COVERAGE IS NOT CONVERSION: A SAME-POOL AUDIT OF LLM SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

An LLM may generate a correct answer yet fail to select it. Consider a gate that switches between a baseline and a proposed selector. Even a perfect gate cannot fix cases where both actions miss a correct candidate. We develop a same-pool audit that compares this missed opportunity with the most a gate could add to the always-deployed proposal. It also separates information loss from errors in using that information. One joint confidence set answers two questions: does the proposal improve the baseline, and does missed opportunity exceed the remaining gate budget? The criteria are sharp for that set. In a secondary analysis of nine fixed LiveAoPS and FEVEROUS comparisons, the budget ordering holds even though no positive gain is certified. A prospective MedMCQA study improves answer accuracy by 1.005 percentage points over direct answering on 20,000 new groups, but loses .470 points to unchanged plurality. A further 4,494-group reasoning study does not confirm improvement over ordinary reasoning. The audit thus identifies limits on gate-only improvement; a proposed intervention still needs validation against a strong control.

1 INTRODUCTION

Test-time sampling gives a model more chances to produce a correct answer. It does not guarantee that the system will choose one. In 500 MATH-500 pools (Hendrycks et al., 2021), appending eight Qwen2-Math-7B candidates to each eight-candidate pool raises coverage from 51.6% to 57.2%. Majority accuracy moves only from 38.4% to 38.6%. The extra candidates create opportunity, but the decision rule converts little of it into better answers.

Selection can help, but its benefit depends on the endpoint. In a retrospective MATH-500 comparison, a target-guided selector raises accuracy from 43.4% to 45.4% under the same final-answer head. On Omni-MATH (Gao et al., 2024), the same comparison raises retained-set coverage from 30.0% to 32.5%, while final-answer accuracy stays at 16.5% (Appendix F). These results motivate a more precise question: *where is the available opportunity lost, and which part can the current decision mechanism still recover?*

An inconclusive gain test leaves that question open. Consider a proposal that sometimes rescues the baseline and sometimes harms it. A perfect gate could keep every rescue and avoid every harm. It still could not recover correct candidates missed by both actions. If this missed capture exceeds the gate's entire remaining budget, better gating alone cannot address the larger loss. This comparison can be statistically resolved even when the proposal's net gain is not. It bounds what a repair could recover; it does not predict the benefit or cost of a replacement.

We formalize this distinction with three losses. Both actions may miss a correct candidate already in the pool: this is *missed capture*. The proposal's rescues and harms may look alike under the available signals: this is *information loss*. The gate may also use those signals poorly: this is *gate regret*. We use these three names throughout. The last two together form the remaining gate budget. Figure 1 shows where each loss occurs and how an intervention is checked.

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.

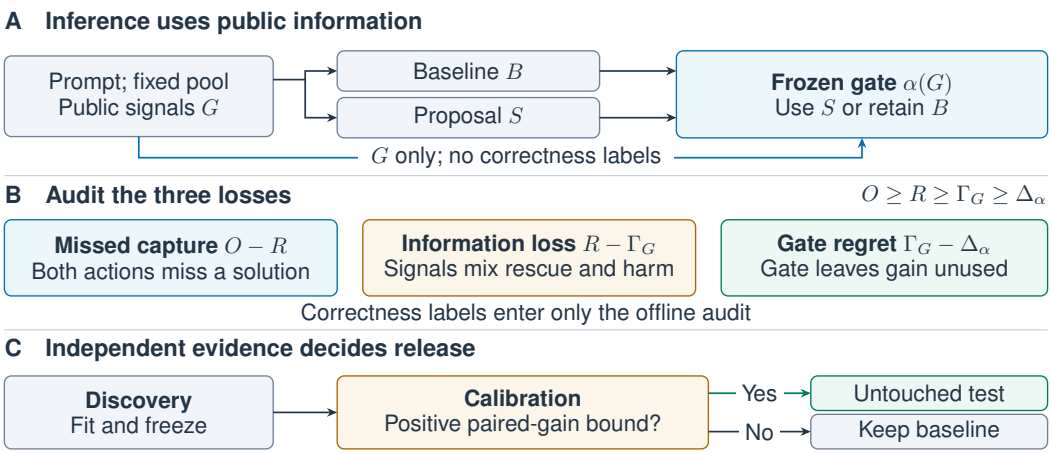

Figure 1: From candidate opportunity to a validated decision. Inference uses only public state and frozen rules. Endpoint labels support the audit and held-out checks. The three losses sum to the opportunity left unconverted. The release path illustrates independent calibration; the later reasoning confirmation uses a separately frozen direct test. A repair budget is an upper limit, not a promised improvement.

Contributions. We give a loss decomposition for a fixed pair of actions and show exactly how added information reduces its information loss. We then derive two sharp decisions from one confidence set: whether the proposal improves the baseline, and whether missed capture exceeds the remaining gate budget. The resulting upper bound can rule out a desired gain from changing only the gate. Independent calibration and learning bounds specify how a proposed change can be validated. Finally, nine fixed comparisons and prospective MedMCQA tests distinguish a resolved limit, an observed benefit and a successful intervention. These are separate empirical claims.

Relation to prior work. Hu measures recoverable mass, capture and harm in fixed candidate pools (Hu, 2026). Shihab et al. distinguish oracle opportunity, signal-restricted Bayes gain and held-out router gain, with selection-valid family inference (Shihab et al., 2026). We build on these decision objects. Our question is narrower: for a specified proposal, is the larger loss caused by the actions it misses or by choosing between the available actions? Table 1 makes the distinction concrete. Two laws have identical opportunity, signal-limited Bayes gain and realized gain, yet opposite budget orderings. Rescue/harm counts resolve this ambiguity; our joint confidence set turns them into a finite-sample decision.

The contribution is this proposal-specific decomposition and decision. Full paired records from earlier frameworks already contain the required counts. The constructed comparison isolates what their three coarse summaries leave open; it does not establish an empirical performance advantage over those frameworks. Section 3 applies the decision to a real FEVEROUS row. The Bayes rule, information monotonicity and multiplicity corrections remain classical ingredients.

Imperfect reward signals already limit inference-time alignment (Huang et al., 2025). Search, self-refinement and representation-based exploration can instead change the candidate law (Snell et al., 2024; Madaan et al., 2023; Tuyls et al., 2026). Our audit holds each configuration fixed and does not optimize how compute should be divided between generation and selection. Its release rule shares the principle of high-confidence policy improvement (Thomas et al., 2015b), with both actions observed here. Paired gain differs from the absolute-risk objectives of selective prediction and conformal risk control (El-Yaniv & Wiener, 2010; Angelopoulos et al., 2024). Appendix B.5 compares the required feedback and guarantees.

2 MEASURING WHERE CANDIDATE COVERAGE IS LOST

2.1 FIX THE ACTIONS, INFORMATION, AND ENDPOINT

A unit consists of a prompt, a candidate pool $Y = \{y_1, \dots, y_K\}$, public state G, and hidden correctness $C \in \{0,1\}^K$. The baseline $B = b(G)$ and proposal $S = a(G)$ select subsets of the same realized pool, with the same declared size and fixed tie-breaking. For a subset A, define

$$Y_A = \mathbf{1}\left\{\sum_{i \in A} C_i > 0\right\}, \qquad \text{sPass}@m(a) = \mathbb{E}[Y_{a(G)}]. \tag{E}$$

This is coverage of the retained set. It is neither fresh-sample pass@m nor final-answer accuracy. For single-answer actions, such as the MedMCQA gate, Y_B, Y_S are ordinary answer correctness. A final-answer head must be evaluated separately when a selector retains multiple candidates.

Public state means the fields available to the decision at inference time. It may contain candidate text, scores and parsed choices. It excludes current correctness, references, execution endpoints and held-out labels. Fitted parameters, feature extraction, ties and the unseen-state fallback are fixed before calibration. For example, the MedMCQA refinement uses a high-consensus indicator derived from eight parsed choices. It changes the representation and gate, not the available information. Appendix B.1 specifies each route.

Let $H = Y_{[K]}$ denote full-pool success. Three event probabilities suffice for the first audit:

$$\begin{aligned} O &= \Pr(H = 1, Y_B = 0), && \text{opportunity}, \\ R &= \Pr(Y_S = 1, Y_B = 0), && \text{rescue}, \\ L &= \Pr(Y_S = 0, Y_B = 1), && \text{harm}. \end{aligned}$$

The proposal's gain is $\Delta = R - L \le O$. Its missed capture is $M = O - R = \Pr(H = 1, Y_B = Y_S = 0)$. A rescue must use an opportunity; if $O = 0$, no selection from this pool can improve the baseline. These are paired outcome differences, not causal treatment effects. We use final-answer accuracy to assess deployed performance. Compute cost can be traded for quality and need not match across algorithms; fixing the pool here defines the diagnostic question.

The identities below need no sampling assumption. Confidence claims use independent discovery, calibration and test roles when used, with within-role i.i.d. units from the same route-specific target law P_0. These are conditional, model-based claims, not design-based guarantees for a fixed benchmark. Changing the actions, endpoint or target law requires a new audit.

2.2 THREE LOSSES AND THE VALUE OF PUBLIC INFORMATION

Write $D = Y_S - Y_B \in \{-1, 0, 1\}$. A gate $\alpha(G) \in [0,1]$ uses the proposal with that probability and otherwise returns the baseline. Any randomization is independent of the endpoint conditional on G. Define

$$\eta(G) = \mathbb{E}[D \mid G], \qquad \Gamma_G = \mathbb{E}[\eta(G)_+], \qquad \Delta_\alpha = \mathbb{E}[\alpha(G) D]. \tag{F}$$

The classical Bayes rule takes the proposal where $\eta(G) > 0$. Hence

$$\sup_{0 \le \alpha \le 1} \Delta_\alpha = \Gamma_G, \qquad 0 \le \Gamma_G \le R \le O. \tag{F2}$$

Every coarsening $Z = h(G)$ has $\Gamma_Z \le \Gamma_G$. The frontier is a population quantity; fitting and evaluating it on the same labels yields only an in-sample diagnostic.

Proposition 1 (Proposal-local loss decomposition)**.** *Every public gate for the fixed baseline–proposal pair satisfies*

$$\underbrace{O - \Delta_\alpha}_{\textit{unconverted opportunity}} = \underbrace{O - R}_{\textit{missed capture}} + \underbrace{R - \Gamma_G}_{\textit{information loss}} + \underbrace{\Gamma_G - \Delta_\alpha}_{\textit{gate regret}}. \tag{I}$$

All three losses are nonnegative.

Proof. A rescue implies an opportunity, so $R \le O$. Conditional Jensen gives $\Gamma_G \le \mathbb{E}[D_+] = R$, and $\alpha\eta \le \eta_+$ gives $\Delta_\alpha \le \Gamma_G$. The sum telescopes. □

Each term concerns a different limit. The proposal controls which baseline failures it can rescue. Public information controls whether those rescues can be distinguished from harms. The gate controls how well the available distinction is used. In particular,

$$\Gamma_G - \Delta_\alpha = \mathbb{E}[\eta(G)_+ - \alpha(G)\eta(G)]. \tag{H}$$

This is an accounting of the fixed decision problem, not a unique causal explanation across all possible selectors.

Information loss has a useful equivalent expression. For any public state W, define rescue–harm overlap by $\mathrm{ov}_W = \mathbb{E}\min\{\Pr(D = 1 \mid W), \Pr(D = -1 \mid W)\}$. It measures how much rescue and harm remain indistinguishable under those signals. The next identity equates this overlap with information loss.

Proposition 2 (Overlap and representation value)**.** *For a fixed proposal and target law,*

$$R - \Gamma_G = \mathrm{ov}_G, \qquad \Gamma_G - \Gamma_Z = \mathrm{ov}_Z - \mathrm{ov}_G \quad \text{if } Z = h(G).$$

Thus refinement is worth exactly the overlap it removes. The frontier equals rescue mass iff rescue and harm have disjoint support in public state.

This identity specializes Blackwell's information ordering to one paired decision (Blackwell, 1953; Reid & Williamson, 2011). Equal frontiers here do not imply equivalent information for other tasks. Nor is Γ_G a computational guarantee: if D is a deterministic measurable function of complete G, then $\Gamma_G = R$, even when computing that function is infeasible. An empirical coarsening estimates Γ_Z, not the unrestricted frontier.

How to measure the losses. Freeze the actions, public coarsening, gate, endpoint and sampling unit. For each audit row, record $(Z_i, Y_{B,i}, Y_{S,i})$ and full-pool success H_i, when available. Let n_z^+, n_z^- count rescues and harms in cell z, and let n_M count joint failures with $H_i = 1$. Then

$$\widehat{M} = \frac{n_M}{n}, \quad \widehat{R} = \frac{\sum_z n_z^+}{n}, \quad \widehat{\Gamma}_Z = \frac{\sum_z [n_z^+ - n_z^-]_+}{n}, \quad \widehat{\mathrm{ov}}_Z = \frac{\sum_z \min(n_z^+, n_z^-)}{n}.$$

All cell counts use the total denominator. These estimates describe the sample; confidence bounds support the decisions below. Partial inspection can lower-bound missed capture when exhaustive labels are unavailable (Section 4.2).

3 Two Decisions from the Same Evidence

3.1 What can a better gate still recover?

The sum of information loss and gate regret is $\mathcal{B}_\alpha = R - \Delta_\alpha$. This is the remaining downstream budget, allowing even endpoint-informed selection between the same actions. It can be computed without fitting η. More generally,

$$\begin{aligned} R - \Gamma_G &\in [0, R - \max\{\Delta_\alpha, 0\}], \\ \Gamma_G - \Delta_\alpha &\in [\max\{-\Delta_\alpha, 0\}, R - \Delta_\alpha]. \end{aligned} \tag{I′}$$

If $O - R > \mathcal{B}_\alpha$, missed capture exceeds both downstream losses combined. For an always-deployed proposal, $\Delta = R - L$, so $\mathcal{B} = L$. An oracle gate can remove its harms; it cannot fix a pool that both actions miss. Equivalently, $M > L$ iff $O > R + L$: opportunity exceeds endpoint disagreement. A rarely active proposal can therefore satisfy this ordering easily. Its meaning depends on which actions are available for repair.

What coverage and gain leave unidentified. Table 1 gives two constructed laws. Both have baseline accuracy .60, proposal accuracy .62 and full-pool coverage .80. Under a constant public representation Z, both therefore have $(O, \Gamma_Z, \Delta) = (.20, .02, .02)$. Yet their capture and gate budgets have opposite orderings. Knowing these three summaries exactly is insufficient to determine $M > L$. Observing rescue mass resolves the ambiguity. The construction in Appendix A.1 proves this limited nonidentification statement; it is not an empirical comparison.

Table 1: Equal coarse summaries, different repair budgets. All entries are probabilities under constructed laws. The proposal is always deployed, $\Gamma_Z = \Delta = .02$, so gate regret under Z is zero and $\mathrm{ov}_Z = L$.

Law	Rescue R	Miss M	Overlap	Budget L	$M > L$?
A	.04	.16	.02	.02	Yes
B	.14	.06	.12	.12	No

Now suppose simultaneous intervals for (M, R, L) cover every member of a fixed finite comparison family, with lower endpoints ℓ and upper endpoints u. Intersect their rectangle with $m + r + l \leq 1$, giving a nonempty confidence set $\mathcal{C}$. Set $\bar{l} = \min\{u_L, 1 - \ell_M - \ell_R\}$. The two decisions reduce to

$$\begin{aligned} &\text{Positive gain:} && \inf_{\mathcal{C}} \Delta = \ell_R - \bar{l} > 0, \\ &\text{Capture-budget dominance:} && \inf_{\mathcal{C}} (M - L) = \ell_M - \bar{l} > 0. \end{aligned} \tag{Q}$$

The feasible law $(\ell_M, \ell_R, \bar{l})$ attains both minima. Thus each criterion is sharp relative to this confidence set: if it fails, a compatible law defeats that conclusion. This is not a claim of optimality over all statistical tests. The family construction and a direct Holm alternative allow dependent comparison rows (Appendix B.7).

One real row, two answers. On 400 FEVEROUS pools, Self-Certainty top-4 versus DPP has 62 missed opportunities, 31 rescues and 19 harms. The proposal gains 3.00 percentage points, but misses 15.50 points of opportunity. On these sampled pools, even an oracle gate can add at most the 4.75 points lost to harm. Under the joint confidence set, $(\ell_M, \ell_R, \bar{l}) = (.10364, .04199, .09008)$, rounded. The lower capture-budget margin is therefore $+1.36$ points, while the gain lower bound is -4.81 points. The evidence resolves the budget ordering without certifying gain. The simplex constraint does not tighten this empirical row; Appendix A gives its full ledger and a constructed case where the constraint matters.

A concrete use: rule out an unattainable target. For a desired additional gain τ over always deploying S, every gate between B and S satisfies $\Delta_\alpha - \Delta \leq L \leq \bar{l}$ on the joint confidence event. Thus $\tau > \bar{l}$ rules out reaching that target by changing only the gate. In this row, the upper budget is 9.01 percentage points of retained-set success. A larger target requires actions beyond this pair. This is a stopping criterion for that restricted design space, not a rule for choosing the best replacement. A cheap improvement within the budget may still be useful. The target check and the positive-gain decision in Equation (Q) answer different questions. The supplement's `audit/diagnostic_walkthrough.py` reproduces all nine rows, these two margins and the constructed example with one offline command.

3.2 Learn a change, then validate it independently

Finer representations can reduce information loss, but learning a separate decision for each cell requires more data. Discovery fits the gate; independent calibration checks its gain. Proposition 6 gives the full finite-sample bound and makes this trade-off explicit. Its discovery lower bound concerns learning with regret small relative to the available Bayes gain. Its calibration lower bound concerns fixed-power validation of a supplied oracle gate. The former does not lower-bound the data needed to validate any arbitrarily small positive gain.

Validation needs only paired outcomes. Once a binary gate $\widehat{\alpha}(G)$ is frozen, its calibration outcome $X = \widehat{\alpha}(G)(Y_S - Y_B) \in \{-1, 0, 1\}$ is one fixed contrast. A valid lower bound on $\Pr(X = 1) - \Pr(X = -1)$ suffices for release. Neither full-pool opportunity nor a fitted frontier is required. Our protected LLM routes use exact count bounds, including Clopper–Pearson (CP) bounds (Clopper & Pearson, 1934), rather than an assumed activity envelope. Continuous features and discovery search size add no calibration penalty when the selected gate is frozen independently. Selecting on calibration does require simultaneous protection.

This is a target-average gain guarantee, not per-prompt safety. Finite action families and learned representations follow the same discovery/freeze/validate logic (Appendix B.2). Bounded quality–cost utilities and justified distribution-shift bounds require the separate assumptions in Appendix B.9.

4 What the Audit Resolves in LLM Selection

The experiments address three questions: whether more coverage improves the final answer, whether the gate's remaining budget can be bounded, and whether an intervention beats a strong control. The nine-row audit measures retained-set success. It compares missed capture with the combined gate budget, without separately estimating information loss and gate regret. Its actions were fixed before endpoint access, but the confidence analysis is secondary. The MedMCQA studies test final answers on new allocations. The evidence map in Table 5 distinguishes these claims and endpoints. Appendix I records sampling units, label-access procedures and analysis families; Table 19 retains earlier stopping decisions. The named selector variants are the objects of evaluation; the audit applies to any fixed baseline–proposal pair.

4.1 Coverage gains require an endpoint check

Table 2 reports every historical math route under one frozen final-answer head. The matched target uses public answer-dependence information in subset selection (Appendix C.1). MATH at $K = 16$ gives a realized $+2.0$-point answer gain, with 13 rescues and three harms. Omni-MATH illustrates the limit: better retained-set coverage produces no answer gain. Thus the endpoint bridge must be measured, even when coverage improves.

Table 2: Historical endpoint bridge: DPP versus the matched-target selector. The first three routes use Qwen2-Math-7B; Cross MATH mixes four outputs each from Qwen2-Math-7B and Qwen2.5-7B. All use a DeepSeek-R1-8B verifier. Omni-MATH is a 200-question numeric-answer subset with difficulty at least six. Coverage refers to the retained set; accuracy uses the same frozen answer head for both selectors. Answer-gain intervals are descriptive paired-bootstrap 95% intervals. All four routes are retrospective; $m = K/2$.

Route	n	Coverage (%)	Accuracy (%)	Gain [interval], pp
MATH, $K = 8$	500	50.2 → 50.2	43.8 → 44.8	+1.0 [$-.4$, $+2.6$]
MATH, $K = 16$	500	55.8 → 56.6	43.4 → 45.4	+2.0 [$+.6$, $+3.6$]
Omni-MATH, $K = 16$	200	30.0 → 32.5	16.5 → 16.5	0.0 [-3.0, $+3.0$]
Cross MATH, $K = 8$	200	58.0 → 58.5	51.5 → 51.0	$-.5$ [-1.5, 0.0]

The MATH gain does not isolate the matched target's added value. Its advantage over a permuted target is only .8 points (4/0, descriptive exact $p = .125$); independence and constant-density controls also improve on DPP. The complete target and final-head ablations remain in Appendix F. These results motivate the audit without serving as prospective validation of it.

4.2 Nine resolved limits despite uncertain gains

The fixed family contains four LiveAoPS rows (Mahdavi et al., 2025): two selectors under two Llama-3.1-8B-Instruct seeds on the same 1,200 prompts, with $K = 8, m = 4$ and a Qwen2.5-7B verifier. It also contains five FEVEROUS proposals (Aly et al., 2021) versus a determinantal point process (DPP) baseline (Kulesza & Taskar, 2012), on 400 pools with $K = 64, m = 4$. Structured candidate answers are scored with Qwen3-4B-Instruct-2507, using Self-Certainty and entailment features over cited premises (Appendix I). Its endpoint requires both the verdict and complete evidence to be correct. The five selectors are target-guided selection (DAKPO), its group-coverage variant (DAKPO-CG), verifier-Borda, Best-of-Majority and Self-Certainty top-4. Figure 2 shows both criteria from the same joint 95% confidence set.

All nine capture lower bounds exceed their harm upper bounds. The simultaneous lower ratios of missed capture to harm range from 1.59 to 9.87; the direct Holm test agrees. Thus every row has more missed capture than the additional gain any gate could recover from its always-deployed proposal.

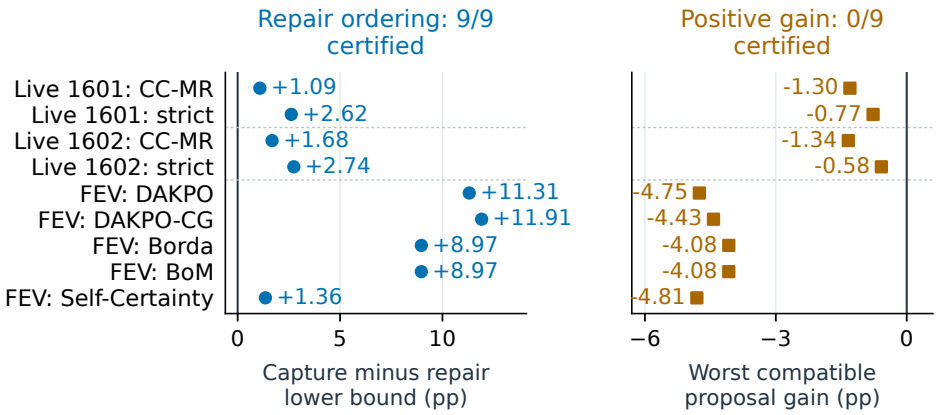

Figure 2: A repair budget can be resolved while gain remains uncertain. The endpoint is retained-set success. Each point is a lower bound from one joint 95% confidence region, not a point estimate with an error bar. All nine missed-capture margins are positive; none of the gain bounds is positive. The two LiveAoPS seeds share prompts, and Borda/Best-of-Majority share FEVEROUS endpoint vectors. This secondary analysis spans two domains, not nine independent benchmarks.

None has a simultaneous positive-gain certificate. The distinction is useful precisely because an unresolved gain test does not force an uninformative diagnosis.

The action family matters. The conclusion is local to each baseline–proposal pair. Declare the actions that an actual repair may use before interpreting the budget: a gate limited to two deployed choices and a router over an existing method family have different ceilings. For a larger family $\mathcal{A}$ containing the baseline, let $R_{\mathcal{A}}$ be the probability that some action rescues it. Relative to a specified reference policy of gain Δ_{ref}, the missed mass and remaining budget are $M_{\mathcal{A}} = O - R_{\mathcal{A}}$ and $\mathcal{B}_{\mathcal{A}} = R_{\mathcal{A}} - \Delta_{\text{ref}}$. Table 3 includes every proposal from the released families. Live's ordering changes: larger-family selection still has headroom. FEVEROUS retains a larger missed mass in the sample. A separate secondary sensitivity covers all 26 reference methods simultaneously. Each observed-best-reference interval crosses zero: neither family-level ordering is resolved. This analysis is separate from the nine-row certificate.

Table 3: Scope check with complete action families. Entries are counts on the same pools; divide by n for rates. The families contain 9, 9 and 5 proposals. The reference has the highest observed gain: Borda on Live and Self-Certainty on FEVEROUS. Conditional 95% intervals for $M_{\mathcal{A}} - \mathcal{B}_{\mathcal{A}}$ cover all reference choices (Appendix I).

Family	n	nO	$nR_{\mathcal{A}}$	$nM_{\mathcal{A}}$	$n\mathcal{B}_{\mathcal{A}}$	Interval (pp)
Live 1601	1,200	64	54	10	29	[−3.90, .76]
Live 1602	1,200	61	51	10	31	[−4.11, .64]
FEVEROUS	400	93	42	51	30	[−4.44, 14.75]

A secondary verifier-Borda comparison beats DPP under a one-sided two-seed conjunction with Holm-9 $p = .02970$; the two-sided sensitivity is .05940. This does not validate our registered proposal, but it prevents a broader claim that selection is exhausted. A rarely active proposal can have little remaining gate budget by construction; an action-family check is therefore essential before generalizing its diagnosis. The full method-level panel and derivation remain in Appendix I.

The diagnostic need not label every candidate. A retrospective masking study lower-bounds missed capture using partial inspection. For DAKPO, 6,898 of 25,600 possible candidate inspections find 41 of 89 misses and retain a positive nominal fixed-family margin. Smaller tested budgets fail; Self-Certainty requires exhaustive residual inspection in this grid. This establishes a label-requirement

example, not measured annotation savings. Appendix K.4 retains all five proposals and all eight budgets.

4.3 WHY THE STRONG CONTROL CHANGES THE CONCLUSION

MedMCQA (Pal et al., 2022) tests whether a promising representation actually supports a better decision. The baseline B is a direct answer; P is semantic plurality over eight Qwen3-4B-Instruct-2507 views. In an earlier calibration role, plurality had 230 rescues and 224 harms, so its gain was uncertified and 4,000 old test labels remained sealed. An exploratory high-consensus cell A required eight valid parses, a unique plurality different from direct, and at least six agreeing views. Its rescue/harm counts were 76/43, versus 154/181 outside the cell. This refinement raised the plug-in frontier from $6/4500$ to $33/4500$. It nominated a repair; it did not authorize deployment.

An independent study fixed a gate S: use P on A, otherwise use B. It allocated 10,000 new calibration groups and 20,000 new test groups, excluding all 12,200 earlier allocations. Independent calibration admitted $S - B$ with a gain lower bound of $+.1805$ percentage points. The new test required both a positive factorized CP bound and one-sided McNemar at .05 against direct answering. A fixed-sequence secondary then tested $S > P$. Both comparisons were required to establish the repair's added value.

Table 4: Prospective MedMCQA final-answer evidence, $n = 20{,}000$. Gain is in percentage points. The gate improves on direct answering but fails the required unchanged-plurality control. The last row is descriptive.

Contrast	Rescues	Harms	Gain (pp)	One-sided p
Gate S vs direct B	383	182	+1.005	9.09×10^{-18}
Gate S vs plurality P	649	743	-.470	.99457
Plurality P vs direct B	1,126	831	+1.475	descriptive

The primary passes, with a factorized 95% gain lower bound of $+.7155$ points. The repair control fails (Table 4). The cell-level result explains the observed loss: on the new test, plurality also helps outside A, with 743 rescues and 649 harms. Falling back to direct answering discards that benefit. The fallback contrast had the opposite sign during nomination, 154/181.

On the new sample, both cell contrasts are positive, so the refined and unrefined plug-in frontiers coincide at .01475. The three descriptive losses for fixed plurality are .11155 missed capture, .04155 overlap and .00470 gate regret. These estimates do not establish zero population refinement value or explain why the orientation changed. They do show why direct-baseline benefit cannot establish the nominated mechanism's value. No threshold is retuned and no old test is reopened (Appendix I).

4.4 A SEPARATE TEST OF REASONING IMPROVEMENTS

A separate confirmation asks whether contrastive short reasoning improves final answers over ordinary short reasoning. This changes generation and therefore tests a different intervention, not the fixed-proposal identity. On 4,494 newly reserved MedMCQA groups from a separate missing-topic population, contrastive reasoning answers 2,910 correctly (64.75%) and ordinary reasoning answers 2,907 (64.69%). The paired counts are 263 rescues and 260 harms: $+.067$ points, one-sided $p = .4652$, with a conservative marginal 95% interval $[-1.52, 1.65]$ points. The development lead is not confirmed.

Contrastive reasoning has the highest point estimate in the declared local panel, including a $+1.51$-point difference from plurality. Its three-answer lead over ordinary reasoning does not establish superiority or equivalence. Because the primary test fails, the twelve pool-control comparisons remain descriptive under the frozen gatekeeping rule. Appendix N reports all thirteen contrasts, the parser sensitivity, the earlier development stop, and the separately frozen confirmation design. Together, the two MedMCQA studies show why a repair claim needs the comparator that isolates its added mechanism.

4.5 SUPPORTING CHECKS AND SCOPE

Earlier studies retain the baseline for different reasons: sparse disagreement on HellaSwag, adjusted cell uncertainty on BFCL, and unresolved rescue–harm overlap on FEVEROUS. HellaSwag has only one favorable discordance; even an exact conditional sign test gives $p = .5$. Two later confidence-feature studies stop at development. Their full decisions and sealed roles are preserved in the appendices.

Retaining a strong baseline can still support an operational improvement. Exact strict-lead plurality stopping reduces measured warm time by 27.50% over 96 paired runs without changing actions or traces. A separate tie-aware replay reduces calls by 32.89% on 3,000 groups with zero action mismatches. The latter is a call-count projection, not a measured latency result. These elementary stopping rules improve cost, not accuracy or the paper's central diagnosis (Appendix L.1).

Controls for the formal mechanism. A synthetic D/C/T study completes the release path: calibration gives 47 rescues and two harms on 800 units, then untouched test gives 39/3 and gain .045. At the same 95% confidence level, CP count bounds admit the calibration contrast with lower gain .03448, while Hoeffding gives -.03029. This example shows the value of count-sensitive inference, not uniform superiority over other risk methods. A separate exact Gaussian control holds $\Gamma_G = .03873$ at equal absolute signal strength while changing gate alignment; gain changes from .03864 to -.01864. Only gate regret changes. Appendices A.5 and B.10 give both controls. Archived code-generation checks on LiveCodeBench, HumanEval+ and DSCodeBench also retain small gains and ties (Appendix J). They are descriptive selector evaluations, not replications of the nine-row confidence analysis. The historical final-answer and TAT-QA ablations have the same limited role: observed benefits do not by themselves establish the value of the proposed mechanism.

5 DISCUSSION AND CONCLUSION

Test-time sampling is useful when better candidates lead to better decisions. Our audit asks how much improvement a fixed pair of actions leaves for a gate and how much it misses altogether. One confidence set can resolve that limit even when it cannot certify positive gain, as shown by nine comparisons. The bound can exclude a desired gain from gate-only changes. Independent MedMCQA tests show why an apparently helpful intervention still needs the strong control that isolates its added value.

The scope is deliberately local. Budgets do not identify the cheapest repair or guarantee a replacement's gain. Public-state frontiers need not be computationally attainable. The nine comparisons cover two domains; they do not establish broad replication. Enlarging the action family changes the Live diagnosis, so the family itself is part of the reported result. Prospective validation of diagnosis-guided intervention choice remains open. The authors managed MedMCQA splits and label access on public benchmark data; these were not independently administered hidden leaderboard tests. Exact inference depends on the declared sampling law; distribution shift requires new labeled checks or a justified transfer bound. Within these limits, the audit gives a concrete next step when a gain test alone gives no direction: measure what the current actions miss, and test a proposed repair against the strong control that could explain its apparent benefit.

AI USE STATEMENT

The authors contributed to drafting and revising the manuscript and made the final decisions about study scope and claims. Generative AI tools assisted methodological discussion, mathematical derivations and proof development, experimental design, implementation, literature search, result interpretation, and drafting and editing. Checks included explicit derivations, deterministic result replays, software tests, and comparison with primary references. AI-generated prose was revised during manuscript development; it was not a source of experimental evidence. The authors take responsibility for the final text, mathematical claims, code and reported results.

ETHICS STATEMENT

This study evaluates model-generated responses on public research benchmarks and introduces no human-subject intervention or human annotation. LiveAoPSBench contains online olympiad problems; we use the released benchmark under its documented research protocol, retain benchmark identifiers only for reproducibility, and do not attempt to identify problem authors or platform users. The main foreseeable harm is overclaiming the reliability of automated selection or LLM-based verification. We mitigate it by separating public mechanism metrics from hidden endpoints, reporting the pre-specified null and boundary routes, and making every certificate conditional on its coverage premise.

REPRODUCIBILITY STATEMENT

Appendix I specifies the pools, public inputs, model versions, seeds, tie rules, statistical families and study records. The anonymous supplement provides pinned dependencies, a file manifest and one `REPRODUCE.py` entry point. It checks the numerical claims, finite-law examples, released LiveAoPS/FEVEROUS projections, MedMCQA development and confirmation panels, and the recorded stopping-cost results. It also recreates the numerical presentation figures from their saved evidence.

For a short entry point, `audit/diagnostic_walkthrough.py` recomputes the nine joint confidence bounds and checks the matched-summary construction and complete-family table. The companion `audit/family_endpoint_audit.py` checks all 26 reference intervals and the historical final-answer counts. These secondary calculations use already released evidence. A separate replay covers the partial-label FEVEROUS study. Source snapshots and protocols document the original studies.

The package reproduces arithmetic and decision consistency. It does not regenerate model traces, repeat original label joins or remeasure latency. Raw benchmark prompts, candidate text, private gold, model weights and the third-party LiveAoPS grader are excluded. Hashes and derivation scripts link the packaged evidence to the research archive. The default replay makes no network or model calls and opens no sealed role.

REFERENCES

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=h-fLVCI1stW.

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

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005. doi: 10.1137/S1052623401399903. URL https://doi.org/10.1137/S1052623401399903.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30(1):160–201, 2002. doi: 10.1214/aos/1015362189. URL https://doi.org/10.1214/aos/1015362189.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Yegor Denisov-Blanch, Joshua Kazdan, Jessica Chudnovsky, Rylan Schaeffer, Sheng Guan, Soji Adeshina, and Sanmi Koyejo. Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness. *arXiv preprint arXiv:2603.06612*, 2026. URL https://arxiv.org/abs/2603.06612.

Qiwei Di, Kaixuan Ji, Xuheng Li, Heyang Zhao, and Quanquan Gu. Best-of-majority: Minimax-optimal strategy for pass@k inference scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a6CVQpjbXq.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Michael Fay. *Exact Conditional Tests for 2 by 2 Tables of Count Data*. CRAN, 2026. URL https://search.r-project.org/CRAN/refmans/exact2x2/html/exact2x2.html. exact2x2 package reference, version 1.6.9; accessed September 6, 2026.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. URL https://www.jmlr.org/papers/v24/21-0048.html.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models, 2024. URL https://arxiv.org/abs/2410.07985.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't Pass@k: A bayesian framework for large language model evaluation. In *International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2510.04265. URL https://arxiv.org/abs/2510.04265.

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

Jie Hu. Oracle gap and signal fidelity: A fixed-pool diagnostic for test-time collaboration. *arXiv preprint arXiv:2607.17531*, 2026. URL https://arxiv.org/abs/2607.17531.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024. doi: 10.48550/arXiv.2403.12031. URL https://arxiv.org/abs/2403.12031.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25c.html.

Bo Jiang. DiscoUQ: Structured disagreement analysis for uncertainty quantification in LLM agent ensembles. *arXiv preprint arXiv:2603.20975*, 2026. URL https://arxiv.org/abs/2603.20975.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.18581.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kzYq2hfyHB.

Peng Kuang, Yanli Wang, Xiaoyu Han, Yaowenqi Liu, Kaidi Xu, and Haohan Wang. Optimal aggregation of LLM and PRM signals for efficient test-time scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=x85kiYqL4y.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044. URL https://arxiv.org/abs/1207.6083.

Guannan Lai and Han-Jia Ye. When routing collapses: On the degenerate convergence of LLM routers. *arXiv preprint arXiv:2602.03478*, 2026. doi: 10.48550/arXiv.2602.03478. URL https://arxiv.org/abs/2602.03478.

Ye Li, Anqi Hu, Yuanchang Ye, Shiyan Tong, Zhiyuan Wang, and Bo Fu. Set-valued prediction for large language models with feasibility-aware coverage guarantees. *arXiv preprint arXiv:2603.22966*, 2026. URL https://arxiv.org/abs/2603.22966.

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*, 2024. URL https://arxiv.org/abs/2401.10480.

Hui Liu, Bin Zou, Kecheng Chen, Jie Liu, Wenya Wang, and Haoliang Li. Task-aware LLM routing with multi-level task-profile-guided data synthesis for cold-start scenarios. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22047–22076. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1007. URL https://aclanthology.org/2026.acl-long.1007/.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b LLM surpass 405b LLM? rethinking compute-optimal test-time scaling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CvjX9Lhpze.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023. URL https://arxiv.org/abs/2303.17651.

Sadegh Mahdavi, Muchen Li, Kaiwen Liu, Christos Thrampoulidis, Leonid Sigal, and Renjie Liao. Leveraging online olympiad-level math problems for llms training and contamination-resistant evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42554–42578. PMLR, 2025. URL https://proceedings.mlr.press/v267/mahdavi25a.html.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. doi: 10.1214/009053606000000786. URL https://doi.org/10.1214/009053606000000786.

Sagi Meir, Tommer D. Keidar, Noam Levi, Shlomi Reuveni, and Barak Hirshberg. More bang for the buck: Improving the inference of large language models at a fixed budget using reset and discard (ReD). *arXiv preprint arXiv:2601.21522*, 2026. doi: 10.48550/arXiv.2601.21522. URL https://arxiv.org/abs/2601.21522.

Shuyin Ouyang, Dong Huang, Jingwen Guo, Zeyu Sun, Qihao Zhu, and Jie M. Zhang. Dscodebench: A realistic benchmark for data science code generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32628–32636, 2026. doi: 10.1609/aaai.v40i38.40540. URL https://ojs.aaai.org/index.php/AAAI/article/view/40540.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Amin Rakhsha, Kanika Madan, Tianyu Zhang, Amir-massoud Farahmand, and Amir Khasahmadi. Majority of the bests: Improving best-of-n via bootstrapping. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1268. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/36556567e8437f137da23047309155dd-Abstract-Conference.html.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011. URL https://www.jmlr.org/papers/v12/reid11a.html.

Ibne Farabi Shihab, Abu Sa-Adat Mohamed Moon-Im Al Ahsan, and Md Najmus Swaqeeb. Opportunity is not realizability: Selection-valid diagnostics for multi-llm routing. *arXiv preprint arXiv:2608.08265*, 2026. URL https://arxiv.org/abs/2608.08265.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015a. doi: 10.1609/aaai.v29i1.9541. URL https://ojs.aaai.org/index.php/AAAI/article/view/9541.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015b. URL https://proceedings.mlr.press/v37/thomas15.html.

Måns Thulin. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, 8:817–840, 2014. doi: 10.1214/14-EJS909. URL https://arxiv.org/abs/1303.1288.

Ming-Hua Tsai and Phat Tran. Reward-based online LLM routing via NeuralUCB. *arXiv preprint arXiv:2603.30035*, 2026. doi: 10.48550/arXiv.2603.30035. URL https://arxiv.org/abs/2603.30035.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794. URL https://doi.org/10.1007/b13794.

Jens Tuyls, Dylan J. Foster, Akshay Krishnamurthy, and Jordan T. Ash. Representation-based exploration for language models: From test-time to post-training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=S4PCF1YxoR.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VNckp7JEHn.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. 2023. URL https://arxiv.org/abs/2305.10601.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. URL https://arxiv.org/abs/2309.03882.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

Guide to the appendix. For the core argument, start with the numerical example and evidence map in Appendix A, then consult the definitions and proofs in Appendix B. To assess the experiments, use the final-answer checks in Appendix F and the study decisions in Appendix I. Both distinguish the measured endpoint from the claim it supports.

The remaining material supports implementation and reproduction. Appendix C begins the selector and target details; Appendix J collects auxiliary domains, including code generation. Appendix K covers partial labels and future study design. Appendices L–N give complete development, cost and reasoning records. These supporting results do not add an independently validated diagnosis-guided repair.

A A Numerical Walkthrough and Implementation Choices

The counts below show how the audit reaches a decision. FEVEROUS supplies secondary evidence; the two-cell example is a constructed law. Neither is a new prospective experiment. The supplement's `audit_walkthrough.ipynb` replays those two examples without data access or statistical recomputation.

A.1 Identical summaries do not identify the repair budget

Table 1 can be realized with three candidates. Let $B = \{1\}$, $S = \{2\}$, use a constant public representation Z, and always deploy S. Give correctness vectors $(1,1,0), (0,1,0), (1,0,0), (0,0,1), (0,0,0)$ probabilities $(.58, .04, .02, .16, .20)$ under law A and $(.48, .14, .12, .06, .20)$ under law B. Both are valid probability laws. In each, $\Pr(Y_B = 1) = .60$, $\Pr(Y_S = 1) = .62$, $\Pr(H = 1) = .80$, $O = .20$, and $\Gamma_Z = (R - L)_+ = \Delta = .02$. But $M - L = .14$ under A and $-.06$ under B. Hence no function of (O, Γ_Z, Δ) alone can determine the sign of $M - L$ over all such laws.

This is an illustrative consequence of the decomposition, not a new statistical guarantee. It uses a declared constant signal; richer public information can change the frontier. Nor does it claim that prior methods retaining full paired outcomes lack this information. Their rescue counts distinguish the two laws. The distinction motivates reporting capture and testing its budget ordering rather than reporting only the coarse summaries.

A.2 From four counts to a diagnosis, not a release

Compare frozen Self-Certainty top-4 with DPP on the same 400 FEVEROUS pools of 64 candidates. Four disjoint categories contain 62 missed opportunities (both selections fail but the pool succeeds), 31 rescues, 19 harms and 288 other outcomes. Thus $O = (62+31)/400 = .2325$, $M = 62/400 = .155$, $R = 31/400 = .0775$, and $L = 19/400 = .0475$. For the declared one-cell coarsening Z,

$$\widehat{\Gamma}_Z = (31 - 19)_+/400 = .0300, \qquad \widehat{\mathrm{ov}}_Z = \min(31, 19)/400 = .0475.$$

Always using the proposal gives $\widehat{\Delta} = .03$, hence $O - \Delta = .2025 = .155 + .0475 + 0$ in this empirical law. These sample summaries do not estimate unrestricted Γ_G.

The nine-row secondary analysis uses 27 intervals, each with marginal miscoverage $.05/27$. Rounded lower bounds for missed capture and rescue are .10364 and .04199; the harm upper bound is .09008. The simplex cap $1 - .10364 - .04199 = .85437$ is inactive, so $\bar{l} = .09008$. The joint-set lower margins are therefore

$$\inf(M - L) = .10364 - .09008 = .01356 > 0, \qquad \inf \Delta = .04199 - .09008 = -.04809.$$

The law $(M, R, L, \text{other}) = (.10364, .04199, .09008, .76429)$ witnesses the unresolved gain. Missed capture exceeds all remaining two-action gate gain, yet the proposal is not authorized. This confidence family contains four LiveAoPS seed–policy and five FEVEROUS comparisons. BFCL and HellaSwag have separate release histories.

For contrast, the constructed intervals $M \in [.50, .60]$, $R \in [.10, .20]$, $L \in [.05, .55]$ do need the simplex: it reduces $\bar{l}$ to .40 and turns the capture lower margin from $-.05$ to .10, while the compatible law $(.50, .10, .40, 0)$ still has gain $-.30$. Here the confidence-set geometry changes the decision. It does not tighten the empirical family's bounds.

A.3 Which empirical question does each result answer?

Table 5 separates the endpoint and evidence role of the main results. This matters because the same symbols can be evaluated on retained-set coverage or on a single emitted answer. Changing the endpoint changes the event counts and requires a new ledger. The fit-free budget certificate measures missed capture against the sum of information loss and gate regret. Separating those two terms additionally requires a declared public representation and evidence for its frontier.

Table 5: Evidence map. Each row identifies what was measured and which stronger conclusion remains unavailable. Historical and secondary evidence does not inherit a later study's prospective status.

Study	Endpoint	Evidence and permitted interpretation
Nine paired audits	Retained-set success	Joint bounds on M, R, L resolve $M > L$, while gain remains uncertified. They do not separately identify overlap and gate regret.
Complete-family check	Retained-set success	Released row projections identify family misses and oracle budgets. A secondary simultaneous sensitivity covers all references; the observed-best-reference orderings remain unresolved.
Historical endpoint bridge	Final answer	Paired head outcomes measure accuracy gains. The FEVEROUS ledger also reconstructs O, R, L, M directly. These descriptive checks do not validate a successful repair.
MedMCQA gate	Final answer	Fresh calibration/test supports gain over direct answering, but not over unchanged plurality. Two-cell frontier and loss values on test are plug-in descriptions, not population identification.
Reasoning confirmation	Final answer	Newly allocated confirmation tests changed generation. The ordinary-reasoning comparison fails, closing the secondary family. It does not test the fixed-proposal overlap identity.

For a concrete use of the budget, specify the deployed reference, the actions a proposed repair may choose, the endpoint and the desired extra gain. For the two-action example above, every such gate

adds at most $\bar{l} = .09008$ on the same joint confidence event. Targets above this bound require a change outside that action pair. This is an exclusion statement, available for any target from the same upper bound; it does not select a new action or establish that a target inside the bound is attainable. Larger action families need their own budget, as Table 3 demonstrates.

A.4 Public information, its representation, and the gate

Consider a constructed law normalized to 1,000 expected units. A binary feature g, available before outcomes, gives the following counts:

Public feature	Units	Misses	Rescues	Harms	Other
$g = 1$	600	40	110	20	430
$g = 0$	400	40	10	30	320
Total	1,000	80	120	50	750

Let the public record be $G = g$. A constant representation Z_0 ignores this feature: its frontier is $(120 - 50)/1000 = .07$ and overlap .05. The representation $Z_1 = g$ retains it: its frontier is $(110-20)_+/1000+(10-30)_+/1000 = .09$ and overlap $(20+10)/1000 = .03$. Always proposing still gives gain .07, so $.20 - .07 = .08 + .03 + .02$. The optimal public gate proposes only when $g = 1$, recovering the .02 gate regret but none of the .08 missed capture. The information G is unchanged; only Z and the gate change. An invertible encoding of g also preserves its frontier. Learning g from the evaluated outcomes would define a different example. Freeze extraction, bins, ties and missing/unseen-cell fallback before calibration. The MedMCQA eight-choice record and high-consensus indicator illustrate the same distinction.

A.5 Choosing a certificate and deciding when it expires

Calibrate one frozen gate's paired contrast directly. Its two CP rate bounds need neither an activity envelope nor a partition-search penalty. On the existing synthetic calibration counts $(n, r, l) = (800, 47, 2)$, the sample mean is .05625. At the same 95% confidence level, the CP gain lower bound is .0344816 (release), while bounded-loss Hoeffding gives $-.0302909$ (retain baseline). This compares precision for one estimand; it does not show general superiority over conformal risk control (CRC). Under paired loss $(1 - X)/2$, removing a rescue raises loss, while removing a harm lowers it. A valid monotone CRC envelope can discard rescue credit. Appendix K.2 gives the construction and its marginal-risk guarantee.

For the frontier-resolution theorem, the conservative envelope $\bar{\nu} = 1$ always works. An independent pilot with k active units out of m can instead supply the one-sided CP upper endpoint $\bar{\nu} = F^{-1}_{\mathrm{Beta}(k+1,m-k)}(1 - \delta_p)$ for $k < m$, and 1 for $k = m$. Give the main procedure error budget δ_m; the combined guarantee has failure probability at most $\delta_p + \delta_m$. The pilot needs extra protection to also select the proposal. Appendix B.9 proves this rule and bounds underestimation.

For a fixed always-propose contrast, suppose the joint categorical law of misses, rescues, harms and other outcomes has justified deployment shift $\mathrm{TV}(P, Q) \le \varepsilon$. The indicator difference $\mathbf{1}_M - \mathbf{1}_L$ lies in $[-1, 1]$. Thus a calibration-law capture margin c transfers as $c - 2\varepsilon$, just as a paired gain bound does. The displayed FEVEROUS margin .01356 remains positive for $\varepsilon < .00678$, a sensitivity threshold, not an estimated shift. Include the shift bound's own failure probability in the joint error budget. A bound on (Z, D) alone is insufficient: when $D = 0$, full-pool success still determines missed capture. Stable public features do not bound this hidden shift. Without a justified bound, renewal needs fresh labels and a prespecified repeated-testing budget. An unlabeled alarm cannot renew a certificate.

Richer decisions. For bounded utility D, replace rescue/harm probabilities by conditional magnitudes $r_G = \mathbb{E}[D_+ \mid G]$ and $l_G = \mathbb{E}[(-D)_+ \mid G]$: the overlap is $\mathbb{E}\min(r_G, l_G)$. For several actions, the exact information gap is $\mathbb{E}\max_j D_j - \mathbb{E}\max_j \mathbb{E}[D_j \mid G]$, including baseline $D_0 = 0$; it is not a sum of binary overlaps. Both extensions fix the actions and utility. Appendix B.2 gives the finite-action certificate; Appendix B.9 gives a cost-aware example.

B PROOFS AND FURTHER PROPERTIES

B.1 NOTATION AND OPERATIONAL PUBLIC STATES

Condition on the frozen route. The record G contains only row-level inputs available to the deployed decision. We use O for solvable opportunity missed by baseline, (R, L) for proposal rescue/harm mass, Γ_G for best G-measurable gain, and Δ_α for the frozen gate's gain. Endpoints never enter G.

Table 6: Operational content of the row-level public state. Global route cards, implementations, and ties are conditioned-on freezes.

Route	Decision-time tuple generating $\mathcal{G}$
Math/Live routes	candidate answer strings/clusters, public verifier scores, and row-level semantic and structured-risk matrices
BFCL	candidate function-call strings, prompt/view identifiers, endpoint-free score summaries, and the row's frozen movement statistic
FEVEROUS	candidate public outputs, full-vocabulary Self-Certainty, cited-premise NLI summaries, and the frozen SC-Guard cell
HellaSwag	the $K = 8$ candidate record, eight-view identifiers and scores, conditional-choice probabilities, and full-vocabulary scores
MedMCQA	eight parsed semantic choices and public tie anchors; allocation-group identifier is stored outside the gate-visible tuple as inaccessible provenance metadata; admission uses the trivial one-cell coarsening in Stage-55 and fixed high-consensus Z_1 in the separate successor

Every row excludes correctness vectors, references, execution results, answer keys, private grader outputs, and calibration/test labels. Features computed after joining endpoints define a different information set and cannot upgrade the protected claim. A deterministic recoding T satisfying $\sigma(T(G)) = \sigma(G)$ preserves Γ_G; rounding, clipping or dropping fields cannot increase the population frontier. Representation can still affect an implemented gate. Freeze serialization and missing-value rules before their authorized role.

For a calibrated aggregation head, condition on its frozen calibration artifact. Its record G contains only the answer strings, multiplicities, and public verifier/PRM scores used at deployment. Calibration and endpoint labels remain outside G.

B.2 ESTIMATING, REFINING, AND MONITORING THE PUBLIC FRONTIER

The population frontier need not be computationally attainable. If correctness is a measurable deterministic function of the complete public record, then $\Gamma_G = R$, even if that function is infeasible to compute. Excluding a label field therefore need not create population overlap. The finite audit identifies Γ_Z, not unrestricted Γ_G. Calibrating one frozen gate requires neither estimate. For one frozen proposal and $Z \in [B]$, let

$$q_z^+ = \Pr(Z = z, D = 1), \qquad q_z^- = \Pr(Z = z, D = -1).$$

Then the directly auditable recipe is

$$R = \sum_z q_z^+, \quad L = \sum_z q_z^-, \quad \Gamma_Z = \sum_z (q_z^+ - q_z^-)_+, \quad \mathrm{ov}_Z = \sum_z \min(q_z^+, q_z^-). \tag{P}$$

Convexity biases the plug-in frontier upward; concavity biases plug-in overlap downward. A higher plug-in frontier after splitting cells therefore need not mean the representation improved.

Finite-family selection and frontier bounds. For actions $D_j = Y_{S_j} - Y_B$, including baseline $D_0 = 0$, define $\Gamma_h = \mathbb{E}[\max_{0 \le j \le J} \mathbb{E}[D_j \mid Z_h]]$. For the finite-family oracle guarantee, discovery selects from prefrozen $Z_h = h(G)$, with $(J+1)^{B_h}$ cellwise policies. A weighted finite-class lower bound selects one representation and policy while paying $\epsilon_h = \sqrt{2\{B_h \log(J+1) + \log(2/(\delta w_h))\}/n}$, where $w_h > 0, \sum_h w_h \le 1$; simultaneously,

$$V_{\widehat{h}}(\widehat{\pi}_{\widehat{h}}) \ge \max\{0, \max_h (\Gamma_h - 2\epsilon_h)\}. \tag{J}$$

Independent calibration alone authorizes it. For a frozen B_Z-cell, J-action design, simultaneous one-sided Clopper–Pearson bounds on the joint rescue/harm masses give

$$L_\Gamma := \sum_z [\max_j \underline{v}_{zj}]_+ \le \Gamma_{\sigma(Z),\mathcal{A}} \le \min\{1, \sum_z [\max_j \overline{v}_{zj}]_+\} =: U_\Gamma \tag{L}$$

with probability at least $1-\delta$. The lower-bound router gains at least L_Γ, has regret at most $U_\Gamma - L_\Gamma$, and needs no positive minimum cell count. The unrestricted Γ_G remains a population ceiling; the audited frontier here is for a discovery-frozen coarsening Z. An in-sample frontier cannot authorize deployment. The construction and proofs follow below and in Appendix B.6.

Construction checklist. Before endpoint access, freeze: (i) the fields actually available before the decision; (ii) deterministic serialization, score normalization, model versions, missing/unseen-state handling, and ties; (iii) the distinction between the full record G and audited coarsening $Z = h(G)$; (iv) a small nested family $Z_0 \preceq \cdots \preceq Z_H$, its cell counts, proposal family, and error weights; and (v) the one-way rule by which discovery chooses one representation and calibration only authorizes it. Missing or unseen states use a frozen catch-all cell that returns baseline. Noise alone does not invalidate the estimand. Changing how proxies are measured defines a new G and requires a new freeze.

Learned trees and continuous signals. For fixed-gate release, fit and prune a tree or score model on discovery. Freeze one winner, feature extraction, ties and baseline fallback before independent calibration. Bounding its gain needs no discretization. Frontier comparison via (J) instead requires an internal discovery split: train candidate trees, freeze a small nested pruning family and weights, then compare penalized values on independent selection data. Conditional on training, (J) applies with n equal to the selection size. Leaf count alone does not account for arbitrary tree search on the same data. Calibration tests the selected gate or budgets all reported frontiers. Merge sparse leaves on discovery or let them return baseline; calibration cannot choose boundaries. This recipe is not a completed real-LLM tree ablation.

Select a representation by Equation (J)'s lower frontier to penalize overfitting. Independent calibration uses Equation (L). Plotting $[L_{\Gamma,h}, U_{\Gamma,h}]$ against B_h shows when uncertainty outweighs estimated refinement value. Freeze h before seeing calibration labels. Selecting cellwise actions on calibration instead requires a simultaneous budget for all BJ choices. These are separate protocols.

Figure 3 shows this tradeoff under a known synthetic law. The population frontier saturates at eight cells, but the discovery plug-in keeps rising. The simultaneous held-out lower envelope peaks at eight and then falls; the overlap interval widens. This diagnostic neither supplies LLM evidence nor permits calibration tuning. Discovery still freezes h; calibration only verifies it.

Overlap can be reported on the same simultaneous event. For a fixed proposal, give both tails of all $2B$ joint Bernoulli masses simultaneous Clopper–Pearson bounds $q_z^\pm \in [\ell_z^\pm, u_z^\pm]$, allocating $\delta/(4B)$ per tail. Besides Equation (L), this gives

$$\sum_z \min(\ell_z^+, \ell_z^-) \le \mathrm{ov}_Z \le \sum_z \min(u_z^+, u_z^-). \tag{P′}$$

Equivalently, if $R \in [L_R, U_R]$ and $\Gamma_Z \in [L_\Gamma, U_\Gamma]$, then

$$\mathrm{ov}_Z = R - \Gamma_Z \in \big[\max\{0, L_R - U_\Gamma\},\ U_R - L_\Gamma\big].$$

For a nested refinement, the population value is $\Gamma_{Z_{h+1}} - \Gamma_{Z_h} = \mathrm{ov}_{Z_h} - \mathrm{ov}_{Z_{h+1}}$. A split-valid feature ablation needs simultaneous frontier intervals or another reserved role. Choosing the larger frontier on the same labeled sample is only an in-sample oracle diagnostic. Empty cells get vacuous bounds, so unsupported strata cannot reward refinement.

Unlabeled monitoring can trigger fallback, but cannot renew a certificate. Prespecified checks may detect public-support changes. Stable $P(G)$, however, need not imply stable $P(D \mid G)$. Renewal needs fresh labeled audit data that were not used to refit the rule. If a domain argument supplies an upper confidence bound $\mathrm{TV}(P_{G,D}, Q_{G,D}) \le \bar{\varepsilon}$, the transferred gain lower bound is reduced by $2\bar{\varepsilon}$, with the shift-bound failure probability included in the family budget. Without labeled renewal or a justified joint-law bound, detected shift requires refusal; importance weighting alone is insufficient. For prespecified labeled monitoring looks t, keep serializer, representation, and gate fixed and allocate

$\sum_t \delta_t \le \delta_{\text{mon}}$; a union bound then makes the fixed-look certificates simultaneous, and a nonpositive lower bound rolls back to baseline. The guarantee excludes unbudgeted repeated looks, accepted-only labels and unspecified dependence. For known clusters, use independent clusters and bounded cluster-level contrasts as the sampling units.

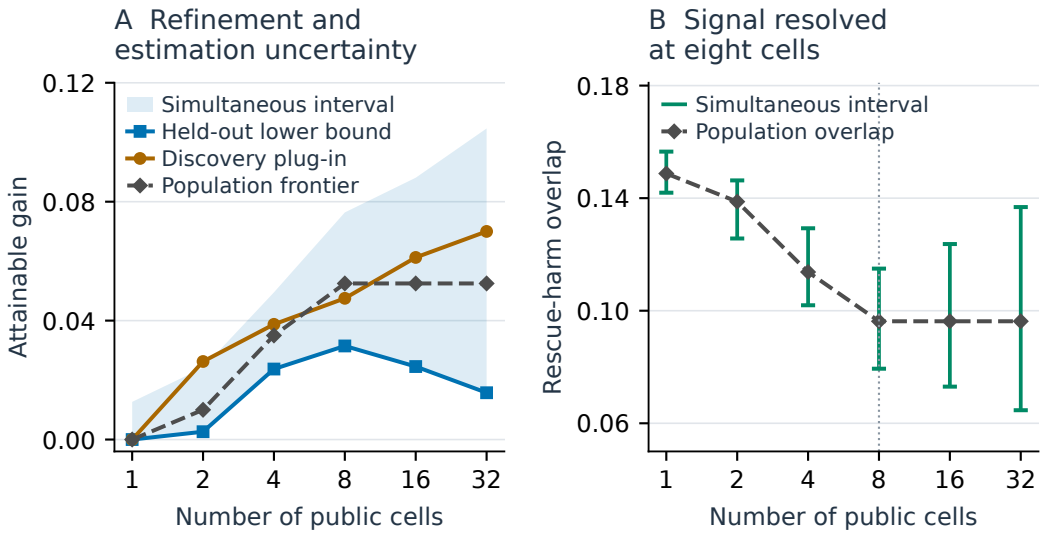

Figure 3: Finite public-state resolution versus uncertainty. Partitions $B = 1, 2, 4, 8, 16, 32$ are nested; discovery ($n_D = 800$) and calibration ($n_C = 30{,}000$) are independent; and the displayed Clopper–Pearson envelopes are simultaneous over every prefrozen partition, cell, sign, and tail (family $\alpha = .05$). After the eight-cell population signal is resolved, finer plug-in frontiers are optimistic while held-out uncertainty grows.

B.3 Exact alternatives and rare-event logging

Conditional on discovery, a deterministic frozen gate produces $X_i = \widehat{\alpha}(G_i)D_i \in \{-1, 0, 1\}$ for binary endpoints. For r positive and l negative contrasts among N independent same-law units, two one-sided CP bounds with tails $\delta/2$ give $L_\Delta = \text{CP}_L(r, N; \delta/2) - \text{CP}_U(l, N; \delta/2)$. Joint coverage is at least $1 - \delta$, both conditional on discovery and unconditionally. Gate complexity adds no factor. For a fractional gate, use a bounded-mean bound on αD, or realized independent Bernoulli gate coins to retain ternary contrasts. For continuous utilities, sign frequencies do not identify mean gain: use the utility magnitudes and Equation (K).

If a new study needs only a sign test, set $k = r + l$. Under $\Delta \le 0$, rescue orientation is at most $1/2$, so $p_{\text{cond}} = \Pr\{\text{Bin}(k, 1/2) \ge r\}$ is valid (set $p_{\text{cond}} = 1$ at $k = 0$). This needs no separate activity bound, but gives no absolute gain lower bound. HellaSwag's published $1/0$ gives .5: sparse discordance still limits inference. Joint exact multinomial inversion is another prefrozen option; we did not evaluate it. Ordinary Wilson/Jeffreys intervals or mid-p modifications can be shorter without uniformly retaining the exact frequentist guarantee (Brown et al., 2002; Thulin, 2014; Fay, 2026). Bonferroni cannot repair invalid base intervals. All protected certificates remain unchanged.

Focused logging may save labels when a public event A structurally contains every nonzero X. For example, if A marks a departure from baseline, a deterministic endpoint gives $X = 0$ on A^c. Representative i.i.d. logging estimates $\Pr(A)$, while independent within-A sampling estimates $\mathbb{E}[X \mid A]$; their product is target-law gain. Positive conditional gain requires positive stratum probability. A magnitude certificate must propagate both uncertainties. Other enrichment changes the target law and requires known positive sampling probabilities plus a valid weighted or stratified bound. Unweighted CP is insufficient. This design is unrun; the existing rare-event nulls remain unchanged.

B.4 Calibration resolution under observed-rate replication

How much calibration would a refused gate need if its observed event mix repeated? For cell size n, rescues r, harms l, and per-bound miscoverage a, define

$$\mathrm{LCB}(n, r, l; a) = F^{-1}_{\mathrm{Beta}(r, n-r+1)}(a) - F^{-1}_{\mathrm{Beta}(l+1, n-l)}(1-a),$$

with the usual zero-success lower and all-success upper boundary conventions. Table 7 exactly replicates each observed calibration block t times, preserving cell prevalence and event rates, and reports the smallest integer t giving a strict positive LCB.

Table 7: Observed-rate calibration resolution. These deterministic replications are planning diagnostics, not power guarantees.

Route	Current LCB	First multiplier	Replicated LCB
BFCL high cell	-.017827	2×	+.007218
FEVEROUS eligible	-.009679	2×	+.000756
HellaSwag router	-.001308	9×	+.0000169

BFCL retains its six-bound family; FEVEROUS and HellaSwag retain their two-bound rules. These calculations repeat observed rates exactly. Future random counts may still fail the rule. HellaSwag barely crosses zero even after ninefold replication, so more labels alone do not give a robust plan. Without cell-prevalence and rounding rules, no unique unrestricted minimum N exists.

B.5 Comparison map: estimand, feedback, and guarantee

Selective prediction and conformal risk control concern risk–coverage or abstention (El-Yaniv & Wiener, 2010; Franc et al., 2023; Angelopoulos et al., 2024). We score both actions per unit and estimate paired incremental utility. CRC's marginal-risk guarantee and monotonicity conditions differ from this high-probability gain event (Appendix K.2). Paired scoring also avoids the counterfactual-support problem of HCOPE/HCPI (Thomas et al., 2015a;b). RouterBench, learned routers, Best-of-Majority, and calibrated voting optimize different policies or cost–quality tradeoffs (Hu et al., 2024; Lai & Ye, 2026; Liu et al., 2026; Tsai & Tran, 2026; Di et al., 2026; Kuang et al., 2026); our audit evaluates a rule after its pool, proposal and public state are frozen. Correlated-error studies motivate the distinction (Kim et al., 2025; Denisov-Blanch et al., 2026). Table 8 compares estimands, not benchmark performance. It shows what each paradigm changes, its evidence, and its relation to Γ_G. These frameworks are complementary (Di et al., 2026; Kuang et al., 2026; Denisov-Blanch et al., 2026; Wu et al., 2025; Liu et al., 2025; Hu et al., 2024; Lai & Ye, 2026; Liu et al., 2026; Tsai & Tran, 2026; El-Yaniv & Wiener, 2010; Angelopoulos et al., 2024; Thomas et al., 2015a;b; Hu, 2026; Shihab et al., 2026). Aggregate AIQ or cost–quality scores do not identify paired (R, L) or Γ_G. A direct comparison needs frozen per-example actions, an independent timestamp/role split, and paired outcomes. Aggregate router scores cannot reconstruct them.

Feasibility-aware set-valued prediction studies the finite-sampling failure floor $a_l = \Pr(H = 0)$ (Li et al., 2026). For a baseline retained from that same pool, $Y_B \le H$, so our opportunity is $O = 1 - a_l - \Pr(Y_B = 1)$, not a_l or $1 - a_l$ alone. This floor measures availability, without fixing rescue mass or identifying paired conversion. We do not use an estimated minimum risk level to certify O. DiscoUQ studies linguistic and embedding-based disagreement features for confidence prediction (Jiang, 2026). These features can enter G if available before the decision and fitted without calibration/test labels. For a fixed proposal, adding them can change overlap. Using them to change the proposal also changes R. Correctness AUROC or ECE alone identifies neither paired quantity. We describe conceptual interfaces here, not implemented comparisons.

Posterior evaluation and allocation-aware coverage. Bayes@N uses a Dirichlet–categorical model to estimate posterior capability and uncertainty for weighted rubrics (Hariri et al., 2026). Its credible intervals are conditional on that model and prior; they are not a frequentist simultaneous upper bound for paired discordance $\nu = \Pr(D \neq 0)$. Reset-and-Discard reallocates attempts among questions and removes solved instances to improve coverage@cost (Meir et al., 2026). This

changes the pool law, which our refinement comparison fixes. Both approaches may inform study design, but neither supplies paired rescue/harm or transports a gain certificate. Fit a learned router on discovery, then freeze its feasible actions, budget, baseline, utility, public inputs and rule. Independent calibration evaluates one paired contrast. Inference over representations is a separate task. We have not prospectively tested an enlarged family; LiveAoPS's expansion was post-endpoint.

Table 8: Adjacent objectives and the role of the public-information frontier.

Paradigm	What changes	Native evidence	Relation to Γ_G
Hu: fixed-pool diagnostic	selector/verifier on a fixed candidate pool	recoverable mass, signal coverage/quality/fidelity, and harm	closest empirical predecessor; no Bayes public-state frontier, whereas this audit further separates frozen-proposal capture from public overlap/admission
Shihab et al.: realizability ladder	model action from a declared pre-answer signal	outcome-oracle, Bayes-signal, and held-out-router gain with selection-valid family intervals	closest formal predecessor; already has a signal ceiling and family selection, while this audit separates losses for one proposal and tests missed capture against the remaining gate budget
Compute-optimal sampling/search	model, budget, generation/search, hence pool law	accuracy–compute or Pareto trade-off	changes O upstream; freeze each configuration before defining its own Γ_G
BoM / calibrated PRM voting	selected-set or aggregation proposal on a realized pool	reward-model regret or held-out accuracy/efficiency	becomes a frozen proposal; audit recomputes R, L, Γ_G, Δ and certifies release
Consensus / correlated-error studies	crowd composition or internal aggregation signal	accuracy and dependence diagnostics	correlation may affect O or overlap, but does not identify $R - \Gamma_G$
Learned model routers	model action under cost/task/user signals	utility, AIQ/Pareto, or online reward	supplies a proposal and G; aggregate utility cannot recover paired R, L, Γ_G
Selective prediction / CRC	accept or abstain using labeled calibration	risk–coverage or monotone-risk control	paired risk below $1/2$ means gain, but native CRC needs pointwise-monotone losses and gives marginal risk control, not capture/overlap allocation
HCOPE / HCPI	target policy from logged action–reward data	high-confidence value or policy improvement	solves partial-feedback support; paired endpoints need no IPS but add conversion accounting
This audit	frozen pool, proposal, public state, baseline, and target law	sharp $O, R, \Gamma_G, \Delta_\alpha$ ledger plus fail-closed admission	localizes capture, public overlap, and target/admission loss

B.6 Public-information frontier proofs

For any $\mathcal{G}$-measurable $\alpha \in [0, 1]$, the tower property gives

$$\mathbb{E}[\alpha(G)D] = \mathbb{E}[\alpha(G)\mathbb{E}[D \mid G]] = \mathbb{E}[\alpha(G)\eta(G)].$$

The integrand is maximized pointwise by admitting when $\eta > 0$, proving Equation (F). Convexity of $z \mapsto z_+$ gives

$$\eta(G)_+ = \big(\mathbb{E}[D \mid G]\big)_+ \le \mathbb{E}[D_+ \mid G].$$

Taking expectations yields $\Gamma_{\mathcal{G}} \le \mathbb{E}[D_+] = R$, and a same-pool rescue lies inside baseline-missed solvable opportunity, so $R \le O$. If $\mathcal{H} \subseteq \mathcal{G}$, then $\eta_{\mathcal{H}} = \mathbb{E}[\eta_{\mathcal{G}} \mid \mathcal{H}]$; conditional Jensen gives $\mathbb{E}[(\eta_{\mathcal{H}})_+] \le \mathbb{E}[(\eta_{\mathcal{G}})_+]$, the decision-specific Blackwell/Bayes-risk comparison (Blackwell, 1953; Reid & Williamson, 2011). All bounds are sharp: equality with R holds when the sign of D is public, while equality with zero occurs when every public-state conditional gain is nonpositive.

For the three-loss decomposition, each integrand $\eta_+ - \alpha\eta$ is nonnegative: it equals $(1-\alpha)\eta$ when $\eta > 0$, $-\alpha\eta$ when $\eta < 0$, and zero otherwise. Hence $\Gamma_{\mathcal{G}} - \Delta_\alpha \ge 0$; adding and subtracting R and $\Gamma_{\mathcal{G}}$ in $O - \Delta_\alpha$ proves Equation (I). The constant reject rule and the reported fixed gate are both $\mathcal{G}$-measurable, so the frontier is at least $\max\{0, \Delta_\alpha\}$; Equation (F2) gives the upper bound R. Substitution proves all endpoints in (I′). Since $B_\alpha = R - \Delta_\alpha$, the upper endpoints reduce to $R - \max\{\Delta_\alpha, 0\} = \min\{R, B_\alpha\}$ and B_α, while the second lower endpoint is $\max\{B_\alpha - R, 0\}$. This proves the general dominance condition. When $\alpha \equiv 1$, $\Delta_\alpha = \mathbb{E}[D] = R - L$, and only then $B_\alpha = L$.

Worked population accounting. Consider 1,000 expected units in a constructed law: 200 full-pool opportunities, 120 rescues and 50 harms. Without instance state, the frontier is 70 and overlap is 50. A pre-existing public refinement removes 20 overlap units, giving $\Gamma_G = 90$, while the frozen gate still realizes 70. Thus $O - \Delta_\alpha = 130 = 80 + 30 + 20$: capture loses 80, overlap 30, and gate regret 20. Refinement recovers only the removed overlap, leaving the proposal's 80 uncaptured opportunities. These population values are illustrative, not fitted frontiers or release evidence. Figure 1 gives a smaller illustration.

B.7 Sharp family-wise admission and repair certificates

Proposition 3 (Sharp family-wise admission and repair-budget certificates)**.** *Let $\mathcal{J}$ be an endpoint-independent family of $J < \infty$ fixed route–proposal populations, each using its proposal as the always-deploy outer action. Write $M_j = O_j - R_j$ and $B_j = R_j - \Delta_j = L_j = \Pr(D_j = -1)$. If all $3J$ Bernoulli rates (M_j, R_j, B_j) have marginal $1 - \alpha/(3J)$ intervals, then with probability at least $1 - \alpha$, simultaneously for all j,*

$$(O_j - R_j, R_j - \Gamma_j, \Gamma_j - \Delta_j) \in [\ell_{Mj}, u_{Mj}] \times [0, \min\{u_{Rj}, u_{Bj}\}] \times [\max\{0, \ell_{Bj} - u_{Rj}\}, u_{Bj}]. \quad (\mathrm{I}'')$$

Intersect each interval rectangle with the exact categorical simplex to form $\mathcal{C}_j = ([\ell_{Mj}, u_{Mj}] \times [\ell_{Rj}, u_{Rj}] \times [\ell_{Bj}, u_{Bj}]) \cap \{(m, r, b) : m + r + b \le 1\}$, and put $\bar{b}_j = \min\{u_{Bj}, 1 - \ell_{Mj} - \ell_{Rj}\}$. Over $\mathcal{C}_j$, the sharp gain range is $[\ell_{Rj} - \bar{b}_j, \min\{u_{Rj}, 1 - \ell_{Mj} - \ell_{Bj}\} - \ell_{Bj}]$. Thus every compatible law has $\Delta_j > 0$ iff $\ell_{Rj} > \bar{b}_j$, and has $M_j > B_j$ iff $\ell_{Mj} > \bar{b}_j$. If either condition fails, $(\ell_{Mj}, \ell_{Rj}, \bar{b}_j) \in \mathcal{C}_j$ is an adverse witness. The first ordering certifies admission; the second certifies that capture loss exceeds the entire endpoint-oracle downstream repair budget and either downstream loss separately. Both remain valid for any reported $\widehat{j} \in \mathcal{J}$, but not after endpoint-dependent family changes. Exact conditional tests of $R_j > L_j$ and $M_j > L_j$, Holm-adjusted over $2J$ values, give a rectangle-free FWER-controlled layer.

For Proposition 3, a union bound makes all $3J$ base intervals valid on one event of probability at least $1 - \alpha$, with no cross-row independence assumption. Equation (I′) then gives (I″). The three cells are disjoint, so their rates also lie in the simplex. Over its intersection with the interval rectangle, the largest feasible B is $\bar{b} = \min\{u_B, 1 - \ell_M - \ell_R\}$. Therefore $\min(R - B) = \ell_R - \bar{b}$, $\max(R - B) = \min\{u_R, 1 - \ell_M - \ell_B\} - \ell_B$, and $\min(M - B) = \ell_M - \bar{b}$. The stated witness attains both minima when the corresponding ordering fails; any nonnegative four-cell probability vector defines a categorical target law. This proves sharpness relative to the reported confidence polytope. Simultaneity permits substitution of any reported $\widehat{j} \in \mathcal{J}$, but not endpoint-dependent enlargement.

For the direct order layer, condition a multinomial row on the total number of observations in cells A and B. The label within this pair is Bernoulli with probability $\theta_{AB} = A/(A + B)$. Under the composite null $A \le B$, its upper-tail probability is maximized at $\theta_{AB} = 1/2$, which proves validity

of the stated conditional p-value. Holm's step-down rule then controls family-wise error for the $2J$ valid p-values under arbitrary dependence.

For the oracle statement, maximize $\beta(D)D$ pointwise. Taking $\beta(1) = 1$, $\beta(-1) = 0$, and either action at zero attains $\mathbb{E}[D_+] = R$, and no rule can exceed it. The always-deploy outer action has value $\mathbb{E}[D] = R - L$, so its maximum increment under any endpoint-informed downstream repair is exactly $B = R - \Delta = L$. Hence $M > B$ means the capture loss exceeds the entire oracle repair budget while holding proposal and baseline fixed.

B.8 Portfolio and split-valid calibration bridges

For the finite-family frontier above, condition on G. The router's conditional gain is linear in the simplex weights $\pi_a(G)$, so a maximizer chooses the action with largest positive $\eta_a(G) = \mathbb{E}[D_a \mid G]$, or the baseline if every value is nonpositive. Integration gives $\Gamma_{\mathcal{G},\mathcal{A}}$. With the endpoint vector revealed, $\max\{0, \max_a D_a\} = 1$ exactly when the baseline fails and at least one family proposal succeeds, giving $R_{\mathcal{A}}$. Adding and subtracting these two frontiers proves the decomposition; nonnegativity follows from nested decision classes and $R_{\mathcal{A}} \le O$. Adding actions enlarges both maximizations, proving monotonicity.

For the exact equality conditions, augment the family by baseline 0, put $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$, $D_0 = 0$, $W = \max_{a \in \mathcal{A}_0} D_a$, $\eta_a = \mathbb{E}[D_a \mid \mathcal{G}]$, and $m = \max_{a \in \mathcal{A}_0} \eta_a$. Then $R_{\mathcal{A}} = \mathbb{E}[W] \ge \Gamma_{\mathcal{G},\mathcal{A}} = \mathbb{E}[m] \ge \Delta_\pi$. The first equality holds iff a $\mathcal{G}$-measurable action is endpoint-oracle throughout each conditional support; the second holds iff $\sum_a \pi_a(G)\{m(G) - \eta_a(G)\} = 0$ almost surely. Finally, $O = R_{\mathcal{A}}$ iff the family captures almost every $\{H = 1, Y_B = 0\}$.

To prove these claims, choose the smallest-index public maximizer $a^*(G) \in \arg\max_a \eta_a(G)$. This is measurable, and

$$\mathbb{E}[W \mid \mathcal{G}] - m(G) = \mathbb{E}[W - D_{a^*(G)} \mid \mathcal{G}] \ge 0.$$

Therefore $R_{\mathcal{A}} = \Gamma_{\mathcal{G},\mathcal{A}}$ if and only if the nonnegative variable $W - D_{a^*(G)}$ is zero almost surely. This proves the common-oracle condition and its converse. Likewise,

$$m(G) - \sum_a \pi_a(G)\eta_a(G) = \sum_a \pi_a(G)\{m(G) - \eta_a(G)\} \ge 0,$$

which vanishes exactly when the router assigns no mass to a strictly suboptimal conditional action. In the binary selected-coverage setting, $W = \mathbf{1}\{Y_B = 0, \max_a Y_a = 1\}$, so $O - R_{\mathcal{A}} = \Pr(H = 1, Y_B = 0, W = 0)$. This proves the capture-equality condition; the full-chain claim follows because all three gaps are nonnegative.

For Proposition 2, define finite measures $\mu^+(E) = \Pr(G \in E, D = 1)$, $\mu^-(E) = \Pr(G \in E, D = -1)$, and let $r = d\mu^+/dP_G$, $l = d\mu^-/dP_G$. Since $\eta = r - l$,

$$R - \Gamma_G = \int \{r - (r - l)_+\}\, dP_G = \int \min(r, l)\, dP_G = \mathrm{ov}_G\,.$$

Also $R - L = \int (r - l) dP_G$ and $V_G = \int |r - l| dP_G$; adding these identities yields $\Gamma_G = (R - L + V_G)/2$. Zero overlap is equivalent to mutual singularity: when $\min(r, l) = 0$ almost surely, the set $\{r > 0\}$ supports μ^+ and its complement supports μ^-, and the reverse implication is immediate. For $Z = h(G)$, pushforward preserves R, L, while $r_Z = \mathbb{E}[r \mid Z]$ and $l_Z = \mathbb{E}[l \mid Z]$. Applying the two identities before and after the pushforward gives $\Gamma_G - \Gamma_Z = \mathrm{ov}_Z - \mathrm{ov}_G = (V_G - V_Z)/2$. This is the decision-specific binary-experiment form of classical variation/Bayes-risk and information-garbling identities (Blackwell, 1953; Reid & Williamson, 2011).

B.9 Bounded utilities and declared-law robustness

Proposition 4 (Weighted positive–negative overlap)**.** *Let all quality, cost, latency, and abstention weights be prefrozen and $D = U_S - U_B \in [-1, 1]$. With $r_u(G) = \mathbb{E}[D_+ \mid G]$, $l_u(G) = \mathbb{E}[D_- \mid G]$, and $\mathrm{ov}^u_G = \mathbb{E}\min(r_u, l_u)$,*

$$\Gamma^u_G = \sup_{0 \le \alpha(G) \le 1} \mathbb{E}[\alpha(G)D] = \mathbb{E}[(r_u - l_u)_+], \qquad \mathbb{E}[D_+] - \Gamma^u_G = \mathrm{ov}^u_G\,.$$

If $Z = h(G)$, then $\Gamma^u_G - \Gamma^u_Z = \mathrm{ov}^u_Z - \mathrm{ov}^u_G \ge 0$. If a declared opportunity functional satisfies $O_u \ge \mathbb{E}[D_+]$, replacing R by $\mathbb{E}[D_+]$ yields the three-loss decomposition; without that domination, only the two downstream losses are universally nonnegative.

Conditionally on G, the value is $\alpha(G)(r_u - l_u)$, so pointwise positive-part maximization proves the frontier. The identity follows from $r_u - (r_u - l_u)_+ = \min(r_u, l_u)$; conditional Jensen proves refinement monotonicity. For $D \in [-C, C]$ the identities remain exact and concentration radii scale by C. Post-label tuning of utility weights changes the estimand and is not covered.

Cost-aware worked example. Consider two equally likely public cells. Their coverage rescue/harm rates are $(.12, .02)$ and $(.08, .01)$. With declared opportunity $O = .14$, the proposal has $R = .10$, $L = .015$, and the public gate uses it in both cells, so $\Gamma_Z = \Delta = .085$. The coverage decomposition is

$$.055 = .14 - .085 = \underbrace{.040}_{O-R} + \underbrace{.015}_{R-\Gamma_Z} + \underbrace{0}_{\Gamma_Z-\Delta},$$

so missed capture exceeds the remaining gate budget. Now predeclare normalized action costs .04 and .10 in the two cells and utility contrast $D_u = .8D_c - c_Z \in [-1, 1]$. Conditional utility gains become .04 and $-.044$. The weighted positive mass is $R_u = \mathbb{E}[(D_u)_+] = .0736$, while $\Gamma_Z^u = .0200$ and $\mathrm{ov}_Z^u = .0536$. If the coverage gate still deploys in both cells, $\Delta_u = -.002$. For a prefrozen broader opportunity family with $O_u = .10$,

$$.102 = .10 - (-.002) = \underbrace{.0264}_{O_u-R_u} + \underbrace{.0536}_{R_u-\Gamma_Z^u} + \underbrace{.0220}_{\Gamma_Z^u-\Delta_u}.$$

The largest gap shifts from capture under coverage to public information under cost-aware utility; gate misorientation adds a further loss. Here R_u is weighted mass, not a probability. Freeze O_u from an action family before labels. Unbounded or unpaired latency measurements do not justify this calculation.

A continuous version adds independent execution noise $E \sim \mathrm{Unif}[-.01, .01]$, unavailable to the gate, and sets $D_u = .8D_c - c_Z - E$. Its support is contained in $[-.91, .77]$. Rescues remain strictly positive and the other contrasts strictly negative; mean-zero noise therefore preserves both positive and negative conditional utility masses. All displayed frontier, overlap, and loss values above are unchanged. The first-cell-only gate has gain $\frac{1}{2}(.8(.12 - .02) - .04) = .020$, but this population value is not a calibration observation. For independent held-out contrasts in $[-1, 1]$, Hoeffding gives $\Pr\{\mathbb{E}X < \overline{X} - \sqrt{2\log(1/\delta)/N}\} \le \delta$, proving the following fixed-gate rule. Freeze utility weights and normalize a known bounded range before calibration. For continuous $D_u \in [-1, 1]$, put $X_i = \widehat{\alpha}(G_i)D_{u,i}$ and use

$$L_u = \overline{X} - \sqrt{2\log(1/\delta)/N}; \qquad \text{release iff } L_u > 0. \tag{K}$$

For dependent prompts within independent, identically distributed clusters, use one prespecified bounded cluster mean as X_i and let N count clusters. This targets mean cluster utility, not automatically prompt-weighted utility. Binomial CP requires binary cluster events, not arbitrary cluster averages. We claim no simulated or real admission from these illustrative values.

Proposition 5 (TV-robust transfer for a fixed gate)**.** *Use* $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. *If calibration law* P *and deployment law* Q *for* (G, D) *satisfy* $\mathrm{TV}(P, Q) \le \varepsilon$*, a gate* $\alpha : G \to [0, 1]$ *is fixed independently of deployment labels, and* $D \in [-1, 1]$*, then*

$$|\mathbb{E}_Q[\alpha D] - \mathbb{E}_P[\alpha D]| \le 2\varepsilon.$$

Hence a valid calibration-law lower bound ℓ_P *transfers as* $\ell_P - 2\varepsilon$*, and positive deployment gain requires* $\ell_P > 2\varepsilon$.

The proof is the sharp bounded-function TV inequality applied to $f = \alpha D \in [-1, 1]$. The same conclusion follows from $\mathrm{TV}(P_G, Q_G) \le \varepsilon$ plus $P(D \mid G) = Q(D \mid G)$. Matching only the covariate marginal gives no guarantee if conditional utility shifts. If ε is estimated, its failure probability must also be charged. This result transfers a fixed-gate gain bound; it does not transport exact CP/McNemar coverage or cover adaptive sampling and feedback.

Split-valid resolution and matching lower bound. The complete statement gives the learning and validation bounds discussed in Section 3.2.

Proposition 6 (Split-valid conversion resolution)**.** *Before the discovery endpoint join, fix a same-pool proposal, a partition* $Z \in [B_Z]$*, and all ties. Put* $D = Y_S - Y_B \in \{-1, 0, 1\}$*,* $\nu = \Pr(D \neq 0) =$

$R + L$, $V(\alpha) = \mathbb{E}[\alpha(Z)D]$, and $\Gamma_Z = \max_{\alpha:[B_Z]\to\{0,1\}} V(\alpha)$. Let $\bar{\nu} \in [\nu, 1]$ be a deterministic prefrozen envelope. On n i.i.d. discovery rows, let $\widehat{\alpha}$ maximize empirical value over the 2^{B_Z} cellwise gates. On a mutually independent calibration role containing N i.i.d. rows from the same target law define

$$x_n = B_Z \log 2 + \log(4/\delta), \quad x_N = \log(4/\delta), \quad e_s = \sqrt{2\bar{\nu}x_s/s} + 2x_s/(3s),\ s \in \{n, N\},$$

and authorize $\widehat{\alpha}$ iff its calibration mean exceeds e_N, otherwise returning baseline. If $\Delta^\dagger$ is final gain, then with probability at least $1 - \delta$,

$$0 \le \Delta^\dagger \le \Gamma_Z, \qquad O - \Delta^\dagger \le (O - R) + \mathrm{ov}_Z + 2e_n + 2e_N. \tag{J′}$$

On the same event, $\Gamma_Z > 2e_n + 2e_N$ forces authorization and $\Delta^\dagger \ge \Gamma_Z - 2e_n > 0$. The bounds give sufficient orders $n = O(\bar{\nu}\{B_Z + \log(1/\delta)\}/\Gamma_Z^2)$ and $N = O(\bar{\nu}\log(1/\delta)/\Gamma_Z^2)$. In the lower-bound construction regime $0 < \Gamma_Z \le \nu/4$, with a tight prefrozen envelope $\bar{\nu} \asymp \nu$, the leading B_Z-dependent discovery order is worst-case necessary for constant-probability $\Gamma_Z - V(\widehat{\alpha}) = o(\Gamma_Z)$ regret. The separate calibration confidence order is necessary for a level-δ, fixed-power test of a supplied oracle gate under calibration-only authority. These necessity statements concern different targets; the discovery statement does not cover authorization of an arbitrarily small positive gain.

For Proposition 6, set $X_\alpha = \alpha(Z)D$ and $q_\alpha = \Pr(X_\alpha \ne 0) \le \nu \le \bar{\nu}$. A direct ternary Bennett calculation gives, for $0 < \lambda < 3$,

$$\log \mathbb{E}e^{\lambda(X_\alpha - \mathbb{E}X_\alpha)} \le q_\alpha(e^\lambda - 1 - \lambda) \le \frac{q_\alpha \lambda^2}{2(1 - \lambda/3)},$$

and the same bound holds for $-X_\alpha$. Thus the displayed e_n (whose linear term is conservative) and a union bound over 2^B gates give, with failure probability at most $\delta/2$,

$$\sup_\alpha |\widehat{V}_n(\alpha) - V(\alpha)| \le e_n,$$

because $2^{B+1}e^{-x_n} = \delta/2$. ERM comparison with a population maximizer gives $\Gamma_Z - V(\widehat{\alpha}) \le 2e_n$. Conditional on discovery, $\widehat{\alpha}$ is fixed; an independent two-sided Bernstein bound gives $|\widehat{V}_N(\widehat{\alpha}) - V(\widehat{\alpha})| \le e_N$ except on an event of probability $\delta/2$. If authorization occurs, its true value is positive and regret is at most $2e_n$. If authorization fails, then $\widehat{V}_N(\widehat{\alpha}) \le e_N$, so $V(\widehat{\alpha}) \le 2e_N$ and $\Gamma_Z \le 2e_n + 2e_N$; baseline has zero gain. Equation (J′) follows from (I) and Proposition 2. The strict margin condition rules out fallback and yields the stated positive lower bound. A data-derived $\bar{\nu}$ would require its own independent validity event; the proposition uses a deterministic prefrozen envelope.

Proposition 7 (Pilot-calibrated activity envelope)**.** *Fix the proposal, endpoint, representation, and gate class, then observe an independent same-law pilot $W_i = \mathbf{1}\{D_i \ne 0\} \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\nu)$, $i = 1, \dots, M$, with $K = \sum_i W_i$. Let*

$$U_\nu(K, M; \delta_p) = \begin{cases} 1, & K = M, \\ F^{-1}_{\mathrm{Beta}(K+1, M-K)}(1 - \delta_p), & K < M. \end{cases}$$

Using $\bar{\nu} = U_\nu$ in Proposition 6 with main error δ_m gives all its conclusions jointly with probability at least $1 - \delta_p - \delta_m$.

One-sided Clopper–Pearson inversion gives $\Pr_\nu\{\nu > U_\nu\} \le \delta_p$. Conditional on the independent pilot and $\nu \le U_\nu$, the envelope is fixed and Proposition 6 fails with probability at most δ_m; a union bound proves the claim. Its leading fixed-margin sample cost relative to known $\nu > 0$ is U_ν/ν, apart from the changed logarithmic error budget. For $K = 0$, $U_\nu = 1 - \delta_p^{1/M} \approx \log(1/\delta_p)/M$. The pilot uses endpoint activity and cannot also select the proposal or representation without independent or simultaneous inference. If its valid upper bound is computationally too loose, the protocol refuses; downward clipping has no nominal guarantee.

Sensitivity to an underestimated activity envelope. Write $e_s(v) = \sqrt{2vx_s/s} + 2x_s/(3s)$. If a justified bound gives $\nu \le \nu_+ = \min\{1, \bar{\nu} + \varepsilon\}$, $\varepsilon \ge 0$, then

$$d_s = e_s(\nu_+) - e_s(\bar{\nu}) = \sqrt{2x_s/s}\,(\sqrt{\nu_+} - \sqrt{\bar{\nu}}) \le \sqrt{2\varepsilon x_s/s}.$$

Substituting ν_+ in the same concentration proof preserves Proposition 6 with radii $e_s + d_s$, provided release requires calibration mean above $e_N + d_N$. Its repair bound and sufficient authorization margin increase by $2d_n + 2d_N$. Keeping the original threshold would only imply $V(\widehat{\alpha}) > -d_N$ upon release on this event, not nonnegative gain. For a multiplicative bound $\nu \le c\bar{\nu}$, $c \ge 1$, the stochastic radius grows by at most $\sqrt{c}$ and the leading fixed-margin sample requirement by at most c. An independent validation event's failure probability must be added. These rules guide design; they cannot validate a failed certificate retrospectively. More observed discordance may warn of change but does not prove it. Stable activity also need not mean stable rescue orientation. A same-law assumption or justified transfer bound remains necessary.

For sharpness, let even $B \ge 4$, Z be uniform, and choose a balanced $\theta \in \{-1, 1\}^B$. For $0 < \gamma \le \nu/4$, set

$$\begin{aligned} &\Pr(D = 1 \mid Z = z) = \frac{\nu + 2\gamma\theta_z}{2}, \quad \Pr(D = -1 \mid Z = z) = \frac{\nu - 2\gamma\theta_z}{2}, \\ &\Pr(D = 0 \mid Z = z) = 1 - \nu. \end{aligned} \tag{J-LB$'$}$$

All hypotheses have $R = L = \nu/2$, $\Gamma_Z = \gamma$, and $R - \Gamma_Z = \nu/2 - \gamma$. One same-pool realization sets $(Y_B, Y_S) = (0, 1)$ for $D = 1$, $(1, 0)$ for $D = -1$, and $(0, 0)$ otherwise; an arbitrary fixed missed-opportunity mark of mass $0 \le m \le 1 - \nu$ can be added inside $D = 0$, preserving the other quantities and setting $O - R = m$. A wrong cell costs $2\gamma/B$, while opposite cell signs have KL at most $16\gamma^2/\nu$. A balanced constant-weight packing and Fano therefore require $n = \Omega(\nu B/\gamma^2)$ for constant-probability $o(\gamma)$ discovery regret. This learning target requires approaching the frontier, rather than merely obtaining some positive value. Even when the oracle gate is given free, a level-δ calibration-only test with fixed power $1 - \beta > 1/2$ and $0 < \delta < 1 - \beta$ requires $N = \Omega\{\nu \,\mathrm{kl}(1 - \beta \| \delta)/\gamma^2\}$ by the corresponding two-point testing bound. These lower bounds hold for the declared one-way protocol; allowing discovery evidence to confer deployment authority would define a different experiment.

For Equation (J), let Π_h be all cellwise maps to baseline 0 and J proposals, with $D_0 = 0$. Write $V_h(\pi) = \mathbb{E}[D_{\pi(Z_h)}]$, let $\widehat{V}_h$ be its discovery mean, and put $\Gamma_h = \max_{\pi \in \Pi_h} V_h(\pi)$. The family and positive weights w_h, with $\sum_h w_h \le 1$, are fixed before these discovery rows. With frozen ties define

$$\widetilde{\pi}_h \in \arg\max_{\pi \in \Pi_h} \widehat{V}_h(\pi), \quad \widehat{L}_h = [\widehat{V}_h(\widetilde{\pi}_h) - \epsilon_h]_+, \quad \widehat{h} \in \arg\max_h \widehat{L}_h.$$

Set $\widehat{\pi}_h = \widetilde{\pi}_h$ if $\widehat{L}_h > 0$, otherwise baseline. Since $|\Pi_h| = (J + 1)^{B_h}$, Hoeffding's inequality for $D_{\pi(Z_h)} \in [-1, 1]$ and a weighted union bound give, with probability at least $1 - \delta$,

$$\sup_{\pi \in \Pi_h} |\widehat{V}_h(\pi) - V_h(\pi)| \le \epsilon_h \quad \text{for every } h.$$

Work on this event. If $\widehat{L}_h > 0$, then $V_h(\widehat{\pi}_h) \ge \widehat{L}_h$, while ERM comparison with a population maximizer gives $\Gamma_h - V_h(\widehat{\pi}_h) \le 2\epsilon_h$. If $\widehat{L}_h = 0$, the output is baseline. Because baseline belongs to Π_h, $\widehat{V}_h(\widetilde{\pi}_h) \ge 0$; ERM comparison and $\widehat{V}_h(\widetilde{\pi}_h) \le \epsilon_h$ then imply $\Gamma_h \le 2\epsilon_h$. Thus the same inequalities hold in both cases. Also $\widehat{L}_h \ge \Gamma_h - 2\epsilon_h$; maximizing these lower bounds proves (J).

The rate is tight for a fixed representation in the bounded-contrast model. There are universal constants $c, c_0 > 0$ such that, for a representation with B equiprobable cells and any discovery learner,

$$\sup_P \Pr_P \left\{ \Gamma_h - V_h(\widehat{\pi}) \ge c \min\left(1, \sqrt{\frac{B \log(J + 1)}{n}}\right) \right\} \ge c_0. \tag{J-LB}$$

For $J \ge 2$, assign each cell a hidden best action; conditional candidate correctness has mean $1/2 + \mu$ for that action and $1/2$ otherwise, with baseline always wrong. A (J)-ary Varshamov packing of the J^B assignments has entropy $\Omega(B \log J)$ and KL divergence $O(n\mu^2)$ from the common center. Generalized Fano for Hamming loss then gives $\Pr\{d_H(\widehat{\theta}, \theta) \ge cB\} \ge c_0$ when $\mu = c \min\{1, \sqrt{B \log J/n}\}$. Each wrong cell incurs weighted regret μ/B, proving (J-LB) for $J \ge 2$. For $J = 1$, encode a rescue/harm sign in each cell with $\Pr(D = 1 \mid z) = 1/4 + \theta_z\mu/2$, $\Pr(D = -1 \mid z) = 1/4 - \theta_z\mu/2$, and apply the binary packing. Assouad's lemma yields the same constant-probability Hamming-loss conclusion with $\mu = c \min\{1, \sqrt{B/n}\}$. A separate two-point experiment at $B = J = 1$, with means $\pm 2\rho$, gives the necessary $\Omega(\log(1/\delta)/\rho^2)$ confidence term for small ρ and δ. Taking the larger of the complexity and confidence requirements is equivalent, within a factor two,

to adding them. Hence fixed- representation sample complexity is $\Theta(\{B\log(J+1)+\log(1/\delta)\}/\rho^2)$, up to constants and the bounded-loss saturation regime. These are standard finite-class ERM and Fano arguments specialized to conversion auditing, not a new general minimax principle (Massart & Nédélec, 2006; Tsybakov, 2009).

Finite-cell inference for continuous utility. Freeze the B-cell representation, J proposals, utility weights and ties before independent calibration. Each i.i.d. unit supplies all paired contrasts $D_{u,ij} \in [-1,1]$. Define

$$m^{\pm}_{zj} = \mathbb{E}[\mathbf{1}\{Z=z\}(D_{u,j})_{\pm}], \qquad \widehat{m}^{\pm}_{zj} = N^{-1}\sum_i \mathbf{1}\{Z_i=z\}(D_{u,ij})_{\pm}.$$

Each summand lies in $[0,1]$. Hoeffding and a union bound over $2BJ$ masses and both tails give, with probability at least $1-\delta$,

$$\max_{z,j,s\in\{+,-\}} |\widehat{m}^{s}_{zj} - m^{s}_{zj}| \le \epsilon_N := \sqrt{\frac{\log(4BJ/\delta)}{2N}}.$$

Clip $\widehat{m}^{\pm} \pm \epsilon_N$ to $[0,1]$, giving $[\ell^{\pm}, u^{\pm}]$, and set $\underline{v} = \ell^+ - u^-$, $\overline{v} = u^+ - \ell^-$. Equation (L) then bounds $\Gamma^{u}_{Z,\mathcal{A}} = \sum_z [\max_j (m^+_{zj} - m^-_{zj})]_+$. The router choosing each cell's largest positive lower endpoint, or baseline if none is positive, gains at least L_Γ, with regret at most $U_\Gamma - L_\Gamma$. For one proposal the same event bounds overlap by

$$\sum_z \min(\ell^+_z, \ell^-_z) \le \mathrm{ov}^u_Z \le \sum_z \min(u^+_z, u^-_z).$$

Dependence among actions on a unit is allowed; independence is across units. For multiple reported frozen partitions, replace δ by δw_h, $\sum_h w_h \le 1$. These are bounded-mean intervals, not binomial CP intervals. Ternary contrasts recover Bernoulli joint masses and their Hoeffding bounds, not CP intervals.

For the transparent un-clipped design envelope, put $v_{zj} = m^+_{zj} - m^-_{zj}$. Then $|\widehat{v}_{zj} - v_{zj}| \le 2\epsilon_N$. The positive maximum is one-Lipschitz in the sup norm, so summing over cells gives $|\widehat{\Gamma} - \Gamma| \le 2B\epsilon_N$. In each cell, choose the empirical maximizing action only when its lower bound $\widehat{v}_{zj} - 2\epsilon_N$ is positive. If admitted, its regret is at most $4\epsilon_N$; if refused, the true cell frontier is at most $4\epsilon_N$. Hence total regret is at most $4B\epsilon_N$, and its gain is strictly positive whenever $\Gamma > 4B\epsilon_N$, equivalently $N > 8B^2\log(4BJ/\delta)/\Gamma^2$. The bound is a transparent design envelope; for binary outcomes, Equation (L) instead uses exact binomial bounds. Calibrating only one discovery-selected gate uses (K), without the frontier's BJ multiplicity factor.

For completeness, an admission budget q gives

$$\Gamma_{\mathcal{G}}(q) = \sup_{0\le\alpha\le 1:\, \mathbb{E}\alpha\le q} \mathbb{E}[\alpha(G)\eta(G)].$$

Suppose an admitted state has lower conditional gain than a non-admitted state. Moving any common amount of admission mass from the former to the latter weakly increases value without changing the budget. Repeating this exchange yields a threshold rule on η, with randomization only at the boundary, with no negative-gain state admitted. For any simultaneous conditional lower surface $\ell \le \eta$, validity implies pointwise $\mathbf{1}\{\ell>0\}\eta \ge \mathbf{1}\{\ell>0\}\ell = \ell_+$; integration gives $\mathbb{E}[\mathbf{1}\{\ell>0\}D] \ge \mathbb{E}[\ell_+]$.

For a separate conditional-cell construction, condition on every observed cell count n_z. Within a frozen cell and action, rescue and harm indicators are Bernoulli under the target conditional law; if $n_z = 0$, use the vacuous Clopper–Pearson interval $[0,1]$. The one-sided Clopper–Pearson bounds therefore miss their respective rates with probability at most $\delta/(2BJ)$. A union bound covers all $2BJ$ rates under arbitrary dependence among actions evaluated on the same rows. On that event,

$$\mathbb{E}[D_j \mid Z=z] = p^R_{zj} - p^L_{zj} \ge \underline{p}^R_{zj} - \overline{p}^L_{zj} = \ell_{zj}.$$

The data-selected action is safe because coverage is simultaneous before the maximization: every admitted cell has nonnegative conditional gain, and a cell without a positive bound returns the baseline. After calibration, the router is fixed and belongs to the public decision class, so $\Delta_{\widehat{\pi}} \le \Gamma_{\sigma(Z),\mathcal{A}}$. The finite-family frontier is at most $R_{\mathcal{A}}$. This is cellwise target-law coverage, not a pointwise claim inside a cell or a guarantee under distribution shift.

For the joint-mass construction in Equation (L), each $\mathbf{1}\{Z = z, D_j = 1\}$ and $\mathbf{1}\{Z = z, D_j = -1\}$ is Bernoulli over the N i.i.d. units. A union bound over their $4BJ$ one-sided lower and upper Clopper–Pearson bounds gives simultaneous coverage. On this event, $\underline{v}_{zj} \le v_{zj} \le \overline{v}_{zj}$, where $v_{zj} = \mathbb{E}[\mathbf{1}\{Z = z\}D_j]$. Since

$$\Gamma_{\sigma(Z),\mathcal{A}} = \sum_z \max\{0, \max_j v_{zj}\},$$

monotonicity of the positive maximum proves (L). Selecting in each cell the largest positive lower endpoint gives gain at least L_Γ; subtracting this from the frontier upper bound proves the regret claim. No division by n_z occurs, so empty observed cells remain valid but uninformative.

For the method-only class, $\Gamma_{\mathcal{G}_0,\mathcal{A}} = [\max_j \sum_z v_{zj}]_+$. Replacing every v_{zj} by its simultaneous lower or upper endpoint gives $[L_0, U_0]$, and the instance-allocation value lies in $[\max(0, L_\Gamma - U_0), \max(0, U_\Gamma - L_0)]$, clipped to $[0, 1]$. The conditional and joint-event constructions each have level $1 - \delta$; claiming them jointly requires sharing the error budget. Finally, a frozen router satisfies $\Delta_{\widehat{\pi}} \le \Gamma \le R_{\mathcal{A}}$, so jointly valid untouched-test bounds on its gain and family capture give the stated outer sandwich.

The executable audit groups paired differences by public state and computes conditional sample means. It sorts these means to solve the budget frontier. Unless estimation and evaluation are separated, the result is marked as an in-sample oracle. The calibration bridge implements both exact split-valid constructions; Equation (L) is the joint-mass version. The anonymous replay checks all 196 selected-set-feasible two-outcome/two-action laws in its grid, 300 seeded laws, the overlap/refinement identities, both Hoeffding constants, and the representation oracle on 81 exact and 100 seeded laws. The aggregate-only BFCL/FEVEROUS replay has SHA-256 `578c8a98497ddd58f96e57a7fcaa4db24c13121fb51312a461af968cd05888cf`. It reads no row-level or protected-test labels and preserves both frozen release decisions.

B.10 Exact opportunity and alignment control

An exact family separates opportunity, public information and gate alignment without model or endpoint calls. Let a fixed proposal have rescue probability $R = .5O$, harm probability $L = .25O$, and tie otherwise. Given $D \in \{-1, 0, 1\}$, a scalar public signal follows $\xi \mid D \sim \mathcal{N}(sD, 1)$. The declared target gate always admits when $\xi > 0$, while the public-information optimum admits when $\mathbb{E}[D \mid \xi] > 0$. Gaussian tail probabilities give both gains exactly.

At $O = .08$, $s = 0$ gives $\Gamma_G = .0200$; with aligned $s = 2$, the frontier is .03873 and the declared gate obtains .03864, near rescue oracle .0400. Reversal keeps $\Gamma_G = .03873$ but changes realized gain to $-.01864$. The losses move from $(.0400, .00127, .00009)$ to $(.0400, .00127, .05737)$: only gate/target regret changes. This exact mechanism control supplies no prospective deployment evidence.

B.11 Why public-only non-regression needs an assumption

Proposition 8 (No-free-lunch over unrestricted endpoint laws)**.** *Fix a public-state realization g and two distinct size-m subsets $A = a(g)$ and $B = b(g)$. In the statistical model class that holds g fixed and imposes no restriction on the conditional endpoint law $\mathcal{L}(C \mid G = g)$, one law has $Y_A > Y_B$ almost surely and another has $Y_A < Y_B$ almost surely. Consequently, no nontrivial public-only departure from B has a selected-coverage non-regression guarantee uniformly over this unrestricted class.*

Because $A \neq B$ and their sizes agree, both $A \setminus B$ and $B \setminus A$ are nonempty. Marking only one candidate in the former correct gives the first labeling; marking only one candidate in the latter correct gives the second point-mass law. These are statistical witnesses in an unrestricted endpoint-law class; we do not claim that both relabelings are semantically admissible for a particular realized problem. Any semantic, structural, or calibration restriction that excludes a witness is precisely an assumption linking public state to endpoints. Such an assumption, labeled calibration, or exact fallback is therefore necessary for a nontrivial uniform guarantee.

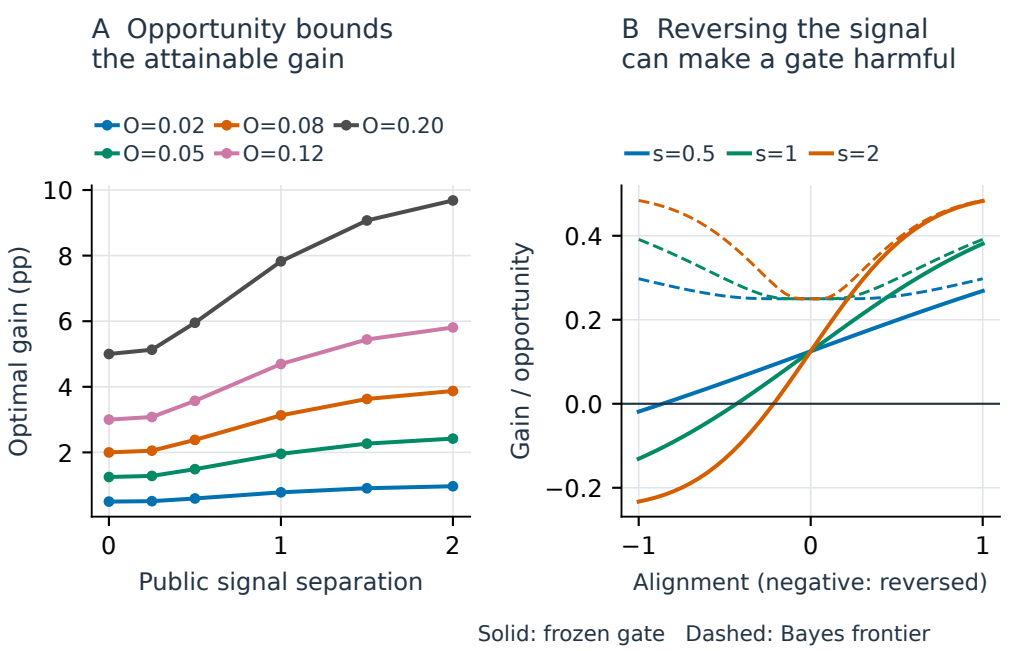

Figure 4: Exact controls. Opportunity and separability set the attainable gain (left). Reversing signal alignment at equal absolute strength makes the declared gate harmful while preserving the optimal ξ-measurable frontier (right). The frontier also varies with signal magnitude.

B.12 Search hardness and deterministic target sensitivity

Even without a quality guard, exact target matching is combinatorial. Given an undirected graph with adjacency G_0, take $S_{\text{obs}} = G_0 + I$, $S^\star = \mathbf{1}\mathbf{1}^\top$, equal public scores, and $\beta = 0$. A size-m subset has $d_{\text{tgt}}(A) = 0$ exactly when it is an m-clique. Hence deciding whether zero mismatch is attainable is NP-complete; this motivates audited approximation rather than a universal search guarantee.

There is nevertheless a deterministic perturbation bound. If true and estimated entries lie in $[0, 1]$, with $\|S_{\text{obs}} - \widehat{S_{\text{obs}}}\|_\infty \le \epsilon_o$ and $\|S^\star - \widehat{S^\star}\|_\infty \le \epsilon_t$, then for every subset

$$|d_{\text{tgt}}(A) - \widehat{d}_{\text{tgt}}(A)| \le 2(\epsilon_o + \epsilon_t).$$

Thus an estimated minimizer over the same feasible set has true target-MSE regret at most $4(\epsilon_o + \epsilon_t)$; an estimated best–runner-up gap strictly above that threshold makes the minimizer invariant. This is a sensitivity statement, not a statistical guarantee: the present target cards do not supply target-domain bounds on ϵ_t.

B.13 Pairwise endpoint bridges

Condition on $G = g$ and a fixed subset A with $m \ge 2$. Write $u_g(A) = \Pr(Y_A = 1 \mid G = g)$; all sums below range over A and probabilities are conditional on g. Let $F_i = \mathbf{1}\{\text{candidate } i \text{ fails}\}$ and $C_i = 1 - F_i$. Since the all-failure indicator is at most the average of its pairwise co-failure indicators,

$$u_g(A) \ge 1 - \frac{2}{m(m-1)} \sum_{i<j} \Pr(F_i F_j = 1).$$

Paley–Zygmund applied to $\sum_i C_i$ also gives

$$u_g(A) \ge \frac{(\sum_i p_i)^2}{\sum_i p_i + 2\sum_{i<j} p_{ij}},$$

where $p_i = \Pr(C_i = 1)$ and $p_{ij} = \Pr(C_i C_j = 1)$, with the second bound defined as zero when its denominator is zero. These bridges require calibrated probabilities; arbitrary similarity matrices do not become endpoint guarantees.

B.14 Separation from monotone diversity

Take $K = 3, m = 2$, equal public scores, and observed off-diagonal similarities $S_{12} = .9$, $S_{13} = S_{23} = .1$. A monotone diversity rule prefers a low-similarity pair. Under a consensus target with off-diagonal one, $\{1, 2\}$ uniquely minimizes squared mismatch; under independence a low-similarity pair does. Thus one fixed repulsion order cannot optimize both, whereas changing $S^\star$ switches the DAKPO objective.

C Target Cards and Matrix Construction

C.1 Which selector is being audited?

DAKPO denotes Dependence-Aware K-sample Post-hoc Optimization. The rules below are alternatives on the same pool. Each robust rule searches its own feasible subsets; it does not certify a subset first chosen by DAKPO. All four DAKPO variants use frozen public features and a DPP public-quality floor. This floor preserves public scores, not hidden correctness.

Table 9: Selector objectives versus the statistical audit. Search limits and numerical tolerances are frozen separately for each route.

Rule	Search or decision	What its check establishes
DAKPO	Maximize the declared diversity–target objective over profile-feasible subsets.	Public mechanism optimization.
DAKPO-CG	Add a floor on distinct public answer groups; the FEVEROUS implementation uses the same beam/one-swap objective as DAKPO.	Group coverage, not correct-answer coverage.
CC-MR	Minimize each subset's worst-case all-failure probability over its moment set; move only if its upper envelope improves on DPP's.	A lower worst-case envelope; subtracting two separately optimized envelopes is not relative safety.
Shared-law	Minimize worst-case excess all-failure probability relative to DPP under one common full-pool law; move only for a strictly negative bound.	Relative coverage gain conditional on true-law inclusion.
Paired release	Freeze any proposal/gate, then lower-bound its mean paired correctness contrast on independent calibration.	Target-average gain under the declared same-law sampling assumptions.

An infeasible moment system, solver failure or failed move criterion returns DPP. Coherence calibration (Appendix D) checks representability of predicted moments, not inclusion of the unknown target law. Paired release tests a different guarantee and cannot supply that premise.

DPP objective and semantic diversity. For the historical math/Live and FEVEROUS implementations, the DPP-style score is

$$J_B(A) = \frac{1}{m} \sum_{i \in A} q_i + .25 \log\det(\widetilde{S}_{\mathrm{sem}}[A, A] + 10^{-8} I).$$

Kernel construction first symmetrizes semantic similarities and maps negative similarities to $[0, 1]$ when present. It then sets the diagonal to one, clips eigenvalues below 10^{-8}, and renormalizes the diagonal. Ties compare the descending quality profile, log determinant, then original indices. The implementation enumerates when $\binom{K}{m} \le 20{,}000$; otherwise it adds candidates greedily. The FEVEROUS $K = 64, m = 4$ baseline is therefore approximate. DAKPO uses a separate diversity term, $V(A) = \{m(m-1)\}^{-1} \sum_{i \neq j \in A} (1 - \bar{S}_{\mathrm{sem},ij})$, with $V = 0$ for a singleton; $\bar{S}$ clips/rescales the selected off-diagonal semantic entries before any spectral projection. This term differs from log determinant.

C.2 HISTORICAL MATH/LIVE TARGET CONSTRUCTION

The target card fixes the public inputs, objective and controls that distinguish target matching from generic diversity or consensus.

Table 10: Historical math/Live target-card fields. Endpoint labels are never selection inputs.

Field	Frozen value or construction
Public quality	Independent LLM verifier score in $[0, 1]$; no reference answer
Observed channels	normalized answer equality; structured risk proxy; cosine or token-Jaccard semantic similarity
Mixture weights	answer/error/semantic $= .4/.4/.2$, renormalized if absent
Structured-error weights	answer/trace/rationale/non-correct-label $= .55/.30/.05/.10$
Main target	risk-conditioned answer-frequency consensus, $\gamma = 1$
Target controls	promptwise permutation, identity, all-ones consensus, constant off-diagonal density
Objective	.25 semantic diversity $-\lambda$ target-MSE; public quality enters only through the hard floor
Frontier	$\lambda = .25, 4$; temporal status is reported per route
Baseline/fallback	DPP quality-diverse using the same public channels
Hard floor	sorted public-score profile weakly dominates DPP
Tie-break	lexicographically smallest original candidate indices

Let $S_{\text{obs}} = .4S_{\text{ans}} + .4S_{\text{err}} + .2S_{\text{sem}}$, renormalizing missing channels, and $d_{\text{tgt}}(A) = m^{-2}\|S_{\text{obs}}[A, A] - S^{\star}[A, A]\|_F^2$. For frozen DPP baseline A_{DPP}, sorted public-score vector $Q(A)$, and mean semantic dissimilarity $V(A)$, the reported probe is

$$A_{\text{DAKPO}}(\lambda) \in \arg \max_{\substack{|A|=m \\ Q(A) \succeq Q(A_{\text{DPP}})}} \{.25V(A) - \lambda d_{\text{tgt}}(A)\}, \qquad \lambda \in \{.25, 4\}.$$

The supplement route cards contain the exact label-only risk mapping, label-conditioned caps/floors, and deterministic ties.

For public verifier score s_i, error risk starts at $r_i = 1 - s_i$, with the route-card corrections just noted. The structural similarity $S_{\text{struct}} = .55S_{\text{ans}} + .30S_{\text{trace}} + .05S_{\text{rationale}} + .10S_{\text{label}}$ is renormalized over present channels and $(S_{\text{err}})_{ij} = (S_{\text{struct}})_{ij} r_i r_j$. If answer a_i appears $n(a_i)$ times, the risk-frequency target is

$$T_{ij} = \mathbf{1}\{a_i = a_j\} \frac{n(a_i)}{K} (1 - r_i)(1 - r_j), \quad i \neq j, \qquad T_{ii} = 1.$$

The implementation symmetrizes the target and clips entries to $[0, 1]$. A route card can also request projection to a positive-semidefinite correlation matrix. The main routes do not use this projection.

FEVEROUS uses a different, explicit card. For a valid candidate, $q_i = \sqrt{p_i^{\text{NLI}} \, r_i^{\text{SC}}}$, where p_i^{NLI} is the cited-premise NLI probability of its stated verdict and r_i^{SC} is its pool Self-Certainty midrank. Invalid parses use the frozen zero-quality sentinel. Off-diagonal answer similarity is equality of valid verdicts; semantic similarity is multiset token-Jaccard of the cited-premise representation. With a_{ij} and s_{ij} these two similarities, the error proxy is $(1 - q_i)(1 - q_j)(a_{ij} + s_{ij})/2$. Channel diagonals are one. The observed mixture is $.4S_{\text{ans}} + .4S_{\text{err}} + .2S_{\text{sem}}$. The fixed target, $.4\mathbf{1}\mathbf{1}^{\top} + .6I$, differs from the math/Live risk-frequency target. DAKPO uses $\lambda = 4$, beam width 64, eight refinement seeds and at most eight one-swap rounds. Its CG variant also preserves DPP's count of canonical (verdict, evidence-set) groups. These public surrogates guide selection. The hidden endpoint requires both the correct label and complete supporting evidence.

D COHERENCE RADIUS AND HIGHER-ORDER EXTENSION

D.1 RADIUS LP AND FINITE-SAMPLE MEANING

Coherence checks whether predicted moments can belong to a joint failure law. It is separate from endpoint accuracy.

For moment vector h and $\phi(z) = (z_i, z_i z_j)_{i,i<j}$, define the correlated-tail coherence debt (Kim et al., 2025; Bertsimas & Popescu, 2005)

$$\rho^\star(h) = \min_{q \in \Delta_{2^d}} \left\| \sum_{z \in \{0,1\}^d} q_z \phi(z) - h \right\|_\infty .$$

Its LP uses variables $(q_z)_{z\in\{0,1\}^d}$ and ρ, with $q_z \geq 0$, $\sum_z q_z = 1$, and

$$-\rho \leq \sum_z q_z \phi_\ell(z) - h_\ell \leq \rho$$

for every moment coordinate ℓ. Thus $\rho^\star(h)$ is the exact ℓ_∞ distance to the marginal polytope, and the LP supplies a coherent witness. When $\lceil (n+1)(1-\alpha) \rceil \leq n$, the frozen order statistic satisfies $\Pr\{\rho^\star(h_{n+1}) \leq s_{(k)}\} \geq k/(n+1) \geq 1-\alpha$ under exchangeability (ties are conservative). If the rank is truncated to n, the maximum guaranteed level is $n/(n+1)$. Neither statement covers unknown true moments.

Table 11: Source coherence calibration and public-only movement audit. These are not realized utility results.

Route	Feasible	Move	Mean s/record	Search
MATH DeepSeek K8	.942	.360	.016	exact
MATH Qwen K16	.880	.322	.103	one-swap
Omni Qwen K16	.910	.355	.091	one-swap

The five fold radii are .015633, .011333, .018222, .007851, .010685. Safety splits contain 196–264 units, with empirical coherence-debt rank coverage .904–.908. These results validate the representability gate; they do not establish coverage of the true target moments.

D.2 Higher-order moments (not evaluated)

Higher-order moments can narrow the ambiguity set at exponential cost. This extension was not evaluated empirically.

Let $X_i = \mathbf{1}\{\text{candidate } i \text{ fails}\}$. For a prefrozen family $\mathcal{F}_r = \{T \subseteq [K] : 1 \leq |T| \leq r\}$, define $\mu_T(P) = \mathbb{E}_P[\prod_{i\in T} X_i]$ and

$$\mathcal{P}_r(g) = \left\{ q \in \Delta_{2^K} : \ell_T(g) \leq \sum_x q_x \prod_{i\in T} x_i \leq u_T(g),\ T \in \mathcal{F}_r \right\}.$$

The worst excess all-failure probability of A over B is

$$R_r(A; B \mid g) := \sup_{q \in \mathcal{P}_r(g)} \sum_{x\in\{0,1\}^K} q_x \left(\prod_{i\in A} x_i - \prod_{i\in B} x_i \right).$$

Proposition 9 (Nested higher-order certificates)**.** *If $\mathcal{P}_r(g) \neq \varnothing$, the atom LP computes $R_r(A; B \mid g)$ sharply. If $P_{0,g} \in \mathcal{P}_r(g)$ and $R_r(A; B \mid g) \leq -\epsilon$, selected-set coverage exceeds the baseline by at least ϵ. If intervals through order $r+1$ are calibrated as one fixed simultaneous family, then $\mathcal{P}_{r+1}(g) \subseteq \mathcal{P}_r(g)$ and $R_{r+1}(A; B \mid g) \leq R_r(A; B \mid g)$.*

The proof uses the atom simplex and nestedness. The LP retains 2^K atom variables and $2\sum_{s=1}^{r} \binom{K}{s}$ moment inequalities. With i.i.d. labeled target units, each prespecified intersection event gives i.i.d. Bernoulli observations. A fixed family of pooled moments can then use Bonferroni–Clopper–Pearson intervals on an independent target split. Finite exchangeability alone does not make these counts binomial. The resulting coverage is marginal and simultaneous; it does not cover every prompt-conditional law. Selecting triples on the calibration split invalidates even this event. Discovery and calibration must therefore remain separate.

Moments through order $K-1$ can still leave the all-failure tail unidentified: even/odd parity laws agree on every proper intersection moment. The full nonempty family through order K identifies the law by Möbius inversion. Block models reduce the parameter count by adding assumptions.

To see conditional shared-law dominance, represent any law on $\{0,1\}^K$ as an atom-simplex vector. If $P_{0,g}$ is feasible, its relative all-failure objective is bounded by the supremum defining R_r. Taking complements gives the selected-coverage claim. Nested ambiguity sets give monotonicity. Lower-order non-identifiability already occurs at $m=3$: the uniform laws on $\{000,011,101,110\}$ and $\{001,010,100,111\}$ share marginals $1/2$ and pairwise moments $1/4$, but their all-failure probabilities are 0 and $1/4$. Appending deterministic failures handles larger m.

E Complete Target Ablation

These controls test whether closer target matching improves endpoint utility. The complete panel separates selected-set coverage from fixed-head final answers.

Table 12: Complete retrospective mechanism and downstream endpoints against same-pool DPP. Target is target-MSE (negative improves match); final-answer uses one frozen verifier-tiebroken plurality head.

Route	Coverage DPP→Sel.	W/L	ΔTarget	Answer Δ (W/L)
MATH $K=8, m=4$	.502→.502	3/3	-.0391	+.010 (10/5)
MATH $K=16, m=8$	.558→.566	4/0	-.0286	+.020 (13/3)
Omni $K=16, m=8$	.300→.325	5/0	-.0334	.000 (4/4)
DeepSeek $K=8, m=4$	.724→.732	4/0	-.0292	—

Table 13: Target-MSE / selected-set coverage means. All MSEs are evaluated against the matched target. CD denotes constant density.

Route	DPP	Matched	Permuted	Indep.	Consensus	CD
MATH K8 MSE	.1029	.0637	.0647	.0711	.2108	.0660
MATH K8 pass	.502	.502	.504	.504	.474	.504
MATH K16 MSE	.1181	.0895	.0917	.0900	.2442	.0918
MATH K16 pass	.558	.566	.564	.566	.510	.566
Omni K16 MSE	.0850	.0516	.0518	.0516	.2157	.0517
Omni K16 pass	.300	.325	.325	.325	.240	.325
Cross MATH N200 MSE	.0912	.0560	.0580	.0646	.1646	.0591
Cross MATH N200 pass	.580	.585	.590	.590	.555	.590

Matched-minus-permuted target-MSE differences (95% paired bootstrap CI) are $-.0009[-.0014,-.0006]$, $-.0021[-.0028,-.0015]$, $-.0002[-.0006,.0001]$, and $-.0020[-.0033,-.0010]$ for the four rows. Matched-minus-constant-density differences are $-.0022[-.0029,-.0016]$, $-.0023[-.0030,-.0016]$, $-.0001[-.0003,.0000]$, and $-.0031[-.0045,-.0019]$. The Omni-MATH controls do not cleanly separate the matched target from these alternatives.

F Secondary Final-Answer Audit

F.1 Historical MATH final-answer heads

This audit tests whether coverage gains survive the final-answer decision. The fixed plurality head selects the largest normalized answer cluster within each subset, breaks cluster ties by the best public verifier score, then uses deterministic ties. Correctness is joined only for evaluation.

On MATH K16, matched-minus-permuted accuracy is $+.008$ (4/0, descriptive exact $p=.125$); matched-minus-independence and matched-minus-constant-density are each $+.006$ (3/0, $p=.25$). The result shows a realized final-answer improvement without establishing a unique advantage for

Table 14: Retrospective secondary single-answer accuracy under the same fixed plurality head. Intervals are descriptive paired-bootstrap intervals.

Route	DPP	Matched	W/L	Difference [95% interval]
MATH K16	.434	.454	13/3	.020 [.006,.036]
MATH K8	.438	.448	10/5	.010 [-.004,.026]
Omni K16	.165	.165	4/4	.000 [-.030,.030]
Cross MATH N200	.515	.510	0/1	-.005 [-.015,.000]

the matched target. The accompanying largest-wrong-cluster share falls by $-.0675[-.0765, -.0588]$, correct-cluster advantage rises by $.067[.057, .077]$, and correct vote share changes by only $-.0005[-.0040, .0028]$. These post-hoc associations describe the change; they do not identify a causal mediator.

F.2 FEVEROUS NATIVE SINGLE-ANSWER COMPARATORS

Table 15 reproduces the released 400-pool discovery panel: one emitted (verdict, evidence-set) answer per rule on the same $K = 64$ pools. It measures single-answer success without new endpoint access.

Table 15: FEVEROUS discovery, native single-answer strict success. W/L and the stored two-sided Holm-4 values compare each row with the DPP head. nM is the descriptive missed-opportunity count for this final-answer pair. Discovery evidence is not prospective superiority.

Rule	Correct / 400	Accuracy	W/L	nM	Holm p
DPP + frozen head	79	.1975	—	—	—
DAKPO + same head	79	.1975	0/0	104	1
DAKPO-CG + same head	79	.1975	0/0	104	1
Majority-of-the-Bests	79	.1975	0/0	104	1
Self-Certainty top-1	48	.1200	7/38	97	1.249×10^{-5}

The released row projection has 183 full-pool successes and 79 baseline final-answer successes, giving $nO = 104$. Thus the first three proposals have $(nR, nL, nM) = (0, 0, 104)$, while native Self-Certainty has $(7, 38, 97)$. These are answer-level counts: the top-4 opportunity count 93 cannot be substituted. Equal correctness vectors leave no observed two-action gate gain for the first three proposals. Their large missed mass does not establish that an alternative answer head can recover it. This descriptive ledger neither inherits the nine-row coverage certificate nor adds a new final-answer significance test.

MoB uses the published bootstrap-output distribution with adaptive $q = .75$. It returns the modal canonical answer group (Rakhsha et al., 2025), without expansion to four outputs. Self-Certainty uses the generator's full-vocabulary probabilities (Kang et al., 2025). Its native output is top-1; top-4 is a set-valued extension. The subset heads choose the largest canonical joint output group. They break ties by group quality sum, group maximum quality, then earliest member. Within-group ties use quality, Self-Certainty, NLI probability, citation relevance and index. The matched-set Best-of-Majority rule (Di et al., 2026) uses mean group reward, frequency threshold $1/64$, and the frozen quality-based backfill to four indices. This adapter does not reproduce every native regime or theoretical assumption of that method.

The four .1975 rows have identical correctness vectors. DPP and native MoB match DAKPO on every unit, so the advantage over Self-Certainty is not DAKPO-specific. Model calls and feature acquisition are shared in this fixed-pool audit. Search and self-refinement change the candidate law (Yao et al., 2023; Madaan et al., 2023; Snell et al., 2024; Tuyls et al., 2026). Comparing those systems requires a separate study that records generation, scoring and selection costs; this panel implies no equal-total-compute superiority.

G MATCHED-HEAD DEVELOPMENT TRANSFER

Question and controlled comparison. This comparison separates better answer scoring from better subset selection. Direct top-1 selection uses all $K = 8$ candidates. DPP and profile-constrained DAKPO first retain $m = 4$, then apply the *same* top-1 head. Ties follow candidate order. We cross these three selection rules with frequency scoring and two fixed linear heads: absolute candidate-F1 regression and within-pool pairwise-F1 regression. Candidate 0 completes the ten-configuration panel. DPP/DAKPO subsets use the fixed answer-frequency quality score and are built once per pool, then reused across heads. All three scored rules evaluate the same eight head scores; candidate 0 uses no learned head. We did not measure inference-cost savings.

The TAT-QA benchmark combines financial tables and text (Zhu et al., 2021). Both learned heads use eight candidate-only features: answer and scale frequencies, reciprocal and normalized candidate rank, log response and answer lengths, and nonempty-answer/scale indicators. Their inputs exclude gold answer-type/source metadata, endpoint scores and cached fitted scores. Four-fold table-grouped development on 400 training questions (359 tables) selected ridge regularization .001. We then fitted each head once and reused it; language-model weights were unchanged. The tuning criterion averaged DPP and DAKPO performance. Direct-K8 is thus a matched-head control, not an exhaustively tuned routing baseline.

Two panels, not pooled replications. The first panel reuses 200 development questions from 149 tables. Direct-K8 was added after the initial head comparison. The second samples 600 training-source tables outside the known head-training and earlier development groups, with one question per table. Its ten methods and four contrasts were fixed before scoring. We collected all 4,800 responses, saved the public choices, then joined training-source labels for evaluation. The second panel is therefore new development, *not* untouched confirmation or an admission study. It uses Qwen3:32b Q4_K_M without thinking, temperature .7, top-p = .95, a 512-token output cap, 16,384-token context and four collection workers. A documented interruption amendment reissued four missing requests once, under unchanged settings. All saved replies were retained; there were no quality-based retries. Historical generation metadata is incomplete. The panels also differ in source split and table-versus-question weighting, so they are not exact replications under a common sampling law.

Table 16: Complete matched-head development panel. F1/EM use the official TAT-QA scoring implementation and are shown on a 0–100 scale. DAKPO denotes the fixed profile-constrained variant. Columns are not pooled.

		Reused dev200		New train600	
Head	Selection	F1	EM	F1	EM
—	Candidate 0	57.630	50.500	58.925	50.667
Frequency	Direct-K8	57.625	50.500	59.437	51.333
Frequency	DPP	57.120	49.500	59.185	50.833
Frequency	DAKPO	57.625	50.500	59.437	51.333
Absolute	Direct-K8	57.120	49.500	59.222	51.000
Absolute	DPP	57.120	49.500	59.378	51.167
Absolute	DAKPO	57.120	49.500	59.385	51.167
Pairwise	Direct-K8	58.835	52.500	59.338	51.167
Pairwise	DPP	58.835	52.500	59.302	51.167
Pairwise	DAKPO	59.335	53.000	59.152	51.000

Attribution and limits. Table 16 retains the initial head improvements: $+2.215$ F1 points within DAKPO and $+1.715$ within DPP on dev200. The same-head DAKPO margin over direct-K8/DPP was $+.500$ points, from one question. On train600, DAKPO falls below both matched-head controls. Neither head-upgrade contrast establishes positive transfer (Table 17). The development gain therefore cannot be attributed wholly to subset selection or assumed to transfer.

The bootstrap uses 20,000 replicates, seed 20260906, common paired table indices across contrasts and linear percentile quantiles. Each algorithm contrast has only three nonzero F1 differences. In the direct-K8 contrast, all three favor that control, so resampling cannot represent an unobserved positive tail. An upper endpoint of zero does not prove population non-improvement. These marginal

Table 17: Four predeclared new-development contrasts. Differences are F1 percentage points. $+/-/=$ counts nonzero positive/negative/tied F1 changes, not changes in candidate identity or binary rescue/harm counts. Intervals are nominal marginal 95% paired-bootstrap diagnostics, not admission certificates.

Contrast	F1 gap	Interval	$+/-/=$
Pairwise DAKPO − pairwise direct-K8	$-.187$	$[-.540, .000]$	0/3/597
Pairwise DAKPO − pairwise DPP	$-.150$	$[-.520, .060]$	1/2/597
Pairwise DAKPO − absolute DAKPO	$-.233$	$[-1.297, .825]$	14/17/569
Pairwise direct-K8 − absolute direct-K8	$+.117$	$[-.907, 1.175]$	14/14/572

intervals assume approximately independent tables; they do not model shared-issuer dependence or provide design-based exact coverage.

The panel supports no equivalence or positive deployment claim, and we do not select a new winner after inspection. It measures realized selected-answer utility, not Γ_G or rescue–harm overlap. It therefore cannot identify the loss responsible for failed transfer or show that algorithmic headroom is exhausted. The supplement retains aggregate panels and contrasts without redistributing prompts, predictions or training-source labels.

H Search Audit

This audit compares practical and exact selection on cached scenarios, separately from the statistical audit's cost.

Table 18: Beam-64 versus exact selection on the seven cached scenarios.

Statistic	Minimum	Mean	Maximum	Required
Exact subset match	.960	.989	1.000	descriptive
Jaccard overlap	.987	.996	1.000	descriptive
Absolute pass delta	0	0	0	0
Absolute target-MSE delta	0	—	9.50×10^{-5}	small
Profile violations	0	0	0	0

Beam search logs $mbK = 8{,}192$ as a proxy for partial-addition work at $K = 16, m = 8, b = 64$. This excludes other evaluations. Refinement uses at most eight seeds and eight improving one-swap rounds per seed. The results support feasibility and fidelity on these scenarios; they do not certify a global DAKPO optimum or performance on untested matrix geometries.

Audit cost is distinct from proposal-search cost. Once paired outcomes and public cells are available, auditing one fixed gate requires one pass over n rows and two counts. For a fixed B_Z-cell partition, $\widehat{\Gamma}_Z = \sum_z [n_z^+ - n_z^-]_+/n$ is computed from counts in $O(n + B_Z)$ time, not by enumerating all 2^{B_Z} gates in the proof. After interval construction, the sharp interval–simplex criteria require constant work per comparison. They do not require the 2^K-atom failure-law LP used by the older full-pool shared-law selector. These costs exclude generation, feature extraction and correctness evaluation. Full-pool diagnosis may still need nK candidate judgments. FEVEROUS demonstrates an audit at $K = 64$; it does not solve the shared-law selector exactly at that scale.

I Prospective Protocol and Post-hoc Audit Details

The LiveAoPS prospective card freezes the data, actions and tests before endpoint access. A seeded hash selects 1,200 LiveAoPSBench-2024 prompts from August–December 2024. Llama-3.1-8B-Instruct generates with temperature .7, top-$p = .95$, and a 1,024-token cap. Prespecified seeds 2026071601 and 2026071602 produce separate pools on the same prompts, with $K = 8, m = 4$. Selection uses an oracle-free Qwen2.5-7B verifier at temperature zero, exact search, DPP as baseline, primary CC-MR, and strict shared-law as secondary. The endpoint uses official math equality. Tests are exact two-sided McNemar, with Holm adjustment for the two strict seed-level tests.

Table 19: Prospective decision/release evidence. D/C/T denotes discovery/calibration/untouched test; counts are rescues/harms, not rates. FEVEROUS and BFCL bounds apply to the listed inferential cells, not their full calibration roles. LiveAoPS seeds reuse the same prompts and are not pooled.

Route	Decision role	Rescue / harm	Certificate and consequence
Synthetic D/C/T	C, $n = 800$	47/2	gain LB $+.0345$; admitted; T gain $+.045$
MedMCQA Stage54	internal C, $n = 1{,}200$	0/0	zero contrast; gain LB $-.00307$; T $n = 1{,}400$ sealed
MedMCQA Stage55	internal C, $n = 4{,}500$	230/224	orientation LB $.4596 < .5$; T $n = 4{,}000$ sealed
MedMCQA successor	new C, $n = 10{,}000$	158/105	gain LB $+.001805$; C admits; T primary passes, repair comparison fails (Table 4)
FEVEROUS	C cell: 498 of 1,000	23/11	cell uncertainty; gain LB $-.00968$; T $n = 1{,}399$ sealed
BFCL	C cell: 136 of 400	7/0	cell uncertainty; gain LB $-.0178$; T $n = 400$ sealed
HellaSwag	C, $n = 2{,}800$	1/0	low activity; gain LB $-.00131$; T $n = 3{,}787$ sealed
LiveAoPS	two seeds, same 1,200 prompts	12/7 and 4/5	seed instability; conjunction IUT $p = 1$

The primary claim is a predeclared conjunction: both seeds must have positive deltas and significant tests. Its IUT p-value is the larger seed-level value; neither row is a separate claim. Gold answers were hash-committed and excluded until all public selections were frozen. Selection and endpoint artifacts contain no raw gold text.

Frozen BFCL route. Before answer access, the BFCL route fixes eight generator identities and seeds, $K = 8, m = 4$, and 100/400/400 prompt-disjoint discovery/calibration/test roles. It also fixes the AST evaluator, public features, five comparator families, prompt-grouped selector, three movement tertiles, and six-bound admission rule. The final public pool hashes are `4be2ef33`, `95da1c0f`, and `32459b59`; the all-pool integrity report hash is `9bd84659`. Discovery authorization precedes model fitting; calibration and test selection hashes `031aff43` and `7ae048f4` are fixed before calibration authorization. The resulting certificate hash is `dcbb2c2b`; it records `any_cell_admitted=false` and `test_answers_opened=false`. These prefixes identify the full SHA-256 hashes in the released revision ledger. All writes use no-replace semantics, and the 73-test implementation suite passes before the integrity audit and label joins.

On discovery, candidate accuracy is .740 and pool opportunity is .890. Prompt-blocked cross-validation selects $C = .1$. The public selector differs from DPP on 383/400 calibration prompts, where pool opportunity is .9325. The predeclared low, middle, and high movement cells have 0/0, 0/0, and 7/0 rescues/harms at sizes 135, 112, and 136; 17 prompts are non-movements. The high cell has empirical gain .05147. Its simultaneous rescue lower and harm upper bounds are .01676 and .03459, giving lower gain -.01783. No cell is admitted. The router returns DPP, and the test remains ineligible for release.

Frozen HellaSwag route. HellaSwag v3 uses one representative per source group. Split-type-stratified salted hashes fix 1,800 discovery, 2,800 calibration, and 3,787 untouched test units. Qwen3-4B generates eight fixed public views, from which the policy returns one primary answer. Raw and exact-HF-scored public pools were completed for every role before discovery correctness was released. Discovery then froze the router, `direct_view` champion, and both C/T public selection ledgers before calibration access. Compatibility bridges changed only serialization, order, and ledger interfaces; all pools, public selections, endpoints, thresholds, comparator membership, and inferential rules remained unchanged.

The local one-way protocol released only binary router/champion correctness, after a role-exact calibration grant, and retained no gold answer. It had no independent custodian or external blinding attestation. The observed 1/0/2,799 rescue/harm/tie counts give empirical gain .000357. With two one-sided Clopper–Pearson bounds at miscoverage .025, the rescue lower bound is .00000904, the harm

upper bound is .00131659, and the simultaneous lower gain is -.00130755. The certificate records `admitted=false`. The raw credential was destroyed after the release receipt; test authorization, payload, receipt, analysis, and state remain absent. CP coverage is exact under the declared i.i.d. source-group Bernoulli superpopulation model. It is neither unconditional design-based coverage of this fixed HellaSwag population nor robust to arbitrary stratum-specific rates.

Frozen MedMCQA Stage-54 route. The prior MedMCQA execution used 700 discovery, 1,200 internal calibration, and 1,400 internal test groups. Its frozen discovery policy selected a fallback identical to the plurality champion. Calibration produced no endpoint discordances: 842 units were correct under both actions and 358 were wrong under both, so $(R, L) = (0, 0)$. The simultaneous lower gain was $-.003069$, the route was nonadmitted, and the 1,400 test labels remain sealed. A classification-equivalent binary64 threshold amendment was publicly timestamped after pool construction but before endpoint access. This preserves the local temporal ordering. The route was nevertheless neither fully frozen before data collection nor independently custodied.

Frozen MedMCQA Stage-55 route. This route asks whether semantic plurality over eight views from Qwen3-4B-Instruct-2507 improves single-answer accuracy over the first canonical view on MedMCQA (Pal et al., 2022). The model snapshot, eight balanced semantic option permutations, sampling settings, parser, direct baseline, plurality action, and public tie rule were fixed before Stage-55 allocation. There is no learned router or instance partition. After excluding all Stage-54 allocations, a public random value orders the unused duplicate components. The first 4,500 fresh groups form calibration; the next 4,000 form untouched test. The author preregistration is publicly timestamped, with a procedurally sealed test and no independent custodian.

Under the declared i.i.d. superpopulation law over eligible source-group units, write $q = \Pr(D \neq 0)$ and $\theta = \Pr(D = 1 \mid D \neq 0)$, so $\Delta = q(2\theta - 1)$. Calibration reserves family error .05 as two one-sided Clopper–Pearson lower bounds with miscoverage .025 each and admits only if $q_L > 0$ and $\theta_L > .5$, equivalently when the factorized strict lower gain is positive. The public $K = 8$ pools for both roles were complete before the sole calibration join. Calibration then records 230 rescues, 224 harms, 2,711 both-correct, and 1,335 both-wrong units. Thus $\widehat{\Delta} = 6/4500 = .001333$, $q_L = .092239$, and $\theta_L = .459604$. The route terminates nonadmitted; no test authorization or result exists and all 4,000 test labels remain sealed. Operational recovery records are retained in the revision ledger but are not scientific evidence.

Retrospective representation diagnostic. After calibration was opened, a binary refinement isolated rows with eight valid parses, plurality different from the direct view, and support at least six. It uses the same public state. The two cells contain 76/43 and 154/181 rescues/harms. Plug-in frontier/overlap counts change from 6/224 to 33/197, with both differences equal to 27. This replayable aggregate result nominates a design; it does not authorize it. A successor needs fresh independent calibration and test groups. The Stage-55 test remains sealed and cannot be reassigned.

Separately preregistered MedMCQA refinement replication. This study tests the refinement nominated by the disclosed Stage-55 calibration summary. That summary serves as retrospective discovery; new calibration/test roles contain 10,000/20,000 groups. This is not a new three-way discovery split. From the 57,562 eligible groups, allocation excludes all 12,200 Stage-54/55 assignments, including their sealed tests. A committed public randomness value ranks the remainder and selects one representative per group, without using answers, explanations, subjects, or difficulty. The remaining 15,362 groups are not used to rescue the study. The model and analysis bundle was timestamped before allocation. This records when the design was committed; it does not establish independent administration or exclude benchmark contamination.

Generation uses Qwen/Qwen3-4B-Instruct-2507, revision `cdbee75f17c01a7cc42f958dc650907174af0554`, through one persistent MPS float16 model (PyTorch 2.8.0, Transformers 4.56.2). Eight balanced option permutations retain the frozen chat template, per-view seeds, strict `FINAL: X` parser, and semantic mapping. Temperature is .7, top-p .8, top-k 20, maximum new tokens 96, context 2,048, and thinking disabled. Invalid or length-capped choices stay in the denominator, with no replacement sampling. All 30,000 public pools and direct, plurality, and gated actions are fixed before any calibration label join. The single-answer endpoint is exact correctness against the source answer, not selected-set coverage.

Let B be view0 and P the unchanged semantic plurality action, with the existing deterministic tie rule. The frozen event A requires eight valid parses, $P \neq B$, and unique plurality support at least six. Set $S = P$ on A, otherwise $S = B$. The public state already contains the eight semantic parses; Z_0 is trivial and $Z_1 = \{A, A^c\}$. The gate acquires no new information and estimates no full-G frontier. It uses no learned head, further threshold search, or post-calibration selection.

Calibration for $S - B$ uses the same factorized certificate as above: q_L and θ_L each have one-sided tail .025, with admission only when both $q_L > 0$ and $\theta_L > .5$. It records 158 rescues, 105 harms, 6,306 both-correct and 3,431 both-wrong groups. The point gain is .0053, $q_L = .0232521$, $\theta_L = .5388082$, and the strict lower gain is .00180474. Calibration therefore authorizes the new test once; it does not authorize reuse of any earlier test.

The primary test requires exact one-sided McNemar $S > B$ at .05 and a strictly positive factorized 95% lower gain. It passes with 383/182 rescues/harms, 13,047 both-correct, 6,388 both-wrong, gain .01005, $p = 9.09075 \times 10^{-18}$, and lower gain .00715475. Primary success opens the registered secondary test $S > P$ at .05. It records 649/743 rescues/harms, 12,781 both-correct, 5,827 both-wrong, gain $-.00470$, and $p = .994569$. The fixed sequence controls this two-test family without assuming independent contrasts. Both contrasts were required for diagnosis-guided repair success, so that claim fails. The direct-baseline benefit remains valid within its scope; it is not an overall positive real-LLM admission or DAKPO claim. The negative secondary point estimate also does not constitute a registered reverse-direction test.

A separate simultaneous 95% confidence family uses four one-sided CP rate bounds, each with tail .0125: rescue lower and harm upper for $S - B$ and $S - P$. The resulting gain lower bounds are .00631802 and $-.01055594$. They are not the primary factorized bound. Write $a = \mathbb{E}[A(U_P - U_B)]$ and $b = -\mathbb{E}[(1 - A)(U_P - U_B)]$. Then

$$\Gamma_{Z_1} - \Gamma_{Z_0} = [a]_+ + [-b]_+ - [a - b]_+.$$

If both simultaneous gain lower bounds are positive, their minimum bounds this increment; otherwise the frozen procedure reports the trivial lower bound zero. Here it reports zero. The observed A and A^c contrasts both favor P, unlike the exploratory fallback cell, so the plug-in increment is also zero. This does not establish equality of population frontiers. Table 20 reports the complete descriptive ledger.

Table 20: Prespecified descriptive ledger for the unchanged plurality proposal versus direct answering. All entries are rates. S is the frozen gate; the last three columns are its empirical three-loss decomposition. These plug-ins are not population confidence bounds.

Role	$\widehat{O}$	$\widehat{R}$	$\widehat{\Gamma}_{Z_0} = \widehat{\Gamma}_{Z_1}$	$\widehat{\Delta}_S$	Capture	Overlap	Gate regret
C	.17040	.05590	.01320	.00530	.11450	.04270	.00790
T	.16785	.05630	.01475	.01005	.11155	.04155	.00470

Inference covers this fixed successor study under the declared same-law, independent-source-group model. It does not provide unconditional design-based coverage on a finite benchmark or an error guarantee over the research history. Earlier negative results remain disclosed. The authors, rather than an independent evaluator, controlled test-label access. Fresh groups are disjoint from recorded Stage-54/55 allocations; this says nothing about model pretraining or all external histories. The benchmark establishes no clinical utility. The anonymous supplement provides terminal aggregates without source labels or candidate text.

Secondary nine-row confidence audit. The secondary audit asks whether missed capture exceeds the full downstream repair budget. Its family follows the pre-endpoint LiveAoPS primary/strict comparisons and the FEVEROUS matched-pass@4 contract, excluding DPP. Four Live seed–policy rows and five FEVEROUS proposals contribute $(M, R, B = L)$ rates. Let $C(k, n; a)$ denote an equal-tailed Clopper–Pearson interval with miscoverage a. Each of the 27 intervals uses $.05/27$. Borda and Best-of-Majority remain separate family members despite identical subsets and endpoint vectors on these 400 units. A union bound gives at least .95 simultaneous coverage. Proposition 3 then allows data-dependent reporting within this nine-member family, but no endpoint-chosen additions.

The simplex $M + R + L \leq 1$ does not tighten any empirical row: $\bar{b} = u_L$ in all nine. Uniformly positive gain is certified in 0/9 rows, each with an explicit compatible nonpositive-gain witness. In

contrast, $\ell_M > \bar{b}$ certifies capture-budget dominance in 9/9. The first result leaves gain unresolved over the confidence polytope; the second establishes a simultaneous budget ordering.

Table 21: Adverse non-admission witnesses in the joint 95% rectangle–simplex event. Each row uses $(M, R, L) = (\ell_M, \ell_R, \bar{b})$; rates are rounded only for display. In every row $\ell_R \le \bar{b} < \ell_M$, so nonpositive gain remains compatible while capture-budget dominance is certified.

Route/proposal	ℓ_M	ℓ_R	$\bar{b}$
Live 1601 CC-MR	.02722	.00335	.01636
Live 1601 strict	.03392	.00004	.00774
Live 1602 CC-MR	.03055	.00035	.01374
Live 1602 strict	.03324	.00000	.00580
FEV DAKPO	.16167	.00105	.04857
FEV DAKPO-CG	.16387	.00046	.04475
FEV Borda	.14200	.01154	.05231
FEV Best-of-Majority	.14200	.01154	.05231
FEV Self-Certainty	.10364	.04199	.09008

The direct order-test sensitivity conditions on $N_A + N_L$ and tests $A \le L$ at the binomial boundary $1/2$, for $A = R$ and $A = M$ in every row. All 18 one-sided exact p-values are Holm-adjusted together. One-sided Clopper–Pearson lower bounds for θ_{AL}, each at miscoverage $.05/18$, are transformed to $A/L = \theta_{AL}/(1 - \theta_{AL})$; a union bound makes the reported ratios simultaneous. Table 22 reports every comparison. This test avoids a penalty for unrelated coordinates; it remains a secondary audit designed after endpoint release.

Table 22: Decision-direct exact order audit. p_A and p_M are Holm values for $R > L$ admission and $M > L$ repair, respectively. Ratio bounds are simultaneous one-sided lower bounds.

Row	R/L	p_A	M/L	p_M	M/L lower
Live 1601 CC-MR	12/7	1	52/7	7.47×10^{-9}	2.59
Live 1601 strict	2/1	1	62/1	1.04×10^{-16}	7.23
Live 1602 CC-MR	4/5	1	57/5	1.84×10^{-11}	3.48
Live 1602 strict	0/0	1	61/0	6.94×10^{-18}	9.87
FEV DAKPO	4/7	1	89/7	2.77×10^{-18}	4.63
FEV DAKPO-CG	3/6	1	90/6	2.25×10^{-19}	5.13
FEV Borda	13/8	1	80/8	3.22×10^{-15}	3.81
FEV BoM	13/8	1	80/8	3.22×10^{-15}	3.81
FEV Self-Certainty	31/19	.535	62/19	8.87×10^{-6}	1.59

Inference assumes i.i.d. prompt units under each fixed row-level target law. It requires no independence among cells, policies, seeds, or tasks. Every pre-endpoint registered comparison is retained. The family has no post-endpoint row filtering, but the confidence analysis itself is secondary, having been formulated after release.

The separate Live complete-family audit is also post-endpoint. It includes all nine non-DPP methods from the endpoint-free public selection freeze, including those originally designated descriptive. For each method and seed, the one-sided exact McNemar null states that rescue probability does not exceed harm probability. The larger seed-level value is an intersection–union p-value for improvement on both seeds. Holm adjusts the nine method-level values without requiring cross-seed or cross-method independence.

Verifier-Borda records 41/16 and 35/15 rescues/harms. The seed values are $.0006318/.0033002$, their conjunction is .0033002, and Holm gives .0297020. Using the registered two-sided per-seed convention instead gives Holm .0594040. Both pools use the same 1,200 prompts, so this is robustness to generation seed, not independent target replication. The unregistered full-family audit cannot change the registered DAKPO verdict. Table 29 retains all nine method-level values.

How the diagnosis changes with the action family. Include the baseline in $\mathcal{A}$ and define $R_{\mathcal{A}} = \Pr(Y_B = 0, \max_{a\in\mathcal{A}} Y_a = 1)$. An endpoint oracle attains gain $R_{\mathcal{A}}$ by retaining the baseline when it is correct and taking a successful family member otherwise. For a reference policy selecting from $\mathcal{A}$, its remaining budget is therefore $R_{\mathcal{A}} - \Delta_{\mathrm{ref}}$, whereas $O - R_{\mathcal{A}}$ is missed by every action. These quantities partition $O - \Delta_{\mathrm{ref}}$. For nested families $\mathcal{A} \subseteq \mathcal{A}'$ and the *same* reference, an increase $d = R_{\mathcal{A}'} - R_{\mathcal{A}} \geq 0$ lowers the capture-minus-budget margin by exactly $2d$. Thus a pair-local ordering cannot be carried to an expanded family without checking its additional rescue mass. This identity gives no public learnability guarantee.

In Table 3, the largest observed method gains are 25/1200 and 20/1200 for Borda on Live, and 12/400 for Self-Certainty on FEVEROUS. Subtracting their numerators from family rescues 54, 51 and 42 gives the displayed budgets 29, 31 and 30. The references are chosen descriptively from the full panel, not independently nominated winners. The nested-family identity fixes its reference and should not be applied across rows whose reference changes.

The portfolio audit asks how much opportunity the complete action family contains. A full-information allocator returns DPP unless a frozen proposal succeeds. The nine Live proposals rescue 54/64 and 51/61 DPP-missed solvable pools; the five FEVEROUS methods rescue 42/93. The gains $.0450, .0425, .1050$ are endpoint-oracle ceilings, not performance of a public policy fitted with endpoint labels.

The family-level diagnosis differs from the 9/9 proposal-local result. Live 1601 and 1602 have family capture losses (10/1200) and (10/1200), below the downstream gaps (29/1200) and (31/1200) relative to Borda. FEVEROUS has a larger sample missed mass, 51/400 versus 30/400. These are sample orderings. The conditional family analysis below leaves all three observed-best-reference orderings unresolved, so the simultaneous nine-row result does not extend to the complete action family.

Simultaneous sensitivity for every reference. The new calculation uses only the previously released endpoint projections. Its plan fixed all families, references and error allocations before this calculation, after historical outcomes were known. It is a secondary fixed-family sensitivity, not a new prospective test. For each unit let $U = \max_{a\in\mathcal{A}} Y_a$ and let Y_j be reference j's outcome. The disjoint Bernoulli events $H - U$ and $U - Y_j$ have probabilities $M_{\mathcal{A}}$ and $\mathcal{B}_j = R_{\mathcal{A}} - \Delta_j$. All 10, 10 and 6 references, including DPP and methods with identical outcomes, are retained. Two-sided CP intervals cover the three shared miss rates and 26 reference budgets. Each of the 58 tails spends $.05/58$. The resulting joint event yields

$$M_{\mathcal{A}} - \mathcal{B}_j \in [\ell_M - \min\{u_{B_j}, 1 - \ell_M\},\ \min\{u_M, 1 - \ell_{B_j}\} - \ell_{B_j}].$$

The simplex caps are inactive for these data. Simultaneity permits reporting the observed-best reference without treating it as independently nominated. Table 3 gives those three intervals. Across all 26 references, 16 intervals support a larger family gate budget (all on Live), ten are unresolved, and none supports a larger missed mass. These are dependent comparisons, not replications. All 26 rows, the plan and the replay are in the supplement's `family_endpoint_audit.py` materials.

Coverage is conditional on the stated historical families and i.i.d. row-level laws. The union bound permits dependence between methods and Live seeds, but does not correct historical route or action development based on endpoints. The calculation changes no registered decision. It quantifies the uncertainty of the family check; it does not validate a learnable allocator or establish equal budgets when an interval crosses zero.

Table 23: Post-endpoint Live action supply and cross-fit allocation. Family capture is the endpoint-oracle $R_{\mathcal{A}}/O$; public increment is public-plus-method minus method-only.

Route	O	Family capture	Borda Δ	Method-only Δ	Public increment
Live 1601	64	54/64	+.0208	+.0200	-.0025
Live 1602	61	51/61	+.0167	+.0167	-.0008

The FEVEROUS discovery card fixes 400 units, (K=64,m=4), Qwen3-4B HSEP-Packing candidates, cited-premise NLI and Self-Certainty public features, the six matched subset methods in Table 26, and strict joint verdict–complete-evidence correctness. All 20 public-feature shards and 20 Self-Certainty shards were completed before a single endpoint join. The endpoint-free seal SHA-256 is:

Table 24: Fit-free algebraic sample bounds for the three losses. Entries are rates; brackets use aggregate summaries and are not confidence intervals.

Route/proposal	Δ	$O - R$	$R - \Gamma_G$	$\Gamma_G - \Delta$
Live 1601 CC-MR	+.0042	.0433	[0,.0058]	[0,.0058]
Live 1602 CC-MR	-.0008	.0475	[0,.0033]	[.0008,.0042]
FEV DAKPO	-.0075	.2225	[0,.0100]	[.0075,.0175]
FEV DAKPO-CG	-.0075	.2250	[0,.0075]	[.0075,.0150]
FEV Borda / BoM	+.0125	.2000	[0,.0200]	[0,.0200]
FEV Self-Certainty	+.0300	.1550	[0,.0475]	[0,.0475]

`8654e219e9300a6d0dc9a96323089e7101216a62b51996f73f88e456e884ec76`. The pre-endpoint analysis-contract SHA-256 is:

`dce02e4a0aedd1d723da553d7694384f9e01841c7e4da9e47083a607609d1d00`. The author-run discovery route is exploratory. In the follow-up route, calibration was accessed only after the double public freeze; the final-test endpoint was not accessed.

Frozen FEVEROUS calibration-law fail-closed audit. The follow-up protocol fixes SC-Guard using the 400 discovery units. Its public feature is the Self-Certainty-top-4 minus DPP range in cited-premise NLI probability. At the discovery-frozen threshold 0.22136104481172092, eligible rows propose Self-Certainty top-4; all others return matched-m DPP. The unit rule samples one claim per evidence-page connected component. Parent discovery/calibration/test splits contain 400/400, 1,000/1,000, and 1,400/1,400 rows/unique components, with maximum multiplicity one and zero cross-split page intersections. The constructible 1,399-unit test also shares no claim or component with discovery or calibration. Before any protected join, we hash-froze and externally timestamped the protocol, implementation surface, split, policy, endpoints, effect floors, and multiplicity families. The protocol SHA-256 is

`ede364de03325297943ec19873459d1884280468af114482f252b790dbffda5f`, with the active freeze receipt timestamped by DigiCert at 2026-08-26 04:39:36 UTC; the endpoint-analysis surface was separately timestamped at 04:40:56 UTC.

Calibration applies a 95% simultaneous pair of one-sided Clopper–Pearson bounds to the eligible cell, requiring at least $n = 100$. The proposal is deployed only if the rescue lower bound strictly exceeds the harm upper bound. Otherwise the policy returns DPP and the final endpoint stays unopened. Admission permits one test evaluation of the externally timestamped router. The primary Holm-2 family contains two-sided exact McNemar utility and an intersection–union admission-predictive-validity test. Utility also requires an observed gain of at least .015. The complete matched-m Holm-5 and native Pass@1 Holm-2 families are reported regardless of sign.

The frozen public builder audited all 1,400 parent test identities before any protected endpoint access: 1,399 were $K = 64$-constructible and one was not. The pre-endpoint v5.3 feasibility amendment excludes every nonconstructible unit without replacement and defines the test target population accordingly. Candidate generation, K, public features, selectors, threshold, endpoints, tests, effect floors, and multiplicity were unchanged. Exact Clopper–Pearson coverage requires independent calibration/test samples, each with i.i.d. component units from the same target law. Component separation removes known shared-page overlap; it does not prove superpopulation independence. The result is neither independently blinded, pointwise, nor robust to arbitrary shift.

The protected calibration join marked 498/1,000 units eligible and observed 23 rescues and 11 harms. Its simultaneous rescue lower and harm upper bounds were .029499 and .039178, respectively, giving conditional-gain lower bound $-.009679$. The certificate therefore selected DPP in both cells and denied final release. The calibration-certificate SHA-256 is

`825f2511f4d0d71a286512d81003b1decbba8a19afa8129a58e20ddf2a8a1f19`; the frozen fail-closed router SHA-256 is

`0d06902f440325cce2da86cefc91d5d5c2d9aec1fc5a80598019e11c407667bd`. Treating the complete frozen SC-Guard policy as one global action, a separately labeled post-calibration audit gives empirical gain/frontier .01200 and the simultaneous exact interval $[0, .031107]$; it remains fail-closed and is not confirmatory. Its audit SHA-256 is

`82f48c606fdb0a26105ac0284fb507e62a9da841d807e1b590e9a4f251050984`. No final endpoint-row or confirmatory-result artifact was produced.

For the descriptive failure audit, candidate correctness is the existing official endpoint boolean. Score discrimination on an opportunity pool averages $\mathbf{1}\{s_{\text{correct}} > s_{\text{wrong}}\} + .5\mathbf{1}\{s_{\text{correct}} = s_{\text{wrong}}\}$ over every correct–wrong candidate pair and then pools pairs within a seed. An identity target has all off-diagonal entries numerically zero. A primary discordance is score-tied only when every candidate added by CC-MR has the same scalar verifier score as every removed candidate. Certificate-gain summaries are unadjusted descriptive means, not hypothesis tests.

The cross-fit diagnostic tests whether the available public features improve action allocation. It was frozen on 2026-08-24, after endpoint release but before row-level computation. A hash assigns each prompt ID to one of two folds. Each target seed–fold is trained on the other seed and opposite fold, excluding both the target prompt and its other-seed pool from training endpoints. All nine proposals are retained. The public vector comprises certificate gain, proposal and relative sharp cofailure bounds, DPP overlap, movement, two validity bits, baseline coherence debt, and calibrated radius. Missing certificate values are mean-imputed from the training cell; features are training-standardized. Fixed ridge $\alpha = 10$ is used without tuning. METHOD-ONLY, PUBLIC-SHARED, and PUBLIC-PLUS-METHOD routers select the largest predicted positive gain and otherwise return DPP. The primary public-plus-method versus method-only contrast is negative on both seeds (3/6 and 3/4). Method-only versus DPP is 38/14 and 35/15. The method-only router chooses Best-of-Majority on 598 of 2,400 rows and verifier-Borda on 1,802; it uses no other proposal. All four frozen contrasts, coefficients, decisions, input hashes, and the failed primary gate are recorded in the executable artifact. Cross-fitting prevents row-level endpoint reuse. The diagnostic form was nevertheless chosen after aggregate endpoint results were known. Its adjusted p-values remain exploratory; they neither estimate $\Gamma_{\mathcal{G},\mathcal{A}}$ nor validate a lower surface.

Table 25: Frozen nine-feature design and post-endpoint failure audit. The diagnostics explain the tested representation; they were not used to select a new feature set.

Design role	Frozen public features	Observed limitation
Action geometry	DPP overlap; moved indicator	Threshold-redundant summaries; slopes do not transfer
Certificate risk	certificate gain; sharp bound; relative sharp bound	One exact signed duplicate on 21,168 finite rows
Validity and coverage	two validity bits; coherence debt; calibrated radius	Radius constant in all four training cells

Seven of the eight nonconstant PUBLIC-SHARED slopes reverse sign across the four target cells. Only quality-profile preservation is sign-stable. The frozen ridge uses shared slopes and method intercepts, without method-by-feature interactions. It therefore cannot express a public signal whose meaning changes by proposal. These observations motivate future tests of verifier-ensemble disagreement, answer-cluster margins, structured-rationale checks, and method-specific slopes. Such features would need to be predeclared and evaluated on a new discovery split; this audit does not validate them.

J Additional Routes and Claim Boundaries

These complete panels retain gains, nulls and reversals. Endpoints are not pooled, and development comparisons remain distinct from prospective claims.

Table 26: Strict joint label-and-complete-evidence success on the same FEVEROUS $K = 64, m = 4$ discovery pools. Intervals are paired bootstrap intervals from DPP; p is Holm-corrected within five matched comparisons. Borda and BoM share an endpoint vector but remain separate family members.

Method	Successes	Rate	Δ [95% interval]	W/L	Holm p
DPP	90	.2250	—	—	—
DAKPO	87	.2175	-.0075 [-.0250,.0075]	4/7	1.000
DAKPO-CG	87	.2175	-.0075 [-.0225,.0075]	3/6	1.000
Borda / BoM	95	.2375	+.0125 [-.0100,.0350]	13/8	1.000
Self-Certainty top-4	102	.2550	+.0300 [-.0050,.0650]	31/19	.595

Table 27: LiveAoPS endpoint. W/L is selector wins/losses against DPP; nO is the count of missed-solvable pools among $n = 1{,}200$; exact p is two-sided.

Seed/method	DPP	Sel.	W/L	nO	κ	exact p
1601 primary	.1767	.1808	12/7	64	.078	.359
1602 primary	.1808	.1800	4/5	61	-.016	1.0
1601 strict	.1767	.1775	2/1	64	.016	1.0
1602 strict	.1808	.1808	0/0	61	.000	1.0

Table 28: Auxiliary robustness and boundary routes. These are descriptive and do not support cross-domain dominance. The N=100 Cross-MATH route has a different frozen verifier/configuration from the N=200 target-control audit. Same-pool DQO/SetPO/QEMPO/GCPO proxies are controls, not post-training reproductions; no universal dominance is claimed, and a faithful large-N evaluation remains open.

Route	Method	DPP pass	Sel. pass	Target change
Cross MATH (N=100)	DAKPO	.5800	.5800	-.0077
LiveCodeBench TOP	DAKPO-CG	.6750	.6800	-.0001
HumanEval+ single	DAKPO	.8415	.8415	.0000
HumanEval+ multi	DAKPO	.9146	.9146	.0000
DSCodeBench (Ouyang et al., 2026)	beam replay	.3500	.3625	-.0056
2WikiMultiHopQA	DAKPO	tie	tie	movement only

Table 29: Complete LiveAoPS frozen-method secondary audit. IUT is the maximum of the two same-prompt generation-seed exact p-values; Holm adjusts over all nine non-DPP methods. The right pair uses two-sided per-seed tests as a registered-test-convention sensitivity, not a registered nine-method family.

Method	One-sided IUT	Holm-9	Two-sided IUT	Holm-9
Constrained DPP	1.0000	1.0000	1.0000	1.0000
MMR	.5000	1.0000	1.0000	1.0000
Quality-only	.4427	1.0000	.8854	1.0000
Best-of-Majority	.0270	.2163	.0541	.4326
Verifier-Borda	.00330	**.02970**	.00660	.05940
Marginal pessimistic	.9558	1.0000	.1934	1.0000
Original DAKPO	.2744	1.0000	.5488	1.0000
CC-MR DAKPO	.7461	1.0000	1.0000	1.0000
Coupled-regret DAKPO	1.0000	1.0000	1.0000	1.0000

K WORKED DECISIONS AND A PROSPECTIVE-USE GUIDE

This guide uses constructed confidence sets, reported synthetic calibration counts, and planning scenarios to explain practical decisions. It adds no protected LLM result. The supplement's `practitioner_worked_examples.py` reproduces the numbers without endpoint payloads.

K.1 A SHARP INTERVAL–SIMPLEX DECISION WITH AN ADVERSE WITNESS

Consider the always-deploy outer gate, so $B = L$. Suppose a valid joint confidence event places the three disjoint masses (M, R, B) in intervals $[\ell_M, u_M]$, $[\ell_R, u_R]$, $[\ell_B, u_B]$. Intersect their rectangle with $m + r + b \le 1$. If nonempty, this set has

$$\bar{b} = \min\{u_B, 1 - \ell_M - \ell_R\}, \qquad M - B \ge \ell_M - \bar{b},$$
$$\Delta \in \left[\ell_R - \bar{b}, \min\{u_R, 1 - \ell_M - \ell_B\} - \ell_B\right].$$

Both lower extrema are attained by the feasible witness $(m, r, b, t) = (\ell_M, \ell_R, \bar{b}, 1 - \ell_M - \ell_R - \bar{b})$, where t is the remaining probability mass. Sharpness is relative to this supplied joint set, not every possible statistical procedure. A nonpositive lower bound therefore has a compatible law; it does not establish impossibility.

Table 30: Constructed interval inputs, not new empirical confidence intervals. Gain admission and capture-budget ordering are separate, potentially simultaneous decisions.

Case	M	R	B	$\bar{b}$	Gain range	$\inf(M - B)$
Capture only	[.50,.60]	[.10,.20]	[.05,.55]	.40	[−.30,.15]	.10
Gain only	[.10,.20]	[.50,.60]	[.05,.55]	.40	[.10,.55]	−.30
Both	[.35,.45]	[.35,.45]	[.05,.40]	.30	[.05,.40]	.05

In the first row, the rectangle alone gives capture margin $.50 - .55 = -.05$. The simplex caps harm at .40 and raises the margin to .10. Yet the witness $(.50, .10, .40, 0)$ still has gain $-.30$. Exchanging M and R in the second row instead certifies gain .10. The third row supports both positive gain and a larger missed-capture budget. These constructed cases show that simplex tightening can matter. It gave no additional tightening in the empirical nine-row audit.

K.2 PAIRED RISK: A NUMERICAL COMPARISON WITHOUT CONFLATING GUARANTEES

This example compares valid bounds on the same paired estimand and their resulting release decisions.

Condition on the synthetic discovery fit. Let $X = U_{\text{routed}} - U_{\text{champion}}$ and $\ell = (1 - X)/2$. The reported calibration counts $(n, r, h) = (800, 47, 2)$ give $\bar{X} = .05625$ and $\bar{\ell} = .471875$. Two one-sided CP bounds with miscoverage .025 give $r_L = .0434830$, $h_U = .0090014$, and $\Delta_L = r_L - h_U = .0344816$. A generic fixed-rule Hoeffding bound instead uses $U_\ell = \bar{\ell} + \sqrt{\log(20)/(2n)} = .5151455$, or $\Delta_L = 1 - 2U_\ell = -.0302909$.

Table 31: Same counts, same fixed-rule paired estimand, same 95% confidence level. This is a CP–Hoeffding precision comparison, not a native CRC contest.

Bound	Risk upper	Gain lower	Strict-positive release
Two one-sided CP rate bounds	.4827592	.0344816	Admit
Bounded-loss Hoeffding	.5151455	−.0302909	Retain champion

Native conformal risk control (CRC) targets marginal expected risk. It requires pointwise-monotone losses or a valid monotone envelope (Angelopoulos et al., 2024). Nested gates alone do not suffice. Removing a rescue raises paired loss from 0 to .5; removing a harm lowers it from 1 to .5. For a two-action adaptation, let action 0 deploy the fixed routed policy and action 1 retain the champion. The suffix-supremum construction gives

$$\tilde{\ell}_0 = \max\{(1 - X)/2, .5\} = .5 + .5\mathbf{1}\{X = -1\}, \qquad \tilde{\ell}_1 = .5.$$

With loss bound 1 and risk target .5, CRC's corrected empirical criteria are $402/801 = .501873$ and $401/801 = .500624$. The feasible set is empty, so the known-safe champion is the fallback. This transparent envelope discards all rescue credit and is not an optimized CRC baseline. Neither its refusal nor Table 31 establishes general superiority over CRC. The audit answers additional questions about the gain permitted by a public state and the remaining repair budgets. A marginal risk guarantee does not by itself answer those questions.

K.3 When full-pool opportunity is not observable

A lower proxy for missed capture can sometimes resolve the repair-budget ordering without labels for every candidate.

For selected-set coverage, $M = O - R = \Pr(H = 1, Y_S = Y_B = 0)$. If a fixed, predeclared subset of candidates is correctness-audited, its coverage indicator H_J obeys $H_J \le H$. Consequently

$$M_J := \Pr(H_J = 1, Y_S = Y_B = 0) \le M.$$

A simultaneous lower bound on M_J and upper bound on harm L can therefore certify missed capture exceeding the gate budget whenever $\ell_{M_J} > u_L$, since total downstream repair relative to always deploying the proposal is $R - \Delta = L$. For another fixed gate, compare with a bound on $R - \Delta_\alpha$, not automatically with L. Report the result as a lower bound on missed capture, not a point estimate of full-pool opportunity or $O - R$. Failure of the inequality is inconclusive.

An imperfect proxy W needs a separate soundness bound. If $\Pr(W = 1, H = 0) \le \varepsilon$, then

$$M \ge \Pr(W = 1, Y_S = Y_B = 0) - \varepsilon.$$

The error bound needs a domain argument or independent calibration, with its failure probability included in the joint budget. Verifier scores, public-example passes and partial inspections do not automatically bound error on the declared endpoint. Without full coverage labels, a one-sided proxy, or a justified proxy-error bound, report only paired R, L, Δ and identifiable frontiers. A numerical capture-loss certificate is then unavailable.

K.4 Retrospective partial-label validation on FEVEROUS discovery

Masking released discovery labels tests how much inspection the budget ordering requires.

Table 32: Retrospective FEVEROUS masking, all five proposals. q^* is the smallest tested extra budget with positive lower diagnostic margin for $M > L$. Inspections include the selected union and are compared with $400 \times 64 = 25{,}600$ exhaustive candidate labels. Margins use the nominal fixed-family 95% i.i.d. model, not prospective validation; retention is descriptive.

Proposal	q^*	Found / full misses	Inspections	Full cost (%)	Margin LB (pp)
DAKPO	16	41/89	6,898	26.95	+1.454
DAKPO-CG	16	41/90	6,852	26.77	+1.834
Borda	32	66/80	12,164	47.52	+6.218
Best-of-Majority	32	66/80	12,164	47.52	+6.218
Self-Certainty	64	62/62	18,734	73.18	+1.623

The analysis uses the supplement's two text-free discovery projections, retaining all 400 pools and five proposals without calibration/test access. The masking plan was fixed on 2026-09-09, before computation but after aggregate outcomes were known. This is secondary analysis, not blinded annotation or pre-endpoint registration. Outcomes use the original strict joint label and complete-evidence endpoint, not a verifier proxy.

Let $U_i = B_i \cup S_i$ be the union of the two selected four-candidate sets. Inspect all its unique members. Let $J_i(q)$ contain the first $\min(q, 64 - |U_i|)$ indices outside U_i, in the original candidate order. Only if both sets fail, inspect this entire extra prefix. No success-based early stopping is credited. The two per-unit quantities are

$$W_i(q) = \mathbf{1}\{Y_{B,i} = Y_{S,i} = 0\}\mathbf{1}\{\exists j \in J_i(q) : C_{ij} = 1\},$$

$$T_i(q) = |U_i| + \mathbf{1}\{Y_{B,i} = Y_{S,i} = 0\} \min(q, 64 - |U_i|).$$

Pointwise $W_i(q) \le \mathbf{1}\{H_i = 1, Y_{B,i} = Y_{S,i} = 0\}$, so $M_q = \mathbb{E}W_i(q) \le M$. Conditional inspection skips labels irrelevant to this event; its probability still uses all 400 units as the denominator. At $q = 64$, $M_q = M$. Cost can nevertheless stay below 400×64, because a successful selected set needs no complementary inspection. This is an exhaustive residual audit, rather than full inspection of every pool.

Each positive budget–method pair receives a one-sided CP lower bound on M_q; each method receives an upper bound on harm L. The $5 \times 7 + 5 = 40$ tails each spend .00125. If the family is fixed independently of outcomes, this gives at least .95 simultaneous coverage under the declared i.i.d. unit model. Masks and method endpoints need not be independent.

Here the analysis followed known aggregate outcomes, so the bounds are nominal fixed-family diagnostics. They do not cover historical route selection, retrospective design choices, or all research attempts unconditionally. The pointwise proxy inequality itself requires no i.i.d. assumption. We use no normal approximation, candidate-level independence, endpoint-trained ordering, missing-unit exclusion, or gate refitting.

Table 33: Complete fixed masking grid. Each cell reports found missed opportunities / total unique candidate inspections, including paired union. These are retrospective counts, not actual annotation expenditures.

Extra budget	DAKPO	DAKPO-CG	Borda	BoM	Self-Certainty
0	0 / 2,002	0 / 1,940	0 / 2,660	0 / 2,660	0 / 2,921
1	0 / 2,308	0 / 2,247	0 / 2,957	0 / 2,957	0 / 3,200
2	0 / 2,614	0 / 2,554	0 / 3,254	0 / 3,254	0 / 3,479
4	0 / 3,226	0 / 3,168	0 / 3,848	0 / 3,848	2 / 4,037
8	23 / 4,450	23 / 4,396	20 / 5,036	20 / 5,036	15 / 5,153
16	41 / 6,898	41 / 6,852	34 / 7,412	34 / 7,412	27 / 7,385
32	74 / 11,794	75 / 11,764	66 / 12,164	66 / 12,164	50 / 11,849
64	89 / 20,093	90 / 20,124	80 / 19,702	80 / 19,702	62 / 18,734

Table 34: Lower margins $\ell_{M_q} - u_L$, percentage points, for the same 40-tail secondary family. Only strictly positive entries certify $M > L$. No budget or method is removed after seeing its result.

Extra budget	DAKPO	DAKPO-CG	Borda	BoM	Self-Certainty
0	-4.749	-4.370	-5.119	-5.119	-8.870
1	-4.749	-4.370	-5.119	-5.119	-8.870
2	-4.749	-4.370	-5.119	-5.119	-8.870
4	-4.749	-4.370	-5.119	-5.119	-8.857
8	-1.921	-1.541	-2.808	-2.808	-7.372
16	+1.454	+1.834	-.273	-.273	-5.327
32	+8.300	+8.895	+6.218	+6.218	-.862
64	+11.576	+12.177	+9.230	+9.230	+1.623

For DAKPO, the paired union costs 2,002 inspections and 306 units have joint failure. At $q = 16$, total cost is $2002 + 306 \times 16 = 6898$; 41/89 misses are found (46.07%). The lower missed-capture rate .062033 and harm upper rate .047489 give margin .014544. This budget uses 26.95% of exhaustive all-pool labels, or 34.33% of this method's 20,093-inspection exhaustive residual audit. Self-Certainty at $q = 32$ finds 80.65% of its misses but does not certify the order, demonstrating that retention alone is insufficient.

Both tables count inspections separately by proposal. Shared labels could be reused, so the columns cannot be summed as disjoint costs. Canonical candidate order may be inefficient; we did not optimize it after replay. The results validate the masking measurement on one released discovery set. They do not establish a new release, actual annotation-time savings, or diagnosis-guided model improvement. The supplement contains `partial_label_missed_capture.py`, its synthetic tests, the full JSON output and the pre-replay plan. Future prospective partial-label audits must freeze their family and inspection policy before their own endpoint access.

K.5 A PILOT-FREE DESIGN AND A PROSPECTIVE REFINEMENT SKETCH

Activity and public-state resolution. A pilot-free activity envelope can use a known structural event $A(G)$. It must satisfy $D \neq 0$ implies $A(G)$, and its target-law probability needs a justified upper bound. For example, identical returned actions have identical utility under a deterministic endpoint; only their complement can be active. A sample frequency of differing answers is not a

population upper bound. Without a useful structural bound, freeze $\bar{\nu} = 1$ or predeclare an exact discordance-count certificate that estimates activity and orientation separately. A pilot-derived bound uses its own error budget. None of these options permits a post hoc switch to a more favorable certificate. Among protected calibration rows with nonzero discordance, observed activity ranges from $1/2800$ (HellaSwag) to $454/4500$ (MedMCQA Stage-55). Relative to those hindsight-tight rates, $\bar{\nu} = 1$ inflates the leading sample term by about $2800\times$ to $9.9\times$; these retrospective ratios are not valid design envelopes or prospective sample-size guarantees.

A lossless, invertible serialization preserves the sigma-field and Γ_G. A fitted tokenizer or gate may still be sensitive to its encoding. Coarsening instead changes the estimand to Γ_Z. Before calibration, freeze normalization, invalid-parse handling, cell boundaries, and unseen-cell fallback. Compare a small discovery-selected partition family. Merge sparse cells on discovery or assign them the baseline. No universal minimum cell count guarantees the correct sign: orientation near zero requires more data regardless of occupancy. Report cell sizes, held-out bounds, and the partition penalty with plug-in frontier changes (Figure 3). The finest partition alone is insufficient.

Strong-control validation after an unsuccessful repair. The completed MedMCQA successor improved direct answering but failed the unchanged-plurality control (Appendix I). A new repair study should anchor calibration and its primary test to a strong comparator fixed before new outcomes. In the same eight-view setting, unchanged plurality can serve as the baseline. A discovery-trained policy may depart from it using public features, with missing states returning the baseline. Before independent calibration, freeze one policy, its features, fitting choices, parser, ties and final-answer utility. Completed successor groups and sealed historical tests cannot serve as fresh confirmation.

With $D^* = Y_{\text{repair}} - Y_{\text{plurality}}$, apply the fixed-gate paired certificate directly to D^*. Calibration authorizes this frozen contrast alone. Refusal keeps the new test sealed; admission permits one evaluation at a separately fixed test size. Before collecting outcomes, prespecify the smallest meaningful effect and power scenarios for plausible activity and orientation. Direct answering supplies context but cannot rescue failure against plurality.

To test representation refinement, retain an unchanged representation with the same pool, action family and fitting budget as a separately budgeted control. Beating plurality establishes repair benefit, but does not isolate the causal value of the diagnostic or feature change. This is a validation design. The separate confidence-feature study below completed development and stopped before confirmation.

Historical calibration-power example. This example uses the earlier exploratory 76/43 cell to show sensitivity to effect shrinkage. It does not give power for the new strong-control contrast. The executed successor used fixed C/T sizes 10,000/20,000 and required superiority over both direct and plurality. The older 4,000 Stage-55 test groups remained sealed.

For a *fixed* gate, let discordance count $K_d \sim \text{Bin}(n, q)$ and rescue count $R_d \mid K_d = k \sim \text{Bin}(k, \theta)$. Let $c_k = 1 + F^{-1}_{\text{Bin}(k,.5)}(.975)$. Its exact calibration admission probability is

$$\mathcal{P}_C(n; q, \theta) = \sum_{k=1}^{n} \Pr(K_d = k) \Pr(R_d \geq c_k \mid K_d = k).$$

The activity lower bound is strictly positive whenever $K_d > 0$. Table 35 fixes $q = 119/4500$ and compares two orientation scenarios. The first copies the retrospective 76/43 cell as an optimistic planning input, not a validated population rate. The second halves gain while holding activity fixed. This power calculation excludes discovery failure, test rejection, model changes, and shift. It gives neither end-to-end success probability nor a recommended sample size.

Table 35: Calibration-only power sensitivity for a fixed successor gate. These assumptions are scenarios, not fresh LLM measurements.

Scenario	θ	Assumed gain (pp)	$n = 4000$	8000	16000
Retrospective-rate input	.63866	.7333	.7936	.9795	.9999
Half gain, same activity	.56933	.3667	.2639	.4962	.8024

For comparison, deterministic replication of the *unrefined* 230/224 mix first raises its nominal orientation bound above .5 at $49\times$: $n = 220{,}500$, 11,270/10,976, lower bound .500015. This fixes future counts and is not a power calculation. The sensitivity table shows why plausible representation gain can matter more than reducing uncertainty around nearly balanced orientation. It does not predict that the new gate will reproduce the retrospective effect.

K.6 Monitoring: an alarm is not a transport certificate

The simplest renewal evaluates the unchanged gate on a fresh labeled window, with a prespecified error budget across monitoring looks. Alternatively, consider a TV bound for finite frozen Z and binary-endpoint contrast $D \in \{-1, 0, 1\}$. Obtain simultaneous probability intervals $[\ell_c^P, u_c^P]$, $[\ell_c^Q, u_c^Q]$ for each joint cell $c = (Z, D)$ under the old and new laws. On that joint confidence event,

$$\mathrm{TV}(P_{Z,D}, Q_{Z,D}) \le \bar{\varepsilon} := \min\left\{1, \frac{1}{2}\sum_c \max\{u_c^P - \ell_c^Q, u_c^Q - \ell_c^P\}\right\}.$$

For a fixed Z-measurable gate, this transfers bounded paired gain with penalty $2\bar{\varepsilon}$. Bounding TV of the entire public record is unnecessary. The intervals do require representative labels under both laws, and may be loose when cells are numerous. Directly certifying gain on the new window is often simpler. Stable public-feature frequencies alone cannot bound changes in rescue/harm orientation. Without labeled renewal or a justified joint-law bound, retain the baseline. An unlabeled drift alarm supplies no certificate.

L FRESH CONFIDENCE-FEATURE STUDY: A DEVELOPMENT-STAGE STOP

Question and separation from prior studies. Could public token confidence help a compact learned selector beat unchanged plurality? We tested this question in the eight-view MedMCQA setting, fixing the protocol and implementation locally before allocation or generation. From the existing 57,562 eligible source/duplicate groups, we excluded all 42,200 groups assigned to the three earlier studies, including their unopened tests. The new allocation reserved 1,500 training, 1,500 selection, 6,000 calibration, and 6,000 test groups; 362 remained unused. Freshness refers only to recorded allocations, not pretraining or all prior exposure. The commitment had neither an independent custodian nor an external timestamp.

Fixed generation and selector panel. We used Qwen3-4B-Instruct-2507, eight balanced option permutations, serial batch-one FP16 generation, temperature .7, top-p .8, top-k 20, and at most 96 new tokens per view. All selectors reused the same pools. Original forward passes supplied raw, pre-warping logits for mean selected-token log probability and Self-Certainty (SC), excluding prompt and special tokens. Production used neither teacher-forced replay nor a second model. The controls were direct view0, plurality, log-probability best-of-eight, SC top-1, SC-rank Borda with rank power .5, and adaptive Majority-of-the-Bests (MoB) with either SC or log probability at $q = .75$. Confidence selectors excluded invalid parses and returned no answer for an all-invalid pool. This validity adapter and local Borda ties differ from the baselines' full original benchmark protocols.

The coarse head used six vote/parse features. The rich head added twelve within-pool confidence summaries (six each for log probability and SC). Both heads selected only observed pool answers. Group-weighted L2 logistic regression used $C \in \{.1, 1, 10\}$, the same 1,500 training groups and the same fitting budget. Each group's distinct answers shared unit total weight. Standardization used training data only. Selection accuracy chose one head per representation, breaking ties toward smaller C, with no subsequent refit. Fitting and development generation add costs, so total compute need not equal direct answering.

Prespecified nomination rule and observed stop. Nomination required all three conditions: positive training gain over plurality, at least five selection net wins over plurality, and positive selection gain over the selected coarse head. These heuristics allocated development budget; they were not independent efficacy tests. All 3,000 pools (24,000 views) preceded development-label access. Both selected heads used $C = 1$. The rich head had training rescue/harm counts 34/28, but selection counts 30/41 versus plurality ($-.733$ pp) and 30/42 versus the coarse head ($-.800$ pp). The latter two conditions failed. Table 36 retains every head and control, including the small coarse and Borda gains.

Evidence boundary and operational accounting. The terminal state was `DEVELOPMENT_NO_GO`. The 12,000 calibration/test groups remained ungenerated and unopened. Borda's one-net-win lead did not trigger promotion, feature expansion or a changed threshold. Had nomination passed, confirmation would have tested the frozen rich policy against plurality, then the coarse control and a fixed competitive family. None of these tests occurred. This stop establishes neither zero population value for confidence nor method equivalence. It does not explain the training/selection difference or validate diagnosis-guided repair.

A disk-full interruption occurred after 1,985 complete pools; recovery reused all 15,882 saved views under the unchanged scientific freeze. The remaining 1,015 pools completed in approximately 26.9 minutes of resumed collection. That duration excludes earlier collection, restart and fitting; it is not total runtime or a speedup. The operational record retains the interruption. The supplement includes the protocol, complete report, terminal state and implementation. The table summarizes aggregates without reanalyzing protected labels.

Table 36: Complete development comparison, with plurality as reference. Each split has 1,500 groups. R/H denotes rescues/harms; gain is $100(R-H)/1500$ percentage points. Training counts are shown for all fitted heads; controls are evaluated on selection. These are development results, including model selection, not independent confirmation or equivalence tests.

Selector	Training R/H	Selection R/H	Selection gain (pp)
Plurality	–	0/0	.000
Direct view0	–	64/80	−1.067
Log-probability best-of-eight	–	38/54	−1.067
SC top-1	–	33/49	−1.067
SC-rank Borda	–	14/13	+.067
MoB (SC)	–	25/36	−.733
MoB (log probability)	–	26/31	−.333
Coarse, $C = .1$	26/16	23/26	−.200
Coarse, $C = 1$ (selected)	23/10	20/19	+.067
Coarse, $C = 10$	20/8	18/19	−.067
Rich, $C = .1$	28/30	27/40	−.867
Rich, $C = 1$ (selected)	34/28	30/41	−.733
Rich, $C = 10$	33/27	28/46	−1.200

Label-free cost control for the retained baseline. After the stop, we fixed a retrospective replay rule before measuring its savings. After t of eight views, stop if the leading valid answer count c_1 satisfies $c_1 > c_2 + 8 - t$, where c_2 is the largest rival count. Otherwise finish all eight and use the unchanged tie rule. Every rival can gain at most $8 - t$ votes, so an early answer equals full-pool plurality for every possible suffix. Invalid outputs consume a slot without voting. This elementary baseline cost control makes no novelty claim or comparison with ESC (Li et al., 2024).

The replay uses public answers and token counts, without labels or model calls. It preserves all 3,000 actions. Exhaustive checks of all 5^8 answer/invalid sequences also found no disagreement. Generation calls would fall from 24,000 to 17,170 (28.46% fewer); training and selection halves save 28.57% and 28.35%, respectively. Prompt tokens would fall from 2,265,604 to 1,620,512, and generated tokens including EOS from 96,000 to 68,680. The stopping counts at $t = 5, 6, 7, 8$ are 1,947/332/325/396. Each view emitted four tokens including EOS, so these results concern short answers. The exact counterfactual counts do not measure prospective latency, energy or accuracy, and establish no SOTA. Online equivalence requires unchanged per-view generation; full historical collection costs were already incurred. The rule applies to plurality deployment, not shared-pool collection or confidence selectors needing all views. The failed nomination stands.

Measured warm-runtime qualification. A separate label-free run fixed 96 development groups by hashed IDs (48 from each original half) before loading the model. Full-eight and early-stop arms alternated full-first/early-first order, 48 pairs each. Both used the persistent Qwen3-4B model and original decoding, without confidence-logit capture. MPS-synchronized timings include prompting, tokenization, generation, parsing and stopping. They exclude disk writes and later trace checks. On one shared M4 Max session, cumulative warm times were 134.75 versus 97.69 seconds (27.50% lower; ratio 1.38), with 768 versus 556 generation calls. Reductions in the two order strata were 31.07% and 23.76%. Every executed view matched its cached raw trace. All 96 actions agreed; no pair failed or was repeated. Model loading took 4.36 seconds and eight synthetic warmup calls took 1.78 seconds, separately counted. The complete qualification took 238.72 seconds. This engineering check uses exposed short-answer development data. It neither tests accuracy independently nor guarantees latency on an isolated host or superiority over adaptive samplers. The failed nomination is unchanged. The supplement retains all timings, protocol, source snapshots and arithmetic replay.

L.1 TIGHTENING THE STOPPING CONDITION WITH THE ORIGINAL TIE RULE

The strict-lead rule can sample unnecessarily. If four initial votes agree, the remaining four can only tie; the original priority may already favor the leader. A second label-free retrospective analysis fixed a refinement before measuring savings. It preserves view order, the eight-view budget, parser and anchored answer priority. Every final plurality decision and the failed development result remain unchanged.

Let $r = 8 - t$ be the remaining slots, c_b the observed valid votes for answer b, and a the current positive-count leader under the fixed tie priority $\prec$ (earlier wins). Stop exactly when, for every $b \neq a$,

$$c_a > c_b + r \quad \text{or} \quad (c_a = c_b + r \text{ and } a \prec b).$$

An all-invalid prefix returns no answer only after all eight slots. Every invalid output consumes a slot without adding a vote. The priority is computed from the original public anchor and group identifier, never chosen from the observed answers or their correctness.

Exactness and the restricted minimality statement. The condition is sufficient because a rival can gain at most r votes, and any reachable tie favors a. It is also necessary for stopping from that prefix without any suffix assumptions: continuing with only invalid outputs leaves a as the final answer, whereas a violating rival b wins if all remaining outputs vote for b. If the prefix has no valid vote and $r > 0$, all-invalid and all-b continuations likewise have different answers. Whenever the rule continues, two admissible suffixes give different final actions. No rule using only that prefix and fixed priority can guarantee the full-budget answer for every suffix. This proves earliest stopping within that class, not novelty, expected-time optimality under a probability model, or latency or accuracy SOTA. Each suffix slot permits four answers or an invalid output. Additional constraints on future outputs could change the bound.

An independent recursive oracle enumerated reachable prefix counts and terminal winners without using the stopping inequality. Across 1,287 states and all 24 priorities (30,888 checks), the formula stops if and only if that winner set is a singleton, including terminal abstention.

Table 37: Exact call counts on the same public development cache. All three rules produce identical full-eight plurality actions on every group. Savings are relative to full-eight calls; these are retrospective counts.

Groups	Full eight	Strict lead	Tie aware	Tie-aware saving
Training (1,500)	12,000	8,572	8,054	32.88%
Selection (1,500)	12,000	8,598	8,053	32.89%
All (3,000)	24,000	17,170	16,107	32.89%

The refinement stops earlier on 883 groups, saving another 1,063 calls (6.19% of strict-lead calls). Its counts at $t = 4, 5, 6, 7, 8$ are 496/1,684/281/295/244. Prompt tokens fall to 1,519,876 and generated tokens including EOS to 64,428. The supplement releases every replay row and original answer priority, with no source identifiers or correctness labels. The 96-pair latency measurement covers only strict-lead stopping; refined-rule latency is unmeasured. This plurality-only cost control cannot replace common-pool collection or implement a confidence selector. It claims no accuracy gain over plurality. Comparing adaptive methods such as ESC needs separate matched accuracy–cost experiments; local minimality proves no superiority over them.

M OPTION-DISTRIBUTION REFINEMENT: A SECOND DEVELOPMENT STOP

Separate population and fixed allocation. The next unfitted selector used a distinct MedMCQA population: single-answer questions with valid question, subject and four distinct options, but missing topic metadata. The model never receives topic metadata. Public-only screening against the entire earlier metadata-complete frame left 55,026 rows in 52,442 groups. Groups are transitive components of equal normalized IDs, equal normalized questions, or character-5gram Jaccard similarity at least .9 on normalized question plus sorted options. Exact prefix-filtered comparisons implement this relation without establishing semantic independence or excluding pretraining exposure. Before generation, one local random anchor fixed group ordering, representatives and disjoint development/calibration/test allocations of 1,000/4,000/7,000 groups. This eligibility law differs from earlier studies and the standard leaderboard, so their results cannot be pooled. The authors recorded the design and controlled staged label access; these steps were not independently administered.

One fixed selector and its public signals. Generation used the same Qwen3-4B snapshot and eight-view settings as Appendix L. At the sampled answer token, we kept the four existing displayed-letter logits, normalized over those options and inverted the known permutation. No extra production forward pass was needed. The contract required exact `FINAL: <letter>` plus EOS tokens. Any missing vector made all new methods fall back to original plurality, retaining the group. Development required no such fallback.

The target maximizes mean semantic option log probability over sampled answers. Its matched information control votes over eight per-vector argmax answers within the same observed support. Arithmetic option-probability averaging and sampled-answer probability-weighted voting join all seven earlier controls. An unrestricted geometric aggregator over all four options is a diagnostic, not an alternative eligible for promotion. New methods share the original anchored tie priority. There is no fitting, temperature search, post-development refit or winner substitution.

Changing both the proposal and its signals does not isolate fixed-proposal information overlap. The unrestricted diagnostic can choose unsampled answers, so the original pool-coverage ceiling does not apply. Its observed agreement with the target does not make the rules identical on every pool.

Permutation aggregation and option-ID debiasing are established ideas (Zheng et al., 2024); averaging logits is not our algorithmic contribution. Indeed, mean log-softmax scores and mean raw logits have identical rankings. For logits $z_j(a) = s(a) + b_{\pi_j(a)}$, where $\pi_j(a)$ is the displayed position of semantic answer a, the eight balanced permutations place each answer twice in every position, so their mean adds the same bias constant to every $s(a)$. This identity neither shows that Qwen follows the decomposition nor guarantees correctness. Our probability vote differs from native CISC's confidence extraction and temperature normalization (Taubenfeld et al., 2025); it is not a native reproduction.

Predeclared nomination and outcome. All 1,000 pools and actions were fixed before one development-label join. Nomination required at least ten target net wins over plurality, at least five over the argmax control, and target accuracy no lower than all eight competitive controls. The target achieved 62.5% accuracy versus plurality's 62.2% (25 rescues, 22 harms; $+.3$ pp), failing the first condition. It passed the argmax hurdle (28/20; $+.8$ pp) but failed the competitive condition, including MoB-logp (19/25; $-.6$ pp). These budget heuristics are not confirmatory tests. No control was promoted afterward.

Table 38: Complete option-distribution development panel ($n = 1,000$). R/H and gain compare the fixed geometric target (62.5% accuracy) against each row. Positive gain favors the target. Intervals are marginal 95% two-rate CP intervals, using four .0125 tails; they are not simultaneous panel certificates. All results are development-only.

Comparator	Acc. (%)	R/H	Gain (pp)	95% interval (pp)
Plurality	62.2	25/22	$+.3$	$[-1.96, 2.56]$
Direct view0	62.8	55/58	$-.3$	$[-3.68, 3.08]$
Log-probability best-of-eight	61.7	29/21	$+.8$	$[-1.52, 3.11]$
SC top-1	62.8	22/25	$-.3$	$[-2.56, 1.96]$
SC-rank Borda	62.5	16/16	.0	$[-1.90, 1.90]$
MoB (SC)	62.6	15/16	$-.1$	$[-1.97, 1.77]$
MoB (log probability)	63.1	19/25	$-.6$	$[-2.78, 1.59]$
Aligned argmax vote	61.7	28/20	$+.8$	$[-1.48, 3.07]$
Arithmetic option mean	62.5	17/17	.0	$[-1.95, 1.95]$
Sample-probability vote	62.6	18/19	$-.1$	$[-2.13, 1.93]$
Unrestricted geometric diagnostic	62.5	0/0	.0	$[-.44, .44]$

What the stop does and does not show. The terminal is `DEVELOPMENT_NO_GO`. The 11,000 calibration/test groups remain ungenerated and unopened. Table 38 retains all comparisons. The nominal one-sided exact McNemar $p = .3854$ against plurality is descriptive, not a confirmatory rejection or an equivalence test. The target and unrestricted diagnostic chose identical answers on all 1,000 groups, giving no evidence that restricting support improved accuracy. The exact-format fallback caused no failures. These observations describe the stop without proving its population mechanism or excluding a smaller true gain.

Had development nominated the target, calibration would have required a positive factorized 95% paired-gain certificate. Test would then have required that certificate and exact McNemar against plurality, followed by the argmax control and a Holm-adjusted eight-comparator family in fixed sequence. None of these stages was reached. Their inference would still require fresh source-group and same-law assumptions; grouping and sampling without replacement do not establish them. This local freeze supplies no global error-rate guarantee across the research history.

Cost and reproducibility. The study completed all 8,000 generation calls, consuming 747,592 prompt tokens and 32,000 generated tokens including EOS. Collection took 1,545.79 seconds (25.76 minutes); recorded model-load-through-terminal runtime was 1,548.35 seconds. Timings exclude preprocessing and qualification: 16 synthetic forwards, then 16 synthetic generation calls in eight matched pairs. All paired traces and prior confidence summaries matched exactly. One persistent model sampled serially, completing every pool without early stopping. These results claim neither equal compute with direct answering nor latency SOTA. The supplement contains complete reports, protocol, code snapshots and aggregate statistical and cost checks. Its offline entry point does not regenerate label joins or model outputs.

This second stopped refinement documents another falsifiable repair attempt. Independent end-system improvement over strong controls remains unestablished.

N REASONING DEVELOPMENT AND UNTOUCHED CONFIRMATION

Question and comparison target. A short contrastive prompt asks the model to distinguish the two most plausible options before answering. We compared it with ordinary short reasoning and twelve answer-only pool controls. Final-answer accuracy is primary; latency and tokens may be traded off and are reported separately. This intervention changes generation, so it cannot isolate the fixed-proposal DAKPO identity or attribute benefit to that component. The controls are declared local implementations, not each published method's best configuration. The missing-topic population differs from the official MedMCQA leaderboard.

Disclosed development sequence. The first pilot reused 128 groups from the earlier option-distribution development set. With its frozen parser, contrastive reasoning answered 71/128 correctly, ordinary reasoning 76/128 and cached plurality 79/128; the recorded terminal was `PILOT_NO_GO`. The parser rejected 22 contrastive and two ordinary responses despite normal EOS endings: each had a unique explicit answer on the same line or with different capitalization. We applied a revised parser equally to both complete saved output sets. It accepts case/whitespace variants but requires one explicit terminal `FINAL:` marker and one A–D choice. Ambiguity, trailing explanation and truncation remain invalid. The post-hoc sensitivity was 83/128 versus 78/128; no valid action changed and no generation was added. This post-hoc scoring sensitivity neither establishes independent improvement nor reverses the no-go.

We froze the repaired contract before allocating 512 fresh development groups, excluding all 12,000 groups assigned to the option-distribution study. The target scored 316/512 (61.72%), ordinary reasoning 309/512 (60.35%), plurality 295/512 (57.62%), and the strongest pool control, MoB-log-probability, 303/512 (59.18%). Both reasoning arms were valid on all 512 groups. The target gained 1.367 percentage points over ordinary reasoning, with 38 rescues and 31 harms (nominal one-sided $p = .2352$). Other nomination conditions passed, but a discretionary two-percentage-point budget hurdle failed. The original `DEVELOPMENT_NO_GO` stands. The calibration 1,500 and test 3,000 groups were never generated or opened. The supplement retains all development contrasts.

After inspecting the complete result, we removed the budget hurdle for a separate confirmation. This choice preceded its new allocation, generation and labels but followed development; it is not retroactive preregistration. The method, parser and controls stayed unchanged. From the 52,442-group public-only census, we excluded all 12,000 old allocations and all 5,012 fresh-development-study allocations, including their unopened roles. A new random anchor selected 4,494 representatives from the remaining 35,430 groups, leaving 30,936 unused. There was no further calibration or selection. Author-managed label access and disjoint allocation do not establish independence, exclude pretraining exposure, or verify the same-law i.i.d. assumptions of paired inference.

Fixed execution and inference. The fixed Qwen3-4B-Instruct-2507 snapshot `cdbee75f17c01a7cc42f958dc650907174af0554` ran in FP16 on MPS with context 2,048, temperature .7, top-p .8 and top-k 20. Each reasoning arm used one call capped at 192 tokens, a common group-derived seed and the same answer contract. Eight balanced answer-only views, each capped at 96 tokens, supplied all twelve pool controls and their confidence/option-logit summaries without extra scoring forwards. The six orders of pool/plain/contrastive blocks each occurred 749 times. Invalid answers count as wrong in the full denominator. All 4,494 groups and actions were complete and hash-bound before one read of the new role's labels. Previously sealed roles stayed unopened.

The primary test was positive gain against ordinary reasoning and exact one-sided McNemar $p \leq .05$, with each reasoning arm valid on at least 4,450/4,494 groups. No minimum positive effect size was imposed. Passing both primary and quality conditions would open twelve Holm-adjusted pool comparisons at .05. This serial procedure controls within-study familywise error under its assumptions, not error across the research history. Separate simultaneous positive-gain certificates allocate .05/13 across contrasts. They differ from the serial tests and marginal intervals below. Optional stopping, added samples and post-test selection were prohibited.

Complete confirmation result. The target achieved 2,910/4,494 (64.753%) and ordinary reasoning 2,907/4,494 (64.686%). The paired difference is +0.0668 percentage points: 263 rescues and 260 harms, one-sided $p = .4652$, with a conservative marginal 95% gain interval $[-1.5214, 1.6548]$

Table 39: Final-answer accuracy on 4,494 newly reserved MedMCQA groups, using the same fixed model and missing-topic population. Δ is contrastive reasoning minus each declared inference configuration, in percentage points.

Method	Correct	Accuracy (%)	Δ (pp)
Contrastive reasoning	2,910	64.75	—
Ordinary short reasoning	2,907	64.69	+0.07
Plurality	2,842	63.24	+1.51
Direct view	2,795	62.19	+2.56
BoN (log-probability)	2,801	62.33	+2.43
Self-Certainty	2,827	62.91	+1.85
Borda-SC	2,863	63.71	+1.05
MoB-SC	2,857	63.57	+1.18
MoB (log-probability)	2,849	63.40	+1.36
Aligned argmax	2,834	63.06	+1.69
Aligned arithmetic	2,846	63.33	+1.42
Sample-probability vote	2,858	63.60	+1.16
Aligned geometric (pool)	2,864	63.73	+1.02
Aligned geometric (all options)	2,864	63.73	+1.02

percentage points. Valid counts were 4,479 and 4,484, respectively, so both met the quality condition. Three pool groups required the frozen option-distribution fallback. The final state is `COMPLETE_NO_PRIMARY_GAIN`; it does not establish equivalence. The target ranks highest in this local panel. A three-item lead establishes neither superiority over ordinary reasoning nor SOTA. Against plurality the observed difference is +1.5131 points, and against direct answering +2.5590 points. These differences remain reported. The failed primary closes the secondary family, so its p-values are descriptive.

Table 40: Complete paired comparisons. Conservative marginal 95% gain intervals subtract rescue/harm Clopper–Pearson bounds with four .0125 tails. They are neither simultaneous across comparisons nor inversions of the McNemar tests. The ordinary-reasoning comparison is the primary one-sided test; Holm adjusts only the twelve pool controls. The primary failed, so secondary values remain descriptive.

Comparator	Rescue/harm	95% interval (pp)	Nominal p	Holm p
Ordinary short reasoning	263/260	[-1.52, +1.65]	0.4652	—
Plurality	418/350	[-0.38, +3.40]	0.0078	0.0623
Direct view	403/288	[+0.76, +4.35]	$< 10^{-4}$	$< 10^{-4}$
BoN (log-probability)	430/321	[+0.56, +4.29]	$< 10^{-4}$	0.0004
Self-Certainty	416/333	[-0.02, +3.71]	0.0014	0.0135
Borda-SC	395/348	[-0.82, +2.91]	0.0457	0.1376
MoB-SC	394/341	[-0.68, +3.03]	0.0275	0.1376
MoB (log-probability)	402/341	[-0.51, +3.22]	0.0138	0.0830
Aligned argmax	421/345	[-0.20, +3.58]	0.0033	0.0301
Aligned arithmetic	402/338	[-0.44, +3.28]	0.0103	0.0718
Sample-probability vote	395/343	[-0.70, +3.01]	0.0302	0.1376
Aligned geometric (pool)	391/345	[-0.83, +2.88]	0.0486	0.1376
Aligned geometric (all options)	391/345	[-0.83, +2.88]	0.0486	0.1376

Engineering continuation and costs. A resource guard stopped collection before charging the next call. The 20-second log omitted the exact triggering resource sample, so the stop cannot be attributed specifically to RAM or disk. Before label access, one diagnosed continuation reused all 29,241 saved calls, including twelve synthetic qualification calls, and generated only the 15,711 missing calls. The original stop and 32,177 pre-recovery files were retained and hash-verified. Prompts, seeds, order, parser, model, sampling and statistical code stayed unchanged. Continuation released unused Python/MPS cache every sixteen new groups and logged any new guard trigger. It repeated no qualification, replaced no group and added no call budget.

The confirmation used 44,952 calls, including twelve synthetic calls; together with fresh development's 5,132 calls this is 50,084, within the original 50,132 allocation. The pilot added 260 calls; its cached pool was paid for by an earlier study, not free at deployment. Wall time from original model loading to terminal was 38,109.98 seconds, including the outage and reload; the continuation took 10,900.61 seconds. The eighteen-hour cap included the outage. Per-method times below count generation only. Costs need not match for an accuracy comparison, and these measurements establish no speedup. Manuscript integration added no model or benchmark calls.

Table 41: Measured generation costs. Direct is already included in the shared pool. Each pool selector uses the full pool; research collection counts it once. Timings exclude setup and synthetic qualification.

Protocol	Calls	Input tokens	Output tokens	Seconds
Contrastive reasoning	4,494	626,238	325,879	12,458.97
Ordinary short reasoning	4,494	513,888	285,666	10,810.45
Shared eight-view pool	35,952	3,392,090	143,823	7,639.35
Direct view	4,494	424,008	17,979	1,064.51

Interpretation and reproducibility. The development lead over ordinary reasoning was not confirmed. Both reasoning arms ranking above answer-only controls is compatible with a broader reasoning benefit, but this experiment isolates neither that explanation nor an algorithm-specific gain. Like the earlier positive direct contrast and failed plurality control, it shows why repair claims need strong unchanged comparators. Historical coverage, direct-baseline and cost findings concern separate estimands and remain in the report.

The supplement includes complete pilot, parser-sensitivity, development and confirmation panels, protocols, source snapshots and a default offline audit. The audit recalculates paired arithmetic, intervals, tests and gatekeeping from sufficient statistics. It checks costs and the preserved original stop. Author-side checks also verified frozen sources, allocation and all cached pre-recovery bytes. Aggregate replay neither regenerates raw traces nor reproduces row-level label joins or verifies independently administered evaluation.