to stop after success and coordinate after failure. Its gains over Stop are .0508 and .0504. Figure 4 (right) attributes

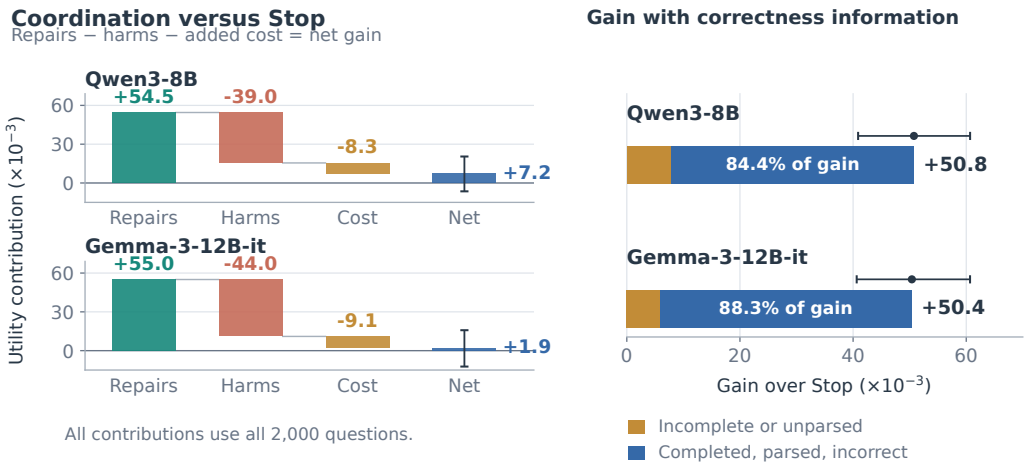


Figure 4: **Coordination outcomes and initial correctness.** Left: Always-Coordination minus Stop, decomposed into repairs, harms, and added cost. Right: Contributions to the privileged correctness-bit policy's gain over Stop; percentages give the share from completed, parsed but incorrect answers. Successful initial answers contribute zero. All contributions use the full 2,000-question denominator from Figure 3. Whiskers show pointwise 95% paired bootstrap intervals for net gains on the left and total gains on the right.

84% and 88% to completed, parsed but incorrect answers. The remaining contribution equals the failure trigger's gain. This diagnostic measures the benefit of unavailable correctness information within a restricted policy class, rather than an upper bound on all routers.

Confidence features contain information about these errors. Initial-error discrimination is measured by the area under the receiver operating characteristic curve (AUC). The selected confidence classifier has AUC .745 on Qwen and .774 on Gemma; among completed, validly parsed responses alone, AUC remains .726 and .759. Its utility advantage over Stop is nevertheless uncertain: .0090 $[-.0026, .0207]$ and .0056 $[-.0063, .0173]$. Initial-error discrimination and the expected benefit of a subsequent action are distinct prediction targets.

Figure 4 (left) shows how these outcomes combine. On the full evaluation set, repairs contribute about .055 on each backbone, but harms and added cost offset most of this gain. The resulting Coordination-minus-Stop contrasts are .0072 and .0019, with pointwise intervals containing zero. Conditional on initial failure, repair rates are 15.8% and 14.1%; conditional on initial success, harm rates are 6.0% and 7.2%. Development-defined confidence bins likewise show increasing error frequency without consistently increasing net utility. These descriptive patterns distinguish initial-error prediction from action-value prediction without identifying a unique cause of router performance.

5.3 Simple Rules Show Narrow Benefits

After observing $A + B_0$, learned routing has gains over EMS3 of .0076 and .0013, with adjusted intervals containing zero (Figure 3 (bottom)). Direct aggregation-minus-EMS3 contrasts are also inconclusive. Because every policy pays for the shared prefix, these comparisons address continuation after acquiring a peer rather than the value of acquiring that peer.

Secondary response-stratum analyses associate continuation value with the observed state. Aggregation exceeds EMS3 voting when the initial response is incomplete or invalidly parsed, based on only 50 and 49 questions; its mean contrasts are negative under valid agreement. Separately, pruning agreement-state Vote decisions to Return-A while holding the learned policy fixed preserves success and saves 66 and 42 output tokens per question. Comparisons of the pruned policy with EMS3 remain inconclusive. Appendix A reports the stratum estimates and distinguishes pruning from refitting a constrained candidate.

Table 1: **Variance-aware bounds on policy gains.** Differences from Stop ($\times 10^{-3}$) on the evaluation set. Empirical-Bernstein (EB) bounds use width 2.04096 and $\delta = .025$ for each of two candidates per backbone. The iid premise approximates the stratified design; these are not additional independent audits.

Backbone	Candidate	Mean gain	EB penalty	EB lower bound
Qwen3-8B	Selected router	$+3.4$	24.6	-21.2
Qwen3-8B	Text-feature ridge	$+3.6$	19.3	-15.6
Gemma-3-12B-it	Selected router	$+1.2$	24.2	-23.0
Gemma-3-12B-it	Text-feature ridge	$+6.8$	20.4	-13.6

The primary four-minus-one-replay gains are $-.0009$ and $-.0051$, and matched training-size comparisons at two replays are inconclusive for both ridge representations. The eight-replay subset has two positive pointwise intervals among eight comparisons, both on Qwen. These extensions do not establish a consistent benefit from the tested increases in supervision, while leaving effects at larger scales unresolved. Appendix A also reports the four-action extension.

5.4 Bounds Limit Claims of Policy Improvement

Rescoring frozen policies over $\lambda \in \{0, .005, .01, .02, .05\}$ changes the balance between accuracy and inference cost. At $\lambda = .05$, the selected routers' gains over Stop are $-.0109$ and $-.0140$, with pointwise intervals below zero. The failure trigger retains positive pointwise gains of $.0055$ and $.0034$, reflecting its low activity. Development-reselected thresholds are reported separately from frozen-policy rescoring. Input-plus-output sensitivity on the supplementary data sets (Appendix F) and the recorded feature-computation budget provide additional accounting perspectives; generation timer sums are not end-to-end latency measurements.

In Table 1, empirical-Bernstein penalties of .019–.025 exceed mean gains of .001–.007. All four lower bounds are negative. These calculations use the evaluation set itself and the paired difference-range width. The iid premise approximates category-stratified sampling without replacement; exact finite-sample coverage is not established for this design.

In the supplementary independent audits, none of the 24 candidates pass, including six MMLU-Pro candidates with no observed departures. Their prespecified 1,000-question Hoeffding penalty is .110. A separate exploratory bound exploiting EMS3's narrow cost-saving range supports answer-preserving savings over Vote, without changing the learned-candidate audit decisions. The appendix reports supplementary fitting and scoring diagnostics; GSM8K uses held-out partitions of its original training split, not the official test set.

6 Conclusion

After an LM agent answers, consulting a peer may improve the final answer, but it may also cause harm and adds inference cost. Predicting an initial error alone cannot establish whether consultation is worthwhile. We define the Causal Value of Coordination (CVC) as the expected difference in cost-adjusted outcomes between actions at the same decision point. Paired replay estimates this difference under a fixed continuation policy. Held-out evaluation then tests complete decision rules.

On MMLU-Pro, confidence predicts initial errors, but the prespecified learned-policy comparisons remain inconclusive. Consultation repairs some errors; it also introduces errors and additional cost. A diagnostic with access to initial-answer correctness produces larger gains, but that information is unavailable at deployment. A simple rule for incomplete or unparsed answers provides modest secondary gains. The reported lower bounds do not establish improvement for learned policies.

The limitations of this work are its focus on selected models, tasks, actions, and budgets. Output tokens are the main cost measure; we do not assess end-to-end latency. The policy bounds use an i.i.d. approximation for category-stratified sampling, so exact finite-sample coverage is unresolved.

AI USE STATEMENT

In this work, we used generative AI tools to draft and edit portions of the manuscript and to create or edit software code. We have not used generative AI tools for other tasks with required disclosure. All authors checked and revised the AI-assisted writing. One author verified the AI-assisted code, and two authors verified the reported results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Pranjal Aggarwal, Aman Madaan, and Yiming Yang. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023. URL https://aclanthology.org/2023.emnlp-main.761/.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pp. 9079–9093, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25cc3adf8c85f7c70989cb8a97a691a7-Abstract-Conference.html.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance, May 2023. URL http://arxiv.org/abs/2305.05176. arXiv:2305.05176 [cs.LG].

Xudong Chen, Yixin Liu, Hua Wei, and Kaize Ding. LEMON: Learning Executable Multi-Agent Orchestration via Counterfactual Reinforcement Learning, May 2026a. URL http://arxiv.org/abs/2605.14483. arXiv:2605.14483 [cs.AI].

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact Is Easier: Credit Assignment for Cooperative LLM Agents, May 2026b. URL http://arxiv.org/abs/2603.06859. arXiv:2603.06859 [cs.LG].

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to Route LLMs with Confidence Tokens, June 2025. URL http://arxiv.org/abs/2410.13284. arXiv:2410.13284 [cs.CL].

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A Unified Approach to Routing and Cascading for LLMs, May 2025. URL http://arxiv.org/abs/2410.10347. arXiv:2410.10347 [cs.CL].

Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. *Advances in neural information processing systems*, 33:22069–22079, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/fb2fcd534b0ff3bbed73cc51df620323-Abstract.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate, May 2023. URL http://arxiv.org/abs/2305.14325. arXiv:2305.14325 [cs.CL].

Wei Fan, JinYi Yoon, and Bo Ji. imad: Intelligent multi-agent debate for efficient and accurate llm inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29403–29411, 2026a. URL https://ojs.aaai.org/index.php/AAAI/article/view/40181.

Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Chengpei Tang, Jian Wang, and Keze Wang. Cost-effective communication: An auction-based method for language agent interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29412–29420, 2026b. URL https://ojs.aaai.org/index.php/AAAI/article/view/40182.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/full/10.1080/01621459.1963.10500830.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002. URL https://dl.acm.org/doi/abs/10.5555/645531.656005.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14751–14768, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024. URL https://aclanthology.org/2024.emnlp-main.992/.

Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. Llm collaboration with multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32150–32158, 2026a. URL https://ojs.aaai.org/index.php/AAAI/article/view/40487.

Yiqing Liu, Zihao Wang, Hantao Yao, Wu Liu, and Yongdong Zhang. Shared Prefixes, Better Credit: Adaptive Routing for Multi-Agent Reasoning, August 2026b. URL http://arxiv.org/abs/2608.02291. arXiv:2608.02291 [cs.AI].

Yiqing Liu, Hantao Yao, Wu Liu, and Yongdong Zhang. EMS: Multi-Agent Voting via Efficient Majority-then-Stopping, June 2026c. URL http://arxiv.org/abs/2604.02863. arXiv:2604.02863 [cs.AI].

Yuhan Liu, Juntian Zhang, Yichen Wu, Martin Takac, Salem Lahlou, Xiuying Chen, and Nils Lukas. Breaking the Martingale Curse: Multi-Agent Debate via Asymmetric Cognitive Potential Energy, March 2026d. URL http://arxiv.org/abs/2603.06801. arXiv:2603.06801 [cs.AI].

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, and Yiming Yang. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Ramtin Zargari Marandi. The impact of multi-agent debate protocols on debate quality: a controlled case study, March 2026. URL http://arxiv.org/abs/2603.28813. arXiv:2603.28813 [cs.MA].

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization, July 2009. URL http://arxiv.org/abs/0907.3740. arXiv:0907.3740 [stat.ML].

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

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial intelligence*, 49 (1-3):361–395, 1991. URL https://www.sciencedirect.com/science/article/pii/000437029190015C?casa_token=gEP3TDyeA64AAAAA:MnvFeISWToz1ZblC6pzKKENQpKcgb3k8zQssg51mId8B-egxKOL4jrLUd7IGH1mJhR5nRI_cv9Iw.

C. Nicolò De Sabbata, Theodore R. Sumers, Badr AlKhamissi, Antoine Bosselut, and Thomas L. Griffiths. Rational Metareasoning for Large Language Models, June 2025. URL http://arxiv.org/abs/2410.05563. arXiv:2410.05563 [cs.CL].

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388. PMLR, 2015. URL http://proceedings.mlr.press/v37/thomas15.html.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, and Ziyan Jiang. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, May 2025. URL http://arxiv.org/abs/2505.09388. arXiv:2505.09388 [cs.CL].

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:141233–141265, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ceaa137fce916aba5c65fceb1309088b-Abstract-Conference.html.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33909–33930, 2026. URL https://aclanthology.org/2026.findings-acl.1694/.

A Primary Evaluation: Design and Additional Results

A.1 Data, Generation, and Candidate Selection

The main evaluation uses the same 2,000 MMLU-Pro questions for Qwen3-8B and Gemma-3-12B-it. Removing all previously used questions leaves 3,131 unique normalized stems. Sampling is without replacement, with category-proportional integer quotas; the target population is this remaining pool. The primary fit retains 850 training and 200 development questions and averages four Coordination replays per training question: two earlier and two new peer–aggregation pairs sharing the initial response. Separate extensions compare 850 with 1,700 training questions at $M = 2$, and $M \in \{1, 2, 4, 8\}$ on a fixed 256-question subset. The primary fit therefore uses 850 questions at $M = 4$, not the largest training set and replay count together.

Generation uses the original bfloat16 (BF16) checkpoints and prompts, with Qwen thinking disabled. Initial and peer responses use temperature .7 and top-$p = .95$; aggregation and self-review are greedy. The per-call cap is 2,048 tokens. Each question has responses A, S, B_0, B_1, C_0, C_1: B_i sees the question and C_i sees (q, A, B_i). Coordination returns C_0; C_1 is a replay diagnostic, not an outcome-selection option. Utility charges the complete path, including A, at $\lambda = .01$ per thousand output tokens; input tokens and feature overhead are recorded separately. Success requires normal completion and a reference match. The normalized parser takes the last standalone uppercase option letter after the first case-insensitive whole word `answer`, allowing multiple letters in that suffix. Strict scoring takes the last line matching `Answer: X`; final-line compliance is recorded separately.

Text features have 55 dimensions. Rich predictors combine frozen question/response embeddings from Qwen3-Embedding-0.6B, six response properties, and thirteen confidence features (2,068 dimensions including the intercept). Confidence features come from teacher-forcing the saved response under its original model and prompt, including token log probabilities, entropy, and answer-letter probabilities. Candidates include ridge and one-hidden-layer value predictors, confidence-based and rich initial-failure classifiers, and category controls. Deployable features exclude reference answers and future responses. The privileged diagnostic receives the initial success bit and learns to Stop after success and Coordinate otherwise.

Preprocessing uses training questions only. Development utility selects configurations and thresholds under frozen tie-breaking rules. Across families, replay counts, seeds, and extensions, 982 configurations per backbone use the same 200 development questions. The selected Qwen and Gemma value models have 256 and 64 hidden units. Candidate parameters were hashed before evaluation; the protocol and amendments were recorded internally, without external preregistration. Five fit seeds reuse the same questions and are not independent evaluation samples.

Generation used two H100 PCIe 80GB GPUs and vLLM 0.11.0, retaining the earlier engine's prompt tokenization and decoding settings. After an engineering failure, Gemma used disabled prefix caching and at most 64 concurrent sequences; rejected runs supply no outcomes. Greedy checks agree with the earlier engine on 97.5% of Qwen and 95% of Gemma answers. On shared training histories, replay-success differences are below one percentage point, with intervals spanning zero and mean token ratios of .98–1.03. These checks do not establish identical generation distributions. The collection contains 45,796 calls across both backbones and all extensions; this is not a policy's online cost.

A.2 Primary Comparisons and Data-Size Sensitivity

The question is the analysis unit. We use 10,000 paired bootstrap resamples, retaining each question's policies and replays together and keeping fitted policies fixed. Table 3 reports six primary contrasts per backbone. Random routing is integrated analytically at development coordination probabilities .49 and .40, without matching evaluation cost. Holm adjustment covers all twelve tests; Bonferroni intervals have level $1 - .05/12$. The minimum unadjusted bootstrap-tail p-value is .108, and all adjusted values equal 1. These approximate tests do not establish equivalence.

Other comparisons use secondary, unadjusted pointwise intervals. At $M = 2$, doubling training size changes utility by $+.0013\,[-.0058, .0088]$ and $-.0024\,[-.0105, .0055]$ for text-feature ridge on Qwen and Gemma, and $-.0014\,[-.0084, .0056]$ and $-.0022\,[-.0106, .0064]$ for embedding ridge. On the 256-question subset, $M = 8$ versus $M = 1$ gives $+.0077\,[.0006, .0147]$

Table 2: **Success and complete-path output cost.** Evaluation set ($n = 2,000$ per backbone). Success requires normal completion and a normalized reference match (%); cost is mean output tokens per question. Selected router denotes the development-selected embedding MLP; EMS3 denotes the three-vote early-stopping adaptation of EMS; the failure trigger is a secondary fixed rule.

	Qwen3-8B			Gemma-3-12B-it		
Policy	Success	Tokens	$J_{.01}$	Success	Tokens	$J_{.01}$
Stop	65.45	535	0.6491	60.95	609	0.6034
Coordination	67.00	1369	0.6563	62.05	1515	0.6053
Vote	66.25	1623	0.6463	62.35	1836	0.6051
EMS3	66.25	1301	0.6495	62.35	1474	0.6088
Selected router	66.15	894	0.6526	61.45	988	0.6046
Failure trigger	66.30	595	0.6571	61.60	671	0.6093

Table 3: All primary contrasts on the evaluation set ($n = 2{,}000$ per backbone). Utility differences and endpoints are in 10^{-3} units. Text-feature denotes the ridge router using text features at $M = 4$; Selected denotes the development-selected primary router. Post-peer comparisons charge the common $A + B_0$ prefix. The last row per backbone compares the selected predictor family at $M = 4$ and $M = 1$. Pointwise intervals have 95% nominal coverage; Bonferroni intervals have level $1 - .05/12$ for the twelve-contrast family. All Holm-adjusted p-values equal 1.

Backbone	Contrast	ΔJ_λ	Pointwise interval	Bonferroni interval
Qwen	Selected − Stop	$+3.41$	$[-6.00, +12.76]$	$[-10.27, +16.52]$
	Selected − Text-feature	-0.23	$[-9.60, +9.23]$	$[-14.12, +13.83]$
	Selected − Random	-0.10	$[-6.84, +6.59]$	$[-10.25, +9.63]$
	Post-peer − EMS3	$+7.60$	$[-1.84, +17.11]$	$[-5.98, +21.64]$
	Aggregation − EMS3	$+6.82$	$[-5.69, +19.68]$	$[-12.10, +25.88]$
	Replay count 4 versus 1	-0.91	$[-8.39, +6.36]$	$[-11.80, +9.63]$
Gemma	Selected − Stop	$+1.20$	$[-7.91, +10.37]$	$[-12.07, +13.85]$
	Selected − Text-feature	-5.62	$[-15.01, +3.78]$	$[-19.37, +7.96]$
	Selected − Random	$+0.43$	$[-6.49, +7.35]$	$[-9.80, +10.32]$
	Post-peer − EMS3	$+1.32$	$[-3.15, +5.88]$	$[-5.21, +8.30]$
	Aggregation − EMS3	-3.41	$[-15.92, +8.98]$	$[-21.42, +14.22]$
	Replay count 4 versus 1	-5.09	$[-13.96, +3.73]$	$[-18.43, +7.34]$

and $+.0035\,[-.0023, .0095]$ for text-feature ridge, and $+.0051\,[-.0016, .0117]$ and $+.0004\,[-.0077, .0085]$ for embedding ridge. Qwen embedding ridge gains $+.0064\,[.0010, .0120]$ at $M = 8$ versus $M = 4$. Only two of eight replay intervals are positive; neither is replicated across backbones. Positive pointwise gains over Stop for Qwen embedding ridge and Gemma text-feature ridge are also secondary.

A.3 Initial Errors, Repair, and Observable Failures

Coordination repairs $109/691$ initially unsuccessful Qwen answers (15.77%) and $110/781$ Gemma answers (14.08%), but changes $78/1{,}309$ successful Qwen answers to failures (5.96%) and $88/1{,}219$ Gemma answers (7.22%). The observable-failure control acts on 49 incomplete Qwen responses plus one completed parse failure, and 49 incomplete Gemma responses. It repairs $17/50$ and $13/49$ cases, so its gains mainly reflect recovery under the generation cap.

Initial-failure AUC is .745 and .774 overall, and .726 and .759 among completed, validly parsed responses. The classifier's utility contrasts with Stop remain uncertain: $+.0090\,[-.0026, .0207]$ and $+.0056\,[-.0063, .0173]$. The privileged diagnostic gains $+.0508\,[.0409, .0607]$ and $+.0504\,[.0406, .0607]$; 84.4% and 88.3% come from completed, parsed but incorrect answers. These gains use unavailable correctness information and are not a deployable performance bound.

Table 4 uses quintile boundaries fixed on development scores. Initial failure increases across bins, but net coordination value is not consistently monotone and every utility interval includes zero. Qwen's highest-error bin has 44 repairs, 36 harms, and $1{,}174$ additional output tokens per question. The analysis is descriptive and does not isolate the cause of router performance.

Table 4: Exploratory confidence-bin analysis. Bin boundaries are development-score quintiles, ordered by predicted initial failure. Error is the initial failure percentage; repairs and harms count opposite success transitions after Coordination; added tokens are mean output tokens beyond Stop. Utility differences from Stop and pointwise 95% intervals are in 10^{-3} units. The bins contain different numbers of evaluation questions and are not used to select a new policy.

Backbone	Bin	n	Error (%)	Repairs	Harms	Added tokens	ΔJ_λ [95% CI]
Qwen	1	394	12.2	8	5	682	$+0.80[-16.98, +18.67]$
	2	403	20.6	12	13	669	$-9.17[-34.10, +15.52]$
	3	362	28.7	17	9	723	$+14.87[-12.61, +42.57]$
	4	347	42.1	28	15	828	$+29.18[-5.66, +66.35]$
	5	494	62.8	44	36	1174	$+4.45[-30.37, +40.62]$
Gemma	1	331	8.2	3	3	489	$-4.89[-20.09, +10.29]$
	2	440	20.5	15	11	762	$+1.47[-21.11, +24.23]$
	3	446	38.1	14	19	794	$-19.15[-44.15, +5.80]$
	4	360	54.4	25	24	943	$-6.65[-45.65, +32.11]$
	5	423	70.4	53	31	1469	$+37.32[-4.46, +79.78]$

A.4 Continuation Policies and Cost Sensitivity

Post-peer policies observe A, B_0 and choose Return-A, aggregation through C_0, or a third response followed by Vote; all pay the $A + B_0$ prefix. Aggregation-minus-EMS3 gains on incomplete or invalidly parsed initial responses are $+.229\,[.089, .369]$ (50 Qwen questions) and $+.109\,[.007, .212]$ (49 Gemma questions). Under valid A/B_0 agreement they are $-.0034\,[-.0090, .0021]$ and $-.0062\,[-.0110, -.0019]$. These prespecified strata are secondary; non-agreement includes invalid responses and differing valid answers.

Under valid A/B_0 agreement, collecting B_1 solely for Vote adds cost without changing the answer. A constrained candidate excludes this action, retains aggregation, and selects its hyperparameters separately; its gains over EMS3 are $+.0076\,[-.0018, .0172]$ and $-.0006\,[-.0090, .0079]$. Post-hoc pruning instead changes only the original policy's agreement-state Vote choices to Return-A. It preserves every success outcome, saves 66.07 and 41.75 output tokens per question, and gains $.000661\,[.000574, .000754]$ and $.000418\,[.000343, .000498]$ utility. Its contrasts with EMS3 still include zero: $+.0083\,[-.0012, .0178]$ and $+.0017\,[-.0027, .0063]$.

The four-action extension adds Self-review and Vote, using the path $A + B_0 + B_1$. Frozen text-feature ridge gains over Stop are $+.0079\,[-.0029, .0187]$ on Qwen and $+.0025\,[-.0063, .0112]$ on Gemma; embedding ridge gives $-.0018\,[-.0122, .0085]$ and $+.0057\,[-.0047, .0161]$. Cost sensitivity separates rescoring frozen decisions at $\lambda \in \{0, .005, .01, .02, .05\}$ from selecting new thresholds on development. At .05, failure recovery gains approximately .0055 and .0034, while always-Coordination differs from Stop by $-.026$ and $-.034$. Output-token utility excludes encoder and scoring-pass overhead.

A.5 Audit Calculation and Verification

The selected primary and text-feature routers are compared with Stop on the same 2,000 questions used for performance estimates, not a separate audit sample. Shared initial cost cancels:

$$D_i = s_{C,i} - s_{A,i} - .01\,(T_{B,i} + T_{C,i})/1000 \quad \text{when Coordination is chosen,}$$

and $D_i = 0$ otherwise. Since each additional call is capped at 2,048 tokens, $D_i \in [-1.04096, 1]$ has width $W = 2.04096$. With sample variance $\widehat{v}$ and $\eta = .025$ per candidate, the empirical-Bernstein calculation (Maurer & Pontil, 2009) is

$$L_{\mathrm{EB}} = \overline{D} - \sqrt{\frac{2\widehat{v}\log(2/\eta)}{n}} - \frac{7W\log(2/\eta)}{3(n-1)}. \tag{7}$$

The two-candidate allowance is .05 per backbone. The secondary Hoeffding bound is $L_{\mathrm{H}} = \overline{D} - W\sqrt{\log(1/\eta)/(2n)}$. Empirical-Bernstein bounds for the primary/text-feature routers are $-.02121, -.01564$ on Qwen and $-.02296, -.01359$ on Gemma; Hoeffding gives $-.05857, -.05834$ and $-.06078, -.05516$. All are negative. The iid formulas approximate category-stratified sampling without replacement; exact finite-sample coverage is unverified. Supplementary independent audits retain their original rules.

Verification corrected the audit support width and a secondary four-action Vote calculation that had duplicated one peer. Frozen parameters and scoring were retained; no candidate was refitted or selected on evaluation outcomes, and primary comparisons are unchanged. A separate sensitivity corrected five completed Qwen responses where a unit letter displaced the explicit answer letter. Fixed-policy success is unchanged; mean EMS3 utility changes by $-.00001024$. Two human reviewers also checked parser extractions for 100 randomly sampled questions per backbone and found no discrepancies in that sample.

B PROOFS

B.1 PROOF OF PROPOSITION 1

Write c_t for the cost incurred between decision events t and $t+1$ (executing a_t and the ensuing generation), so that $C_{t:}(\tau) = \sum_{s=t}^{T-1} c_s$, and let c_0 be the cost of the prefix before the first decision, which does not depend on the coordination policy. For a terminal history h_T with final answer y, set $V^{\pi}_{\lambda,T}(h_T) = U(y,x)$ for every π. Because every component other than the coordinator is fixed, the distribution of (c_t, H_{t+1}) given $(H_t, A_t) = (h,a)$ is the same regardless of which policy chose a, and Equation 2 satisfies

$$Q^{\pi}_{\lambda,t}(h,a) = \mathbb{E}\big[-\lambda c_t + V^{\pi}_{\lambda,t+1}(H_{t+1}) \;\big|\; H_t = h,\, \mathrm{do}(A_t = a)\big]\,. \tag{8}$$

Let $\tau \sim \pi'$ and $t < T$. By Equation 8 and the tower property, $\mathbb{E}_{\pi'}[Q^{\pi}_{\lambda,t}(h_t,a_t)] = \mathbb{E}_{\pi'}[-\lambda c_t + V^{\pi}_{\lambda,t+1}(h_{t+1})]$, where the expectation is over trajectories of π' and a_t is the action π' takes at h_t. Since $A^{\pi}_{\lambda,t}(h_t,a_t) = Q^{\pi}_{\lambda,t}(h_t,a_t) - V^{\pi}_{\lambda,t}(h_t)$, summing over $t = 1,\dots,T-1$ telescopes:

$$\begin{aligned}
\mathbb{E}_{\pi'}\left[\sum_{t<T} A^{\pi}_{\lambda,t}(h_t,a_t)\right] &= \mathbb{E}_{\pi'}\left[-\lambda\sum_{t=1}^{T-1} c_t + \sum_{t=1}^{T-1}\big(V^{\pi}_{\lambda,t+1}(h_{t+1}) - V^{\pi}_{\lambda,t}(h_t)\big)\right] \\
&= \mathbb{E}_{\pi'}\left[-\lambda\sum_{t=1}^{T-1} c_t + U(y,x)\right] - \mathbb{E}\big[V^{\pi}_{\lambda,1}(h_1)\big]\,.
\end{aligned} \tag{9}$$

The horizon is bounded by the hard budget, so all sums are finite. The first term equals $J_\lambda(\pi') + \lambda\,\mathbb{E}[c_0]$, and, because the distribution of h_1 does not depend on the coordination policy, the second equals $J_\lambda(\pi) + \lambda\,\mathbb{E}[c_0]$. Their difference is $J_\lambda(\pi') - J_\lambda(\pi)$. □

B.2 PROOF OF COROLLARY 1

At a history where π' deviates, the summand in Equation 4 is $A^{\pi}_{\lambda,t}(h_t,a_t) > 0$ by construction. At a history where π' acts as π, the action satisfies $a_t \sim \pi(\cdot \mid h_t)$, and the summand has conditional expectation $\mathbb{E}_{A\sim\pi(\cdot|h_t)}[Q^{\pi}_{\lambda,t}(h_t,A)] - V^{\pi}_{\lambda,t}(h_t) = 0$ by the definition of $V^{\pi}_{\lambda,t}$. The expected sum is therefore nonnegative, and strictly positive whenever a deviation occurs with positive probability. □

B.3 PROOF OF PROPOSITION 2

Fix one audited candidate and let $\mu = J_\lambda(\pi_{\mathrm{cand}}) - J_\lambda(\pi_k)$ on the audit distribution, so that $\mathbb{E}[D_i] = \mu$. Because the candidate is frozen before its audit and the audit items are not used in its construction, $D_1,\dots,D_n$ are independent and identically distributed on $[-R_\lambda, R_\lambda]$ by assumption. Hoeffding's inequality (Hoeffding, 1963) gives, for $\varepsilon > 0$,

$$\Pr\big(\overline{D} - \mu \ge \varepsilon\big) \le \exp\left(-\frac{2n\varepsilon^2}{(2R_\lambda)^2}\right) = \exp\left(-\frac{n\varepsilon^2}{2R_\lambda^2}\right). \tag{10}$$

Choosing $\varepsilon = R_\lambda\sqrt{2\log(K/\delta)/n}$ makes the right-hand side equal to δ/K, so $\Pr(\mathrm{LCB}_{\mathrm{audit}} > \mu) \le \delta/K$. A candidate with $\mu < 0$ can be accepted only if $\mathrm{LCB}_{\mathrm{audit}} \ge 0 > \mu$, an event of probability at most δ/K. A union bound over at most K audited candidates bounds the probability that any accepted candidate has $\mu < 0$ by δ. The strict rule $\mathrm{LCB}_{\mathrm{audit}} > 0$ used for the LM audits accepts a subset of what the nonnegative rule accepts, so the same bound applies. □

C Simulation Studies

C.1 Replay Coupling and Estimator Variance

The reference and treatment utilities are Bernoulli with $p_0 = .60$ and $p_1 = .65$, so $\Delta_U = .05$. We compare shared-uniform, independent, and antithetic couplings. At $M = 32$ and 100,000 estimates per setting, the respective means are .04989, .04948, and .05057; root-mean-square errors are .0384, .1213, and .1532; paired-harm rates $\Pr(Z_U < 0)$ are 0%, 21%, and 35%. Coupling changes variance and paired harm, not the estimand. Across $M \in \{1, 2, 4, 8, 32, 128\}$, shared-uniform root-mean-square error falls from .216 to .019. These simulations do not estimate LM replay noise.

C.2 Replay-Label and Conditional-Mean Coverage

Let $Z \in \{-1, 1\}$ with $\mathbb{E}[Z] = .98$; each replay label averages M independent draws. We calibrate a one-sided 90% split-conformal endpoint around prediction 1 on 200 items and repeat the experiment 5,000 times. At $M = 4$, coverage is 95.90% for a new replay label but 0.02% for the conditional mean .98: the residual quantile is zero when more than 90% of calibration labels equal 1. Mean coverage at $M = 1$, 8, and 32 is 0%, 14%, and 100% (Figure 6, left). Replay-label coverage does not imply mean coverage; the full-policy audit requires neither.

C.3 Coverage under Adaptive Selection

With 5,000 calibration and 100,000 test items, a utility endpoint covers 97.96% of test labels. Selecting items with positive endpoints retains 2,036 items whose true effect is -1.02, with zero coverage (Figure 6, middle). Marginal coverage need not hold after selection.

C.4 Audit Error Control, Power, and Pseudoreplication

Each audit family has $K = 3$ samples of n independent items, with $R_\lambda = 1.06$, $\delta = .05$, and acceptance $\mathrm{LCB}_{\mathrm{audit}} \geq 0$. We simulate 20,000 families per setting. Table 5 compares Hoeffding and empirical-Bernstein acceptance (Maurer & Pontil, 2009); Figure 5 gives the full grid. Exact binomial tails confirm nonpositive-effect acceptance below δ. Equating the mean to the Hoeffding penalty, effects of $+0.005$, $+0.01$, $+0.02$, $+0.05$, and $+0.10$ require 368,033, 92,009, 23,003, 3,681, and 921 items at $R_\lambda = 1.06$, or 344,533, 86,134, 21,534, 3,446, and 862 at $R_\lambda = 1.0256$. For a harmful effect of -0.02 on 180 items, false acceptance is 0.185% with item clustering but 52.36% when 32 replays per item are counted as 5,760 observations (Figure 6, right). Declaring $K = 3$ looks while performing 30, 300, or 3,000 raises false acceptance from 0.23% to 2.3%, 20%, and 90%.

Table 5: Audit simulations: fraction of families (of 20,000) accepting at least one of $K = 3$ looks, with 95% Wilson Monte Carlo (MC) intervals, and the corresponding empirical-Bernstein (EB) rate. $R_\lambda = 1.06$, $\delta = .05$. A true effect of zero is the boundary of non-degradation.

True ΔJ_λ	n per look	Hoeffding acceptance (%)	95% MC interval	EB acceptance (%)
-0.02	180	0.170	[0.122, 0.237]	0.010
-0.02	1,000	0.110	[0.073, 0.167]	0.005
-0.02	20,000	0.000	[0.000, 0.019]	0.000
0.00	180	0.565	[0.470, 0.679]	0.000
0.00	1,000	0.610	[0.511, 0.728]	0.020
0.00	20,000	0.615	[0.516, 0.733]	0.150
$+0.02$	180	1.245	[1.100, 1.408]	0.005
$+0.02$	1,000	3.330	[3.090, 3.588]	0.315
$+0.02$	5,000	17.535	[17.014, 18.068]	5.865
$+0.02$	20,000	80.460	[79.905, 81.004]	62.945
$+0.05$	1,000	22.875	[22.298, 23.462]	3.460
$+0.10$	1,000	90.440	[90.025, 90.840]	52.135

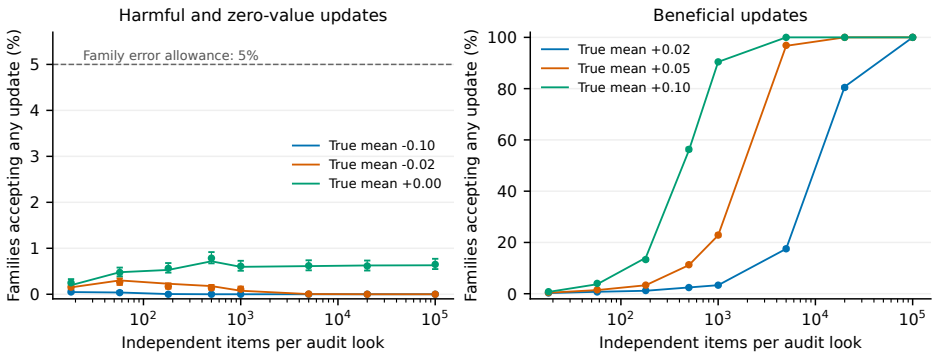


Figure 5: Simulation results for the audit rule. Points are simulated family acceptance rates (20,000 families per setting, three looks of n items each, $R_\lambda = 1.06$, $\delta = .05$), bars are 95% Wilson intervals, and lines are exact binomial-tail probabilities. Left: harmful and zero-effect updates stay below the 5% allowance. Right: beneficial updates are certified only at sample sizes of order 10^4.

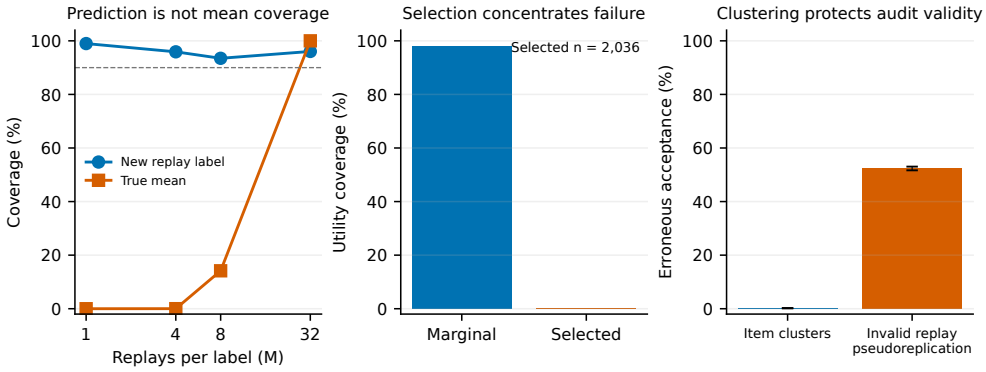


Figure 6: Simulation diagnostics of coverage and audit error. Left: nominal 90% coverage of finite-M replay labels versus coverage of the true conditional mean. Middle: 97.96% marginal utility coverage coexists with zero coverage on 2,036 selected items. Right: false acceptance of a harmful update at $n = 180$ with correct item clustering versus counting 32 replays per item as independent.

D Actions, Diagnostics, and Assumptions

D.1 Actions and References

Table 6 lists the candidate–reference pairs considered in this paper. Generation cost is attributed to the candidate only when generation is causally downstream of the decision.

Table 6: Coordination actions and their references.

Decision time	Candidate a	Reference a^0
Before generation	Solicit and reveal a peer response	Skip the peer response
After private generation	Reveal the available response	Withhold it; generation cost is sunk
Before critique	Request and expose a critique	Skip the critique
After a debate round	Continue for one round	Aggregate or stop now

D.2 Diagnostics

Three post-hoc diagnostics use item-clustered intervals: *harmful persuasion*, the fraction of answer-changing pairs with lower treatment utility; *high-influence negative value*, the fraction of answer-changing events with negative CVC; and *correct-minority suppression*, where a private agent and the reference are correct but the treatment is wrong. They depend on the declared replay coupling or deterministic decoding and are not online features or coupling-invariant causal parameters. Examples appear in Appendices E.1 and F.

D.3 Identification Assumptions

Equation 3 assumes well-defined feasible interventions, consistent isolated branches, fixed system components and continuation, and no interference between tasks. Generalization requires coverage and exchangeability across replay, calibration, audit, and deployment histories. All events from a question stay in one split and one uncertainty-estimation cluster. CVC describes a specified software intervention; it neither identifies a stronger semantic cause nor transfers unchanged to different models, tools, references, or continuations.

E Supplementary GSM8K Experiments

Common protocol. GSM8K uses Qwen3-8B in bfloat16 (BF16), with thinking disabled and a 1,024-token cap per role. Initial and peer answers use temperature 0.7 and top-$p = 0.95$; self-review and aggregation are greedy. All GSM8K partitions come from the original training split, not the official test set. Routing reuses 850 training questions from the development study; development, calibration, evaluation, and audit partitions are mutually disjoint and exclude training. New questions were selected deterministically after excluding previously used groups. Success requires normal termination and a frozen normalized-parser match to the unchanged reference. Strict extraction, formatting, parse failures, and truncation are recorded without repair; MMLU-Pro uses the answer-line sensitivity described in Appendix F. Online paths are A (one call), $A{+}S$ (two), $A{+}B{+}C$ (three), and $A{+}B_0{+}B_1$ (three). Offline totals count unique calls, not per-policy cost: GSM8K has 29,600 calls (1,200 shared-history, 12,000 development, and 16,400 routing/two-node), with 5,104,900 input tokens, 2,578,415 output tokens, and 69,169 summed generation seconds for the two-node experiment. Supplementary MMLU-Pro adds 17,820 calls per backbone on 2,970 shared questions, for campaign totals of 83,060 calls and 6,270 questions. These are not independent evaluation sample sizes. MMLU-Pro uses a 2,048-token cap. Input tokens and per-call generation times are separate accounting measures, not end-to-end latency.

Estimator objective. With $\gamma, \eta \geq 0$ weighting the cost and harm heads, the general estimator minimizes

$$\mathcal{L}(\theta) = \ell(\mu_U, \widehat{\Delta}_U) + \gamma\, \ell(\mu_C, \widehat{\Delta}_C) + \frac{\eta}{M} \sum_{m=1}^{M} \mathrm{BCE}(p_{\mathrm{harm}}, \mathbf{1}\{Z_{U,m} < 0\}) \tag{11}$$

where BCE is binary cross-entropy, ℓ is squared or Huber loss, and $Z_{U,m}$ is the paired utility difference. The supplementary LM fits use squared-error ridge heads and $\eta = 0$ (no harm head).

E.1 GSM8K: Complete Actions at a Shared History

The shared-history study collects 200 A, 200 S, 400 B, and 400 C generations (Table 7). Each S is reused across paired references. Of the 3.0-point $C - S$ gain, 2.0 points come from four self-review outputs accepted by strict extraction but rejected by normalized parsing; 0.5 points come from preserving initial answers changed by self-review, and 0.5 from additional matches on two ambiguous questions. Strict-only success is 91.5% for S and 92.5% for C. Among 400 A/B branches, C preserves all ten correct-A/incorrect-B cases and all 358 both-correct cases. On a disjoint 100-question sample, $C - S$ is -0.75 points ($[-1.75, 0]$); a symmetric reference-label sensitivity leaves the three-arm contrasts unchanged. Against Vote, C has five favorable and three adverse transitions across five questions.

Table 7: Four complete actions at a shared initial answer on 100 GSM8K questions. Top: tokens and generation time sum the complete online path, including A. Bottom: paired differences with 95% intervals from 10,000 question-bootstrap resamples ($n = 100$); $C - V$ is the prespecified primary contrast.

Action	Success (%)	Input tokens	Output tokens	Gen. time (s)	J_λ
Stop (A)	92.00	156.32	162.67	8.25	0.918373
Self-review (S)	89.50	501.32	325.28	16.44	0.891747
Coordination (C)	92.50	821.09	480.94	24.32	0.920191
Vote (V)	92.00	468.96	481.58	24.39	0.915184
Contrast	**Success difference (pp)**		**ΔJ_λ**		
$C - V$	$+0.50$ $[-1.25, +2.50]$		$+0.005006$ $[-0.012535, +0.025009]$		
$C - S$	$+3.00$ $[+0.75, +6.00]$		$+0.028443$ $[+0.005917, +0.058444]$		
$C - A$	$+0.50$ $[0.00, +1.25]$		$+0.001817$ $[-0.003249, +0.009330]$		
$S - A$	-2.50 $[-5.50, -0.50]$		-0.026626 $[-0.056676, -0.006531]$		
$V - A$	0.00 $[-2.00, +2.00]$		-0.003189 $[-0.023315, +0.016853]$		

E.2 GSM8K: Learned Routing

Features and fitting. The 55-dimensional design contains an intercept, up to 1,024 training-only word/bigram term frequency–inverse document frequency (TF–IDF) terms projected through a fixed 48-dimensional Rademacher map and $\tanh$, and six standardized response properties: input/output length, termination, and normalized-parser, strict-parser, and format validity. Ridge heads minimize summed squared error plus α times squared non-intercept coefficients, using five shared question-bootstrap resamples. Fit seeds 6101–6105 repeat this procedure; 6101 is primary. Value-router variants select among $\alpha \in \{.01, .1, 1\}$ and thresholds $\{0, .01, .03\}$; SD-margin gates use the same α grid and $\beta \in \{.5, 1, 2\}$. The error trigger uses a ridge Brier objective, clipped predictions, thresholds $\{.05, .15, .3\}$, and fixed Coordination. Development J_λ selects each candidate, with lower-cost tie-breaking. Routing takes about 9.5×10^{-5} CPU seconds per question, excluding loading and generation.

Calibration and audit. After fitting and development selection, the conformal gate uses 200 calibration questions. Each residual is maximized over three non-Stop actions; the order statistic is $\lceil (n+1)(1-\alpha) \rceil$, with $\alpha_U = \alpha_C = .05$. On 500 evaluation questions, utility/cost coverage is 98.6%/97.9% marginally and 96.0%/95.8% simultaneously across actions; joint coverage is 92.8%. The primary fit selects four questions with zero joint coverage. Across five fits, selected counts are 4, 2, 8, 4, and 2, all with zero coverage. The 20 pooled question–fit records can repeat questions and do not establish a population selected-coverage rate. Frequency controls match development action rates, not evaluation rates (49.5% versus 50.0% for the SD-margin comparison). Three frozen single-node candidates share a 500-question audit with $R_\lambda = 1.02048$, $K = 3$, $\delta = .025$, and acceptance $\mathrm{LCB}_{\mathrm{audit}} > 0$. The two-node family uses a disjoint 100-question audit with the same K, δ; total error allowance is .05.

Table 8: Learned routing on 500 independent GSM8K questions (Qwen3-8B; training size 850; fit seed 6101; hyperparameters frozen on development). Active is the fraction of questions on which the policy does not stop at A. ΔJ_λ is the paired difference from Stop in units of 10^{-3} with a 95% interval from 10,000 shared question-bootstrap resamples. The value router's direct, paired-difference, and decomposed variants make identical decisions.

Policy	Success (%)	J_λ	Output tokens	Active (%)	ΔJ_λ vs. Stop ($\times 10^{-3}$)
Stop (answer alone)	94.20	0.9404	158.7	0.0	—
Self-review	94.40	0.9408	321.1	100.0	$+0.38$ $[-11.61, +12.39]$
Coordination	94.45	0.9397	478.2	100.0	-0.70 $[-8.68, +8.35]$
Vote	94.70	0.9422	480.2	100.0	$+1.78$ $[-5.28, +9.74]$
Value router (direct = paired = decomposed)	94.70	0.9435	354.6	72.8	$+3.04$ $[-2.99, +10.95]$
Value router, rate-matched	94.70	0.9441	291.2	52.6	$+3.68$ $[-2.35, +11.58]$
CVC gate, ensemble-SD margin	94.35	0.9406	290.1	50.0	$+0.19$ $[-3.27, +3.76]$
CVC gate, conformal endpoints	94.20	0.9404	159.7	0.8	-0.010 $[-0.020, -0.002]$
Error trigger	94.05	0.9387	183.1	5.8	-1.74 $[-4.30, -0.18]$
Random, matched to the SD-margin rate	94.38	0.9409	291.5	49.5	$+0.47$ $[-2.88, +4.22]$
Random, matched to the conformal rate	94.20	0.9404	160.3	0.5	-0.003 $[-0.043, +0.042]$

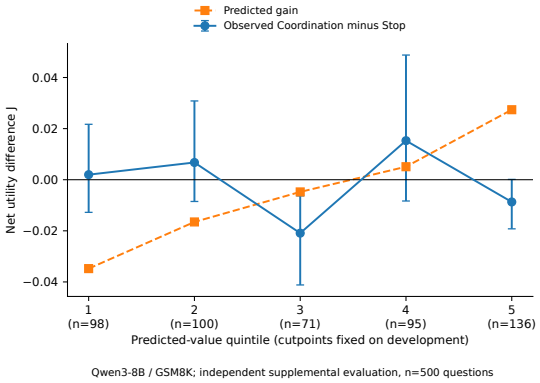


Figure 7: GSM8K routing: predicted versus observed Coordination-minus-Stop value by quintile of the fitted pre-decision value (500 evaluation questions; 95% bootstrap intervals). The observed gains are $+2.0$, $+6.7$, -20.9, $+15.3$, and -8.7×10^{-3} across the five groups; the highest-scoring group ($n = 136$, predicted $+27.4 \times 10^{-3}$) has observed value -8.7×10^{-3} with interval $[-19.2, +0.1]$, and only group 3 excludes zero, in the harmful direction.

Table 9: Calibration and audit outcomes. Top: joint utility–cost coverage of replay labels over all evaluated items versus over the items the conformal gate selected, for the five LM fit seeds, the classifier proxy (BC denotes Breast Cancer), and an analytical counterexample. GSM8K counts selections per fit on the same 500 questions; its pooled 0/20 denotes question–fit records, not independent questions. Bottom: every frozen audit, with the one-sided Hoeffding lower bound of Equation 6; LM audits use $\delta = .025$ per family and the strict rule $\mathrm{LCB} > 0$, the proxy audits use $\delta = .05$ and $\mathrm{LCB} \geq 0$.

Selected-event coverage Setting	Gate	Marginal joint coverage	Events selected	Selected coverage
GSM8K routing, five fit seeds	conformal endpoints	92.4–94.2%	4, 2, 8, 4, 2	0/20
Classifier proxy (BC / Digits / Wine)	conformal endpoints	90.4 / 87.8 / 100%	2 / 38 / 0	0/2, 21/38, —
Analytical counterexample	conformal endpoints	97.96%	2,036 of 100,000	0/2,036
Frozen audits Family / incumbent	Candidate	n	Mean ΔJ_λ	$\mathrm{LCB}_{\mathrm{audit}}$
LM single node / Stop	value router	500	-0.0034	-0.1446
LM single node / Stop	CVC gate, SD margin	500	-0.0017	-0.1429
LM single node / Stop	CVC gate, conformal	500	-0.0000	-0.1412
LM two-node / Review-then-Stop	CVC, both nodes	100	$+0.0214$	-0.2975
LM two-node / Review-then-Stop	CVC, node 1 only	100	$+0.0114$	-0.3076
LM two-node / Review-then-Stop	CVC, node 2 only	100	$+0.0198$	-0.2991
Proxy / always-coordinate	CVC gate, conformal	57 / 180 / 18	$+0.024$ / -0.031 / -0.107	-0.378 / -0.257 / -0.822

Learning curves and replay counts. Figure 8 compares nested training sizes (100, 250, 500, 850) and replay counts on the same 500 evaluation questions. SD-margin mean J_λ is .941718, .941660, .944253, and .943442, with fit-seed SD .001551, .000890, .002885, and .002684; neither this curve nor the value-router/error-trigger curves is monotone. At $M = 1, 2, 4$, Stop-versus-Coordination utility is .939018, .939910, and .939914; four-action utility at $M = 2, 4$ is .943214 and .943442. Vote requires two peers, so $M = 1$ is undefined for the four-action menu. These analyses reuse collected labels. Figure 9 shows the frozen threshold grid; no operating point was selected on evaluation.

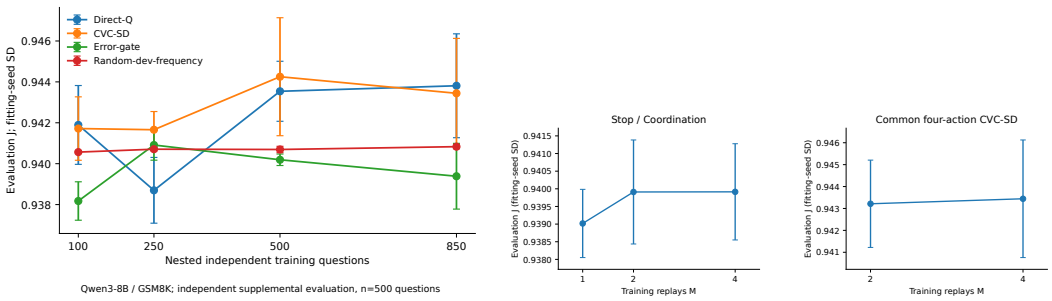


Figure 8: GSM8K routing efficiency diagnostics on the 500 evaluation questions. Left: five-seed mean J_λ against nested training size; bars are fitting-seed standard deviations, not confidence intervals. Right: J_λ against the number of coordination replays per training label.

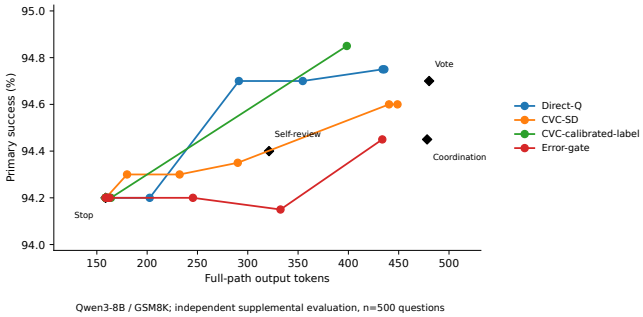


Figure 9: GSM8K routing: success against complete-path output tokens on the 500 evaluation questions for each router's frozen threshold grid, with the fixed actions marked.

E.3 GSM8K: Development Fit

This fit uses 1,000 training and 200 development questions collected before routing: 12,000 calls comprising 1,200 A, 1,200 S, 4,800 B, and 4,800 C. The design uses a 2,048-term vocabulary, a 64-dimensional projection, $\alpha \in \{.001, .01, .1\}$, and $\beta \in \{.5, 1, 2\}$. Development selects $\alpha = .01$, $\beta = 1$ for the CVC gate. Table 10 gives the primary fit (1,000 training questions, seed 6101); Figure 10 gives five-fit curves. Primary gains versus Stop, Vote, and the value router are respectively $+0.50$, $+0.25$, and $+0.25$ success points, and $+3.71$, $+4.27$, and $+3.26$ in 10^{-3} utility units; all corresponding intervals include zero. Averaging each question over fits before resampling gives CVC-minus-value-router utility $+3.13 \times 10^{-3}$ $[+0.16, +6.63]$ and success $+0.225$ $[-0.075, +0.575]$ points. These are development diagnostics. All 60 matched-α checks give identical value-router and $\beta = 0$ decisions; direct-J_λ and decomposed CVC outcomes also match. The primary gain over Stop comes from three questions routed to Vote, for a net one success-equivalent question in 200. Of six initial failures, three have a successful peer.

Table 10: GSM8K development fit on 200 development questions (training size 1,000, fit seed 6101, hyperparameters selected on the same development questions). Values are development diagnostics, not independent estimates.

Policy	Success (%)	J_λ	Output tokens	Calls
Stop	97.000	0.968471	152.9	1.000
Self-review	96.000	0.956882	311.8	2.000
Coordination	96.500	0.960363	463.7	3.000
Vote	97.250	0.967910	459.0	3.000
Value router	97.250	0.968926	357.4	2.290
CVC gate, SD margin ($\beta = 1$)	97.500	0.972182	281.8	1.795
CVC gate, $\beta = 0$ (= value router)	97.250	0.968926	357.4	2.290
CVC gate, no cost head	97.625	0.973369	288.1	1.835
CVC gate, direct J_λ	97.500	0.972182	281.8	1.795
Error trigger	97.000	0.968471	152.9	1.000
Random, matched frequency	96.875	0.966045	270.5	1.795

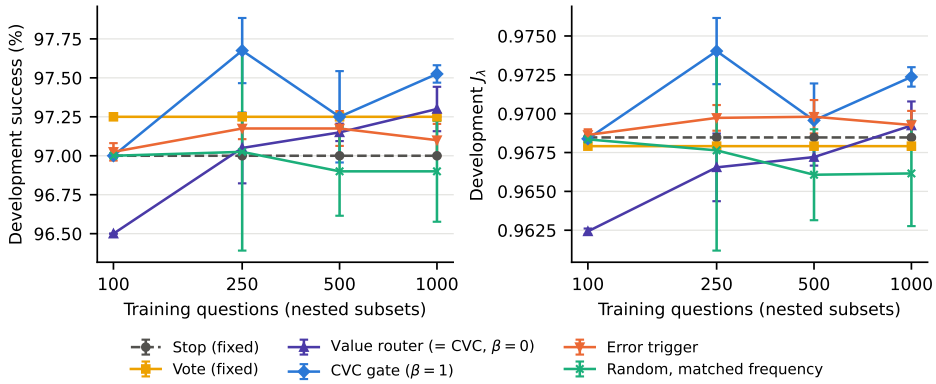


Figure 10: GSM8K development fit: five-fit-seed mean success and J_λ on the 200 development questions against nested training size. Bars are fitting-seed standard deviations on one LM collection; the development questions also selected hyperparameters.

E.4 GSM8K: Two-Node Routing

The call tree is $A, S(q, A), B(q), D(q, S, B)$. Node 1 chooses Stop_A or review using q, A and a frozen node-2 continuation; node 2 chooses Stop_S or D using q, A, S. Decisions are recorded before the relevant generations, and the full tree is collected for policy replay. Node 2 is fitted first; node 1 targets use its in-sample predictions without cross-fitting. Both use the ridge design with $\alpha = .1$. Table 11 and Figure 11 report 200 evaluation questions.

Table 11: GSM8K two-node routing on 200 evaluation questions (fit seed 6101). Node-2 reach is the fraction of questions on which a policy reviews; D is the fraction on which it coordinates after reviewing. ΔJ_λ is the paired difference from Stop_A in units of 10^{-3} with 95% question-bootstrap intervals.

Policy	Success (%)	J_λ	Output tokens	Node-2 reach (%)	D (%)	ΔJ_λ vs. Stop_A ($\times 10^{-3}$)
Stop_A	94.50	0.9434	157.4	0	0	—
Review, then Stop_S	94.50	0.9418	318.9	100	0	-1.62 [-21.62, $+18.38$]
Review, then D	96.00	0.9539	613.6	100	100	$+10.44$ [-9.60, $+35.31$]
Value router, both nodes	95.00	0.9475	253.4	33.5	11.5	$+4.04$ [-1.11, $+14.15$]
CVC gate, node 1 only	94.50	0.9433	172.3	12.0	0	-0.15 [-0.21, -0.09]
CVC gate, node 2 only	95.50	0.9517	325.5	100	2.0	$+8.32$ [-1.72, $+23.33$]
CVC gate, both nodes	94.50	0.9434	158.9	1.0	0.5	-0.015 [-0.043, 0.000]

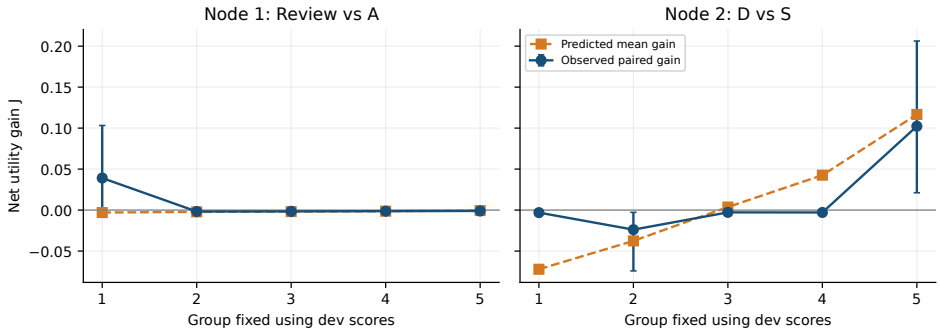


Figure 11: GSM8K two-node routing: predicted and observed paired value at each node on the 200 evaluation questions, grouped by development-fixed scores. Node 1 compares reviewing under the frozen node-2 continuation with Stop_A; node 2 compares D with Stop_S.

F Supplementary MMLU-Pro Experiments

R1 and R2 denote the Stop-reference and Vote-reference candidate families, respectively.

Data collection and partitions. All backbones share a category-stratified question list (seed 20260916) with disjoint splits: 20 smoke, 200 pilot, 850 training, 200 development, 200 calibration, 500 evaluation, and 1,000 audit questions. Each question contributes A, S, B_0, B_1, C_0, C_1, totaling 17,820 calls per backbone. Settings follow Appendix E; peers use separate seed streams, and C_i sees only (q, A, B_i). Six primary candidates were frozen and hashed before opening evaluation and audit, each once. Shared questions use the same 10,000 bootstrap indices. Fit seeds 6101–6105 are refits, not independent experiments.

Pilot amendment. The 200-question pilot checked initial success, complementary opportunity (at least 5%), parsing, formatting, truncation, and budget. The original format gate (c) required under 5% final-line failures per role. Two independent pilots had under 2% normalized parse failure but 8–15% format failure on sampled roles. Because scoring does not require final-line compliance, the gate was amended to under 5% normalized parse failure, retaining format failure as a diagnostic. This amendment occurred after the original pilot decision but was recorded and hashed before collection release.

Excluded collections. Excluded collections remain documented: an initial Qwen run failed strict parsing ($9/20$ versus a $17/20$ requirement); a later Qwen run failed gate (c) and budget; Gemma failed budget only; and a Qwen3-4B GSM8K pilot had 4.5% complementary opportunity, below the 5% threshold. Phi-4 stopped after smoke/pilot collection because of evaluator serialization defects: 1,320 verified calls (120 on 20 smoke questions, 1,200 on 200 pilot questions), with no verified efficacy, router, or audit. A heterogeneous Llama-3.1-8B peer pilot failed format, truncation, and budget checks. Its GSM8K fallback uses model-specific initial answers on previously examined questions and is exploratory only.

Mistral-Nemo. This exploratory run uses the Tekken tokenizer and checkpoint `mistralai/Mistral-Nemo-Instruct-2407` (revision `04d8a905`). It shares the question list, settings, scoring, and bootstrap indices; because it lacks a system role, instructions are folded into the first user turn. A prespecified collection-only protocol collected all 2,970 trees regardless of pilot outcomes. The retrospective pilot failed the initial-success floor (0.39 versus 0.40), passed complementary opportunity (0.175) and amended parsing (maximum 4.0%, role B_0), but would have failed the original format gate (6.0% for B_0). Other engineering checks passed. Results are therefore exploratory. On evaluation, $93/500$ questions (18.6%) have a complementary peer answer; this question-level rate is not comparable to the GSM8K branch-level rate. Coordination repairs 7.5% of initial answers and harms 5.4%. Of repairs, 81% change the categorical answer and 15% repair parsing/formatting; 96% of harms are categorical. On audit, Stop scores 42.9%

($J_\lambda = 0.4262$), Vote 46.3% (0.4546), and EMS3 46.3% (0.4556); all six candidate audits decline (Table 18). Post-hoc oracle and ranking analyses were not run for this backbone.

Table 12: **Qwen3-8B: policy outcomes.** MMLU-Pro ($n = 500$, fit 6101). Success (%), calls per question, and ΔJ_λ against the stated reference (10^{-3}; 95% paired bootstrap intervals). †: development-selected; Random matches its development action frequencies.

Policy	Reference	Success	Calls	ΔJ_λ [95% interval]
Stop	Stop	67.8	1.00	$0.0\,[0.0, 0.0]$
Self-review	Stop	63.0	2.00	$-51.3\,[-81.3, -21.3]$
Coordination	Stop	68.0	3.00	$-5.9\,[-25.1, +13.2]$
Vote	Stop	69.4	3.00	$+5.6\,[-15.7, +27.3]$
EMS3	Stop	69.4	2.24	$+8.8\,[-12.5, +30.5]$
$R1$ Value router	Stop	68.4	2.04	$+1.6\,[-16.3, +19.6]$
$R1$ CVC, SD margin†	Stop	68.1	1.79	$-0.4\,[-15.6, +15.0]$
$R1$ CVC, conformal	Stop	67.8	1.00	$0.0\,[0.0, 0.0]$
$R1$ Error trigger (control)	Stop	67.6	2.35	$-7.5\,[-24.5, +9.1]$
$R2$ Value router	Vote	67.9	2.35	$-10.0\,[-30.2, +9.8]$
$R2$ CVC, SD margin	Vote	68.7	2.50	$-3.1\,[-22.4, +15.7]$
$R2$ CVC, conformal	Vote	69.4	3.00	$0.0\,[0.0, 0.0]$
$R2$ Error trigger (control)	Vote	68.0	1.09	$-4.6\,[-24.7, +15.4]$
Random (selected rate)	Stop	67.6	1.92	$-6.3\,[-15.3, +2.7]$

Table 13: **Gemma-3-12B-it: policy outcomes.** MMLU-Pro ($n = 500$, fit 6101). Success (%), calls per question, and ΔJ_λ against the stated reference (10^{-3}; 95% paired bootstrap intervals). †: development-selected; Random matches its development action frequencies.

Policy	Reference	Success	Calls	ΔJ_λ [95% interval]
Stop	Stop	65.8	1.00	$0.0\,[0.0, 0.0]$
Self-review	Stop	66.0	2.00	$-1.3\,[-23.2, +19.0]$
Coordination	Stop	67.5	3.00	$+8.1\,[-15.1, +31.7]$
Vote	Stop	68.0	3.00	$+9.9\,[-12.7, +33.8]$
EMS3	Stop	68.0	2.28	$+13.5\,[-9.2, +37.3]$
$R1$ Value router	Stop	67.5	2.23	$+10.7\,[-8.8, +30.6]$
$R1$ CVC, SD margin	Stop	66.8	2.02	$+4.6\,[-14.2, +23.0]$
$R1$ CVC, conformal	Stop	65.8	1.00	$0.0\,[0.0, 0.0]$
$R1$ Error trigger (control)	Stop	66.3	1.04	$+4.6\,[-0.3, +11.4]$
$R2$ Value router	Vote	67.3	2.33	$-1.6\,[-23.8, +20.0]$
$R2$ CVC, SD margin†	Vote	66.6	2.49	$-10.0\,[-29.9, +9.1]$
$R2$ CVC, conformal	Vote	68.0	3.00	$0.0\,[0.0, 0.0]$
$R2$ Error trigger (control)	Vote	66.0	1.04	$-8.5\,[-32.0, +14.0]$
Random (selected rate)	Vote	67.1	2.46	$-4.2\,[-15.0, +6.4]$

Early-stopping contrasts. EMS3 preserves Vote's answer while reducing token cost. Utility gains over Vote are $+3.3\,[3.0, 3.6]$, $+3.5\,[3.2, 3.9]$, and $+0.9\,[0.8, 1.1]$ in 10^{-3} units on Qwen, Gemma, and Mistral-Nemo. Mistral-Nemo gains over Stop are $+25.5\,[1.5, 51.3] \times 10^{-3}$ on evaluation and $+29.3\,[11.2, 47.5] \times 10^{-3}$ on audit. These fixed-policy bootstrap contrasts do not constitute audit acceptance.

Scoring-rule dependence. The normalized parser takes the last standalone uppercase `A`–`J` after the first case-insensitive whole word `answer`, without rejecting conflicting letters. Strict extraction takes the last case-sensitive `Answer: X` line, allowing the specified whitespace and parentheses, wherever it appears. Strict success requires normal completion and a reference match, not normalized agreement; final-line compliance and unique answer markers are separate diagnostics. Qwen Stop scores 67.8% normalized versus 35.0% strict. Coordination-minus-Stop changes from $+0.2\,[-1.7, +2.1]$ to $+26.6\,[+22.6, +30.6]$ success points. Gemma Stop scores 65.8% versus

Table 14: **Mistral-Nemo: policy outcomes.** MMLU-Pro ($n = 500$, fit 6101). Success (%), calls per question, and ΔJ_λ against the stated reference (10^{-3}; 95% paired bootstrap intervals). †: development-selected; Random matches its development action frequencies. Mistral-Nemo is a collection-only probe.

Policy	Reference	Success	Calls	ΔJ_λ [95% interval]
Stop	Stop	41.0	1.00	$0.0\,[0.0, 0.0]$
Self-review	Stop	42.0	2.00	$+8.1\,[-8.1, +26.0]$
Coordination	Stop	43.1	3.00	$+17.1\,[-9.0, +43.2]$
Vote	Stop	44.0	3.00	$+24.6\,[+0.6, +50.4]$
EMS3	Stop	44.0	2.51	$+25.5\,[+1.5, +51.3]$
$R1$ Value router†	Stop	41.6	2.13	$+3.3\,[-15.4, +22.4]$
$R1$ CVC, SD margin	Stop	41.2	1.85	$0.0\,[-15.5, +15.9]$
$R1$ CVC, conformal	Stop	41.0	1.00	$0.0\,[0.0, 0.0]$
$R1$ Error trigger (control)	Stop	43.0	2.35	$+17.0\,[-5.4, +39.0]$
$R2$ Value router	Vote	42.1	2.43	$-16.9\,[-39.1, +5.4]$
$R2$ CVC, SD margin	Vote	43.0	2.65	$-8.6\,[-27.6, +9.5]$
$R2$ CVC, conformal	Vote	44.0	3.00	$0.0\,[0.0, 0.0]$
$R2$ Error trigger (control)	Vote	43.4	2.35	$-4.9\,[-18.8, +7.2]$
Random (selected rate)	Stop	42.4	2.15	$+11.6\,[-0.2, +23.6]$

Table 15: **Qwen3-8B: complete-path inference costs.** Means on 500 evaluation questions, primary fit. ΔJ_λ is relative to Stop ($\times 10^{-3}$; 95% paired bootstrap intervals). Model-native tokens are comparable only within a backbone.

Policy	Calls	Output tokens	ΔJ_λ vs. Stop
Stop	1.00	520	$0.0\,[0.0, 0.0]$
Self-review	2.00	851	$-51.3\,[-81.3, -21.3]$
Coordination	3.00	1314	$-5.9\,[-25.1, +13.2]$
Vote	3.00	1565	$+5.6\,[-15.7, +27.3]$
EMS3	2.24	1235	$+8.8\,[-12.5, +30.5]$
$R1$ Value router	2.04	958	$+1.6\,[-16.3, +19.6]$
$R1$ CVC, SD margin	1.79	855	$-0.4\,[-15.6, +15.0]$
$R1$ CVC, conformal	1.00	520	$0.0\,[0.0, 0.0]$
$R1$ Error trigger (control)	2.35	1072	$-7.5\,[-24.5, +9.1]$
$R2$ Value router	2.35	1069	$-4.5\,[-23.5, +14.0]$
$R2$ CVC, SD margin	2.50	1175	$+2.5\,[-16.8, +21.6]$
$R2$ CVC, conformal	3.00	1565	$+5.6\,[-15.7, +27.3]$
$R2$ Error trigger (control)	1.09	623	$+1.0\,[-5.2, +8.4]$

Table 16: **Gemma-3-12B-it: complete-path inference costs.** Means on 500 evaluation questions, primary fit. ΔJ_λ is relative to Stop ($\times 10^{-3}$; 95% paired bootstrap intervals). Model-native tokens are comparable only within a backbone.

Policy	Calls	Output tokens	ΔJ_λ vs. Stop
Stop	1.00	589	$0.0\,[0.0, 0.0]$
Self-review	2.00	922	$-1.3\,[-23.2, +19.0]$
Coordination	3.00	1478	$+8.1\,[-15.1, +31.7]$
Vote	3.00	1797	$+9.9\,[-12.7, +33.8]$
EMS3	2.28	1442	$+13.5\,[-9.2, +37.3]$
$R1$ Value router	2.23	1221	$+10.7\,[-8.8, +30.6]$
$R1$ CVC, SD margin	2.02	1129	$+4.6\,[-14.2, +23.0]$
$R1$ CVC, conformal	1.00	589	$0.0\,[0.0, 0.0]$
$R1$ Error trigger (control)	1.04	630	$+4.6\,[-0.3, +11.4]$
$R2$ Value router	2.33	1254	$+8.3\,[-12.0, +28.5]$
$R2$ CVC, SD margin	2.49	1400	$-0.1\,[-21.2, +21.5]$
$R2$ CVC, conformal	3.00	1797	$+9.9\,[-12.7, +33.8]$
$R2$ Error trigger (control)	1.04	642	$+1.5\,[-0.7, +5.6]$

Table 17: **Mistral-Nemo: complete-path inference costs.** Means on 500 evaluation questions, primary fit. ΔJ_λ is relative to Stop ($\times 10^{-3}$; 95% paired bootstrap intervals). Model-native tokens are comparable only within a backbone.

Policy	Calls	Output tokens	ΔJ_λ vs. Stop
Stop	1.00	276	$0.0\,[0.0, 0.0]$
Self-review	2.00	466	$+8.1\,[-8.1, +26.0]$
Coordination	3.00	665	$+17.1\,[-9.0, +43.2]$
Vote	3.00	814	$+24.6\,[+0.6, +50.4]$
EMS3	2.51	722	$+25.5\,[+1.5, +51.3]$
$R1$ Value router	2.13	544	$+3.3\,[-15.4, +22.4]$
$R1$ CVC, SD margin	1.85	479	$0.0\,[-15.5, +15.9]$
$R1$ CVC, conformal	1.00	276	$0.0\,[0.0, 0.0]$
$R1$ Error trigger (control)	2.35	578	$+17.0\,[-5.4, +39.0]$
$R2$ Value router	2.43	600	$+7.8\,[-13.1, +28.6]$
$R2$ CVC, SD margin	2.65	679	$+16.0\,[-5.0, +37.8]$
$R2$ CVC, conformal	3.00	814	$+24.6\,[+0.6, +50.4]$
$R2$ Error trigger (control)	2.35	708	$+19.7\,[-0.6, +41.5]$

64.8%, with contrasts $+1.7\,[-0.6, +4.1]$ versus $+2.6\,[+0.2, +5.1]$. Strict audited gains likewise reach $+0.094$–$+0.143$ on Qwen but only $+0.011$–$+0.015$ on Gemma (Table 18). Extraction sensitivity is thus backbone-specific; Appendix E.1 reports the related GSM8K effect.

Error-trigger baselines. Error-trigger controls in Tables 12, 13, and 14 have no positive interval excluding zero. On Gemma, both act on only 3.6% of questions.

Pilot versus evaluation. Pilot effects were larger than evaluation effects. Qwen's pilot complementary-opportunity rate was $34/200 = 17.0\%$, with Coordination-minus-Stop success $+8.5\,[+4.25, +13.0]$ points and utility $+0.0768\,[+0.0349, +0.1217]$. On 500 evaluation questions, the same contrasts were $+0.2$ points and -0.0059 utility. Gemma's pilot opportunity rate was $26/200 = 13.0\%$, with utility $+0.0362\,[-0.0085, +0.0813]$, versus $+0.0081$ at evaluation. Evaluation opportunity rates were $47/500 = 9.4\%$ and $53/500 = 10.6\%$. These split comparisons are descriptive; no test of population change was performed.

Selected-event coverage. Across thirty conformal gate–seed combinations, marginal replay-label coverage is 85.6–93.4% against nominal 90%. None selects a question, so selected coverage is undefined. SD-margin gates select 45–311 of 500 questions. Pooled selected-label coverage is $48/2{,}650$ on Qwen, $51/2{,}206$ on Gemma, and $41/1{,}551$ on Mistral-Nemo. These question–fit–family records reuse questions across five fits and two reference families. Pooled coverage is below 3%, although individual gates differ: primary Mistral-Nemo $R1/R2$ coverage is $9/255$ (3.53%) and $10/202$ (4.95%). SD margins have no conformal guarantee, and replay-label coverage does not establish mean coverage.

Table 18: MMLU-Pro independent audits: 18 candidates, 1,000 questions per backbone, $K = 3$ per reference family, $\delta = .025$, $R_\lambda = 1.12288$. Acceptance requires a positive primary Hoeffding bound; none passes. EB bounds and strict-scoring gains are prespecified sensitivities, not acceptance rules. Declining a candidate does not establish harm.

Candidate	Ref.	Mean ΔJ_λ	Hoeff. LCB	EB LCB	Strict ΔJ_λ	Decision
Qwen3-8B						
Value router	Stop	+0.00517	−0.1047	−0.0475	+0.1187	decline
CVC gate, SD margin	Stop	+0.01165	−0.0982	−0.0373	+0.0942	decline
CVC gate, conformal	Stop	+0.00000	−0.1099	−0.0287	+0.0000	decline
Value router	Vote	+0.00643	−0.1034	−0.0462	+0.1429	decline
CVC gate, SD margin	Vote	+0.00817	−0.1017	−0.0424	+0.1137	decline
CVC gate, conformal	Vote	+0.00000	−0.1099	−0.0287	+0.0000	decline
Gemma-3-12B-it						
Value router	Stop	+0.00826	−0.1016	−0.0451	+0.0108	decline
CVC gate, SD margin	Stop	+0.01082	−0.0991	−0.0399	+0.0128	decline
CVC gate, conformal	Stop	+0.00000	−0.1099	−0.0287	+0.0000	decline
Value router	Vote	+0.01235	−0.0975	−0.0397	+0.0149	decline
CVC gate, SD margin	Vote	+0.00946	−0.1004	−0.0409	+0.0120	decline
CVC gate, conformal	Vote	+0.00000	−0.1099	−0.0287	+0.0000	decline
Mistral-Nemo						
Value router	Stop	+0.00902	−0.1009	−0.0426	+0.0210	decline
CVC gate, SD margin	Stop	+0.00918	−0.1007	−0.0391	+0.0207	decline
CVC gate, conformal	Stop	+0.00000	−0.1099	−0.0287	+0.0000	decline
Value router	Vote	−0.01152	−0.1214	−0.0662	+0.0005	decline
CVC gate, SD margin	Vote	−0.00712	−0.1170	−0.0576	+0.0044	decline
CVC gate, conformal	Vote	+0.00000	−0.1099	−0.0287	+0.0000	decline

Audit resolution. At $n = 1{,}000$, $K = 3$, and $\delta = .025$ per family, the Hoeffding penalty is 0.110, versus positive Qwen/Gemma audit means of about 0.005–0.012. Holding the largest observed effects fixed gives bound-crossing sample sizes of roughly 79,000–89,000; also holding variance fixed gives 7,100–7,500 for empirical Bernstein. These are conditional sensitivity calculations, not prospective power estimates. They change no audit decision, and candidate audits are not independent certification trials.

Structural and cost sensitivity. Six primary conformal candidates make no observed departures and have zero gains, but retain $K = 3$ family sizes and lower bounds $-.10988$. Observed inactivity does not establish policy identity on all histories. EMS3's difference from Vote lies in $[0, \lambda L/1000]$,

where L is the third-response cap (1,024 on GSM8K; 2,048 on MMLU-Pro). Exploratory bounds use this smaller cost-saving range without adding EMS3 to the original audit families.

Pricing input and output tokens equally at .01 per thousand gives Qwen's selected router $-.0056\,[-.0211, .0095]$ against Stop and Gemma's $-.0122\,[-.0317, .0072]$ against Vote ($n = 500$ each). EMS3-minus-Vote is $+.00523\,[.00485, .00563]$ and $+.00533\,[.00491, .00576]$. These exploratory contrasts retain the output-only primary objective. On GSM8K, mean-over-fits $M = 4$ minus $M = 2$ at 850 training questions is $+.00023\,[-.00207, .00230]$. Five SD-margin fits select three hyperparameter combinations and disagree on 33.28% of evaluation actions across fit pairs, measuring fitting instability on shared questions.

Compute and cost accounting. Qwen ran on six L40S GPUs; Gemma and Mistral-Nemo each ran on separate eight-A100-SXM4-80GB systems. For 3,000 evaluation calls per backbone, input/output tokens and summed generation seconds are: Qwen, 2,135,852/1,218,868 and 33,731; Gemma, 2,313,636/1,349,515 and 75,288; Mistral-Nemo, 1,474,596/621,994 and 15,388. Per-call times are not end-to-end latency and do not support cross-hardware speedup claims. Offline collection counts unique calls; online cost follows the executed path. Model-native tokens and utility contrasts are comparable within backbones; cross-backbone costs use calls and output characters.

G Post-Hoc Value Diagnostics

Formal definition. Let H be the question and realized initial answer A, and $j_0(H)$ the fixed Stop utility. Because the reference answer is a function of the question, unrestricted functions of H may implicitly use it: $V^*_{\cdot,H}$ is an oracle benchmark, not a finite-data routing target. Let Z_0, Z_1 be utilities of two coordination paths under the same frozen continuation, $D_r = Z_r - j_0(H)$, and $G = \frac{1}{2}\big(\mathbf{1}\{D_0 > 0\}D_1 + \mathbf{1}\{D_1 > 0\}D_0\big)$.

Proposition 3. *If Z_0 and Z_1 are conditionally i.i.d. given H with $\mathbb{E}|Z_r| < \infty$, then with $p(H) = \Pr(D_r > 0 \mid H)$ and $\delta(H) = \mathbb{E}[D_r \mid H]$, $\mathbb{E}[G \mid H] = p(H)\,\delta(H) \le \max\{\delta(H), 0\}$, and therefore $\mathbb{E}[G] \le \mathbb{E}\big[\max\{\delta(H), 0\}\big] = V^*_{\{\mathrm{Stop},\mathrm{Coord}\},H} - V_{\mathrm{Stop}} \le V^*_{\mathcal{A},H} - V_{\mathrm{Stop}}$, where $V^*_{\cdot,H}$ is the value of the best policy measurable from H over the indicated action set.*

In particular, the expected cross-replay gain satisfies

$$\mathbb{E}[G] \le V^*_{\mathcal{A},H} - V_{\mathrm{Stop}}. \tag{12}$$

Proof. Conditional independence gives $\mathbb{E}[\mathbf{1}\{D_0 > 0\}D_1 \mid H] = p(H)\delta(H)$ and symmetrically, so $\mathbb{E}[G \mid H] = p(H)\delta(H)$. For $p \in [0, 1]$ and real δ, $p\delta \le \max\{\delta, 0\}$. Taking expectations and noting that for a finite action set the measurable conditional-mean argmax with deterministic tie-breaking attains $\mathbb{E}[\max\{\delta(H), 0\}]$ over $\{\mathrm{Stop}, \mathrm{Coord}\}$ gives the identity, and enlarging the action set cannot decrease the optimum. □

Scope and assumptions. The bound concerns expectations, not individual outcomes or sample means. For $D = Z - c$, $Z \sim \mathrm{Bernoulli}(c)$, and $c = 1/4$, the optimal binary gain and $\mathbb{E}[G]$ are zero, yet $G = 3/4$ when both replays succeed (probability $1/16$). Independence is essential: reusing one outcome gives $\mathbb{E}[G] = c(1 - c) = 3/16 > 0$ despite zero optimal gain. Separate seed streams with shared A implement the intended design but do not prove stochastic independence. The optimum $V^*_{\mathcal{A},H}$ ranges over unrestricted histories, not the fitted 55-dimensional feature map, ridge family, label budget, or selection rule.

The post-hoc selector uses one reference-scored replay to choose an action and another to score it. Displayed cost excludes the selecting replay, so this is not a deployable policy at that cost. Under Proposition 3, a valid lower confidence limit for $\mathbb{E}[G]$ also bounds the oracle advantage. Reported intervals are pointwise and exploratory, not a multiplicity-adjusted existence test or safety certificate.

Policy and selector definitions. Table 19 covers Qwen and Gemma only. L1 answers alone; L3_dev is selected on development and frozen. L3 selects the best frozen router on evaluation and re-maximizes within each bootstrap resample; its intervals are not fixed-policy intervals. Fixed-winner contrasts with Stop include zero on both backbones. L4 selects an admission offset on evaluation

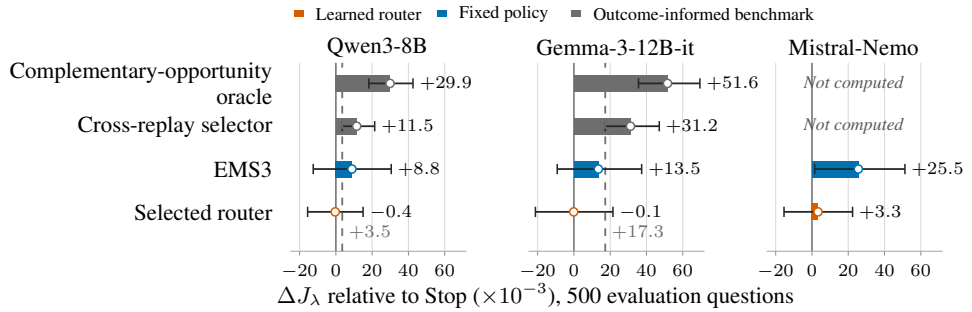


Figure 12: **Post-hoc value diagnostics on MMLU-Pro.** Paired ΔJ_λ versus Stop ($\times 10^{-3}$; pointwise 95% question-bootstrap intervals). The top two rows use outcome information and do not represent deployable policies; the cross-replay selector excludes acquisition of the selecting replay from its displayed cost. Dashed lines mark the selector's lower interval endpoints, whose interpretation for the unrestricted-history optimum requires Proposition 3. These exploratory intervals are not multiplicity-adjusted.

Table 19: **Post-hoc value diagnostics.** Normalized scoring on 500 Qwen/Gemma evaluation questions; ΔJ_λ is relative to Stop (10^{-3} units; pointwise bootstrap intervals). *: evaluation-selected or outcome-informed; ‡: cross-replay selection, excluding the selecting replay's cost; $^\circ$: outcome-informed. Policy definitions below explain the selection biases. Proposition 3 concerns the expectation of L6, not its sample mean. Fixed EMS3 results are in Tables 12 and 13.

		Qwen3-8B		Gemma-3-12B-it	
ID	Policy	J_λ	ΔJ_λ vs. Stop	J_λ	ΔJ_λ vs. Stop
L1	Stop (answer alone)	0.6728	+0.0 [+0.0, +0.0]	0.6521	+0.0 [+0.0, +0.0]
L3_dev	Development-selected candidate (frozen)	0.6724	−0.4 [−15.6, +15.0]	0.6520	−0.1 [−21.2, +21.5]
L3	Best learned router, evaluation maximum*	0.6784	+5.6 [+0.9, +29.5]	0.6628	+10.7 [+2.7, +35.9]
L4	Best admission threshold, evaluation-optimized*	0.6767	+3.9 [+0.0, +24.1]	0.6623	+10.2 [−3.7, +36.5]
L6	Cross-replay selector‡	0.6843	+11.5 [+3.5, +21.4]	0.6833	+31.2 [+17.3, +47.1]
L5	Complementary-opportunity oracle$^\circ$	0.7027	+29.9 [+18.2, +42.5]	0.7037	+51.6 [+35.6, +69.5]
L7	Five-action oracle on realized outcomes*	0.7520	+79.2 [+57.4, +103.0]	0.7350	+82.9 [+61.2, +106.7]

while fixing coefficients, conservative multiplier, eligibility mask, ranking, and tie rules. Offsets $+0.142$ and $+0.135$ yield 98.8% Stop on Qwen and 97.2% Vote on Gemma, with 0.4% Coordination each; this does not rule out other threshold changes. L5 coordinates when a complementary peer exists; L6 is the cross-replay selector; L7 chooses the best of five realized actions. L5/L7 use outcome information, and L7 also selects on realized noise. Ratios among these diagnostics do not estimate a learnable fraction of the optimum.

Ranking diagnostics. Using the frozen router's unpenalized Coordination-minus-Stop score, complementary-opportunity AUC is $0.481\,[0.388, 0.576]$ on Qwen and $0.550\,[0.462, 0.637]$ on Gemma. Masking to the six response properties without refitting gives $0.553\,[0.457, 0.650]$ and $0.579\,[0.491, 0.667]$; masking to TF–IDF gives $0.408\,[0.323, 0.495]$ and $0.530\,[0.445, 0.614]$. Five intervals contain 0.5; Qwen TF–IDF lies below it. For positive net gain, AUC is $0.4802\,[0.3729, 0.5872]$ and $0.5842\,[0.4886, 0.6784]$. Spearman correlations with realized gain are $-0.049\,[-0.135, +0.040]$ and $-0.076\,[-0.165, +0.015]$. These results do not establish chance equivalence or absence of predictive information. Complementary peer correctness differs from net benefit because aggregation can lose the answer and incur cost.

H Controlled Classifier Proxy

The proxy uses scikit-learn Breast Cancer (569 items), Digits (1,797), and Wine (178), with stratified 50/20/20/10 training/calibration/test/audit splits (test $n = 114/360/36$; audit $n = 57/180/18$; seeds 2201/2203/2202). Agent A is a depth-8 tree, B standardized logistic regression, and C a 300-tree random forest. Coordination calls B, then C on disagreement with A, at costs .02 and .04 with

$\lambda = 1$ and $R_\lambda = 1.06$. The gate compares Coordinate with Stop at h_A; the audit incumbent always coordinates ($K = 3$, $\delta = .05$, acceptance $\mathrm{LCB}_{\mathrm{audit}} \geq 0$). Separate 400-tree regressors predict $\widehat{\Delta}_U, \widehat{\Delta}_C$, with Bonferroni-corrected split-conformal endpoints at $\alpha = .10$. Local adaptations, not official reproductions, provide baselines: TreeCredit-adapt fits positive best-sibling ΔJ_λ (Liu et al., 2026b); iMAD-adapt predicts initial error before $B{+}C$ (Fan et al., 2026a); EMS-adapt calls B, then C on disagreement (Liu et al., 2026c). Implementation: Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0; 10,000 paired bootstrap resamples.

The conservative gate beats frequency-matched random routing on Breast Cancer and Digits but trails the strongest adaptive baseline on all tasks (Table 20; Figure 13). Coordination costs at most .06, so the best policy nearly always coordinates; oracle values are .9442, .9508, and .9950. Marginal joint utility–cost coverage is 90.35%, 87.78%, and 100%, versus selected coverage of $0/2$ and $21/38$ (undefined on Wine). A point-estimate gate attains .9235, .9301, and .9800, matching the best baseline. Referencing Stop makes the conservative gate favor silence; referencing the always-coordinate incumbent would favor coordination unless skipping is confidently safe (Corollary 1). Audit gains are $+.0239$, $-.0311$, and $-.1067$, with intervals $[-.0288, .0600]$, $[-.0763, .0096]$, and $[-.2733, .0600]$; all Hoeffding bounds are negative. Holding the Breast Cancer mean fixed, the bound crosses zero at roughly 16,000 items.

Table 20: Controlled proxy ($\lambda = 1$). Best adapt. is the largest J_λ among the TreeCredit (T), iMAD (I), and EMS (E) adaptations; the conservative CVC gate is compared with a random policy matched to its coordination frequency; the last two columns audit the frozen gate against the always-coordinate incumbent on independent items. Selected coverage is joint utility–cost coverage conditional on a positive coordination decision (undefined without triggers).

	Best adapt.	Conservative CVC gate					ΔJ_λ vs. random	Selected	Independent audit	
Dataset	J_λ (method)	Acc.	Cost	J_λ	Calls	Coord.	[95% CI]	coverage	ΔJ_λ	LCB
Breast Cancer	.9235 (E)	.9123	.0011	.9112	1.0351	.0175	+.0168 [.0002,.0418]	.0000	+.0239	-.3779
Digits	.9301 (T/E)	.8472	.0051	.8421	1.1806	.1056	+.0316 [.0107,.0541]	.5526	-.0311	-.2572
Wine	.9800 (T)	.9167	.0000	.9167	1.0000	.0000	.0000 [.0000,.0000]	—	-.1067	-.8216

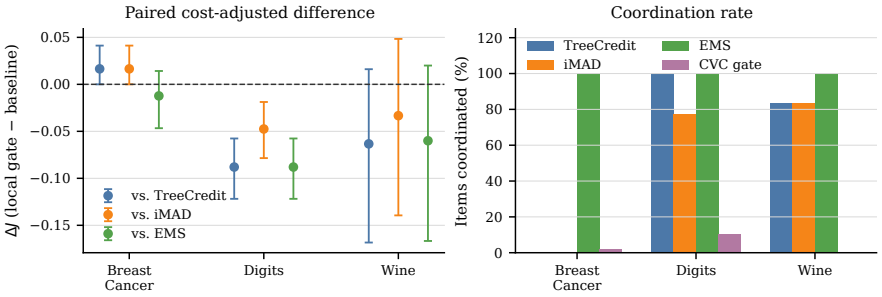


Figure 13: Controlled proxy. Left: Paired ΔJ_λ of the conservative CVC gate against each adaptive baseline, with 95% paired-bootstrap intervals. Right: Fraction of test items on which each policy coordinates: the baselines coordinate on most or all items, the conservative gate on 1.8%, 10.6%, and 0%.