COVER: Identifiable Evaluation of Structured Routing Interventions

Anonymous authors
Paper under double-blind review

Abstract

A routing system may choose a model, a team of agents, a tool bundle, or a complete model configuration. Exhaustively executing every legal action can be costly when the goal is only to compare two frozen routing policies. We introduce COVER, an identification-first framework that asks which actions must actually be executed to determine that policy difference while keeping the downstream execution process fixed. Under a declared linear action-value representation, a policy contrast is exactly identifiable when the difference between the policies' expected features lies in the span of the executed action features. This yields a minimum identifying support and an explicit conditioning measure for noise amplification. Controlled experiments reduce a 220-action coalition problem to 66 actions for exact second-order table recovery and to 8 actions for exact frozen-policy contrasts. On 43 untouched MuSiQue tasks, COVER uses 9 actions per task versus 11 matched and 15 exhaustive, with all 43 contrasts within $\epsilon = .01$; however, the content-only ablation has only three distinct table patterns, so this cohort demonstrates prospective identification under favorable natural structure rather than broad natural-table compressibility. On 42 fresh MMLU-Pro tasks over 12 hosted configurations, COVER uses 5 actions per task versus 7.24 matched and 12 exhaustive, with 38/42 contrasts within $\epsilon = .05$ and fresh token/latency reductions of 29.9%/40.5%; charging shared calibration reduces savings to 11.6% actions, 12.8% tokens, and 14.7% latency. ToolSandbox and MMLU-Pro expose the failure boundary: a single-check safeguard is conservative, while a stronger residual-correction safeguard can cost more than the assumption-free baseline. COVER therefore turns routing evaluation into an auditable question about what information is actually required for the comparison.

1 The problem: what must we execute to compare routers?

Consider a system in which a router decides which expert should handle a task. Router π may send a question to Expert 1 while router ρ sends the same question to Expert 2. To attribute any difference to routing, everything after that decision must follow the same rules. If prompts, workers, aggregation, resource limits, or the evaluator also change, the final accuracy difference is no longer a clean comparison of the routing choice. COVER therefore fixes the downstream execution process and changes only the routed action.

For task x, let $\mathcal{U}_x$ be the finite set of legal actions. An action $a \in \mathcal{U}_x$ can be a worker team, a tool bundle, an expert subset, or a model configuration. Executing one chosen action is what we call an intervention. Let G denote the fixed downstream stack: prompts, workers, communication, resource policy, finalizer, judge, and scoring rule. Executing action a under this common stack produces an observed value $y^G_{x,a}$. Figure 1 shows the comparison we want to isolate.

If every legal action is executed, the finite benchmark value is simply

$$V_x^{\mathrm{bench},G}(a) = y^G_{x,a}, \tag{1}$$

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.

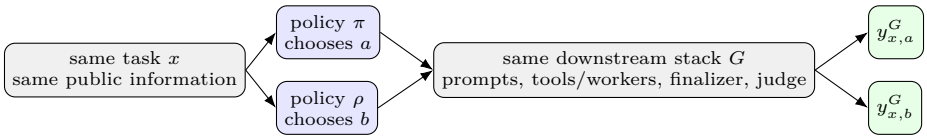


Figure 1: COVER changes the routed action while keeping the downstream execution contract fixed. Behavior that changes because a different action was selected is part of the routing effect; changing the rules of G is not.

and a deterministic router r has average benchmark regret

$$\mathcal{R}^{\mathrm{bench},G}(r) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left[\max_{a \in \mathcal{U}_x} y^G_{x,a} - y^G_{x,r(x)} \right]. \tag{2}$$

Lower regret is better. This quantity is deliberately finite and stack-conditional. If execution is stochastic, the population value $V^{\mathrm{pop},G}_x(a) = \mathbb{E}_\omega[y^G_{x,a,\omega}]$ requires repeated executions; a single provider call does not identify it.

The key distinction is between finding the best action and comparing fixed policies. Without assumptions, an unexecuted action might be the best one, so exact oracle regret generally needs complete coverage. But two frozen policies may depend on only a much smaller part of the action-value table. COVER asks:

Which actions must be executed to determine the policy difference?

Throughout the paper, we say that a comparison is identified when the observations determine it uniquely. Concretely, if two different action-value tables could both agree with everything we executed but imply different values for the policy comparison, then that comparison is not yet identified.

This paper contributes three pieces. First, we state an evaluation contract that makes the comparison auditable. Second, we give necessary-and-sufficient identification conditions, minimum support laws, and a stability criterion for structured action spaces. Third, we test both sides of the claim: cases where a small support succeeds prospectively and natural cases where the structural assumption fails and must be detected or abandoned. COVER is an evaluation methodology, not a new universal router.

2 Related work and novelty boundary

Model routing and multi-agent systems primarily ask which model, worker, tool, role, or team should be selected (Ong et al., 2024; Wu et al., 2023; Du et al., 2023; Wang et al., 2024a). Recent work also studies profile-conditioned counterfactual skill routing, large-scale retrieve-and-rerank skill selection, query-conditioned compatibility among retrieved skills, and retry-aware workflow cost (Wang et al., 2026a; Zheng et al., 2026; Wang et al., 2026b; Fu et al., 2026). These works primarily improve or analyze the routing decision itself. Set encoders and combinatorial methods similarly learn to score groups whose members may interact (Zaheer et al., 2017; Lee et al., 2019). COVER starts after that choice mechanism has been defined: the policies are frozen, and the question is how many interventions are needed to compare them under one execution stack.

Off-policy evaluation usually starts from an existing log and asks what can be inferred about another policy from actions that happened to be recorded (Wang et al., 2017; Rebello et al., 2023; Shimizu et al., 2024; Khan et al., 2024; Narang et al., 2024). For example, if an old router happened to send 70% of tasks to model A and 30% to model B, off-policy evaluation asks what that log can tell us about a new 40/60 policy. COVER instead designs which actions should be deliberately executed. The mathematical ingredients have classical precedent: feature expectations, row-space estimability, optimal design, transductive measurement, and optimal recovery all separate a target quantity from the full parameter vector (Abbeel & Ng, 2004; Searle, 1971; Pukelsheim, 1993; Fiez et al., 2019; Traub et al.,

1988; Foucart, 2021). We do not claim that "one linear functional can be estimated without reconstructing every parameter" is new. The contribution is the routing-evaluation formulation, the minimum intervention support induced by frozen policies, the coalition-specific laws below, and an empirical protocol that treats structural misspecification as falsifiable rather than silently extrapolating through it.

3 The COVER evaluation contract

Before any held-out outcome is examined, COVER fixes five objects. The legal action set $\mathcal{U}_x$ is declared; the router's information boundary excludes private evidence, answers, hidden labels, and current test outcomes; all actions share the same downstream stack G; the intervention support is chosen for the estimand—the quantity we actually want to report, such as oracle regret or the difference between two policies—and the policies, scoring rule, tolerance, and decision criterion are frozen. If several actions are executed for the same task, uncertainty is summarized at the task level rather than pretending those correlated outcomes are independent samples.

The coverage requirement depends on the question. Let $O_x \subseteq \mathcal{U}_x$ be the observed actions and $m_x^G = \max_{a \in O_x} y_{x,a}^G$. If values are known only to lie in $[L, U]$ and the router's selected action is observed, then

$$m_x^G - y_{x,r(x)}^G \le \mathcal{R}_x^{\mathrm{bench},G}(r) \le U - y_{x,r(x)}^G. \tag{3}$$

For example, with values in $[0, 1]$, a selected action worth .6 and a best observed action worth .8 imply regret in $[.2, .4]$; the unseen action prevents an exact oracle claim. Relative policy comparison can be cheaper because actions that cannot affect the contrast need not be identified.

3.1 Identifying a contrast without reconstructing the table

Assume each action has a declared feature vector $\phi_x(a) \in \mathbb{R}^q$ and that, for the structural class being tested,

$$V_x(a) = \phi_x(a)^\top \theta_x. \tag{4}$$

The features are fixed before the fresh evaluation; the task-specific coefficients θ_x are unknown. A stochastic policy simply assigns probabilities to actions—for example, .7 to A and .3 to B instead of always choosing one action. Its expected feature vector is

$$M_{\Phi,x}(\pi) = \sum_{a \in \mathcal{U}_x} \pi_x(a)\phi_x(a). \tag{5}$$

For two frozen policies π and ρ, define the feature difference

$$g_x = M_{\Phi,x}(\pi) - M_{\Phi,x}(\rho). \tag{6}$$

Their value contrast is $\Delta_x = g_x^\top \theta_x$.

The following two-agent example is the central idea in miniature. Let $\phi(\{A\}) = [1, 0]$, $\phi(\{B\}) = [0, 1]$, and $\phi(\{A, B\}) = [1, 1]$. Suppose the policies differ by $g = [.5, .5]$. Executing only the joint action is enough because $g = .5[1, 1]$. If $y_{\{A,B\}} = .8$, then $\theta_A + \theta_B = .8$ and the policy contrast is $.5(.8) = .4$. We did not need to recover θ_A, θ_B, or the other action values separately. In the general theorem, "row span" means exactly this: the target vector g must be obtainable by scaling and adding the feature rows of the actions we actually executed.

Stack the feature rows of the executed actions into $\Phi_{x,O}$. The example generalizes exactly:

Theorem 1 (Finite-action policy contrast certificate). For the linear value class in Equation 4, the contrast between π and ρ is identified from the executed actions O_x for every value function in the class if and only if

$$g_x \in \mathrm{rowspan}(\Phi_{x,O}). \tag{7}$$

Equivalently, there exists a coefficient vector λ satisfying $\Phi_{x,O}^\top \lambda = g_x$, and then

$$\Delta_x^G(\pi, \rho) = \lambda^\top y_{x,O}. \tag{8}$$

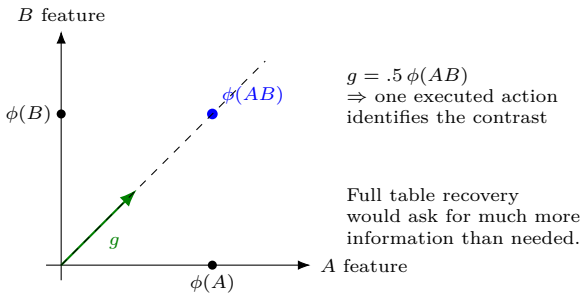


Figure 2: Identifying a policy contrast is a target-specific problem. The target direction g can lie in the span of a small executed support even when that support cannot reconstruct the full action-value table.

The vector λ is not another learned model parameter; it is simply the recipe for combining observed outcomes. If, for example, $g = 2\phi(a_1) - .5\phi(a_2)$, then $\Delta = 2y_{a_1} - .5y_{a_2}$. For a family of frozen policies Π, let $\mathcal{H}_{\Phi,x}(\Pi)$ be the span of all pairwise feature differences. The minimum common support is

$$\kappa_{\Phi,x}(\Pi) = \min\{|O| : \mathcal{H}_{\Phi,x}(\Pi) \subseteq \text{rowspan}(\Phi_{x,O})\}. \tag{9}$$

In plain language, κ is the smallest number of actions we must execute so that every policy comparison in the frozen family can be reconstructed. If all policy-difference directions can be formed from three executed feature rows, for example, then three observations are sufficient even if the legal action space contains hundreds of actions. Appendix A gives proofs, the affine-dimension law, and the general support-design hardness result.

For deterministic routers, there is an especially simple geometric interpretation. If their selected feature vectors have affine dimension r_x and the feature map contains a constant coordinate, then

Corollary 1 (Affine-dimension law). If $r_x > 0$, the minimum common support is $\kappa_{\Phi,x}(\Pi) = r_x + 1$. If $r_x = 0$, all declared policy contrasts vanish and no intervention is required.

For example, three selected actions that lie on one line have affine dimension one, so two well-chosen observations can identify every pairwise contrast among them even though three different actions are present.

3.2 Coalition structure, stability, and the assumption-free endpoint

Coalition routing gives interpretable features. A first-order model contains individual-agent indicators; a second-order model also contains pair indicators; a d-additive model includes interactions up to order d. Equation 9 remains policy-family specific. The following closed form instead asks for the universal support needed to identify all contrasts among deterministic routes on the full size-k action layer; call this family Π_{all}:

Corollary 2 (Fixed-cardinality phase law).

$$\boxed{\kappa_{\Phi,x}(\Pi_{\text{all}}) = \kappa^{\text{all}}_{d,k}(n) = \min\left\{\binom{n}{k}, \binom{n}{d}\right\}}, \qquad 1 \le d \le k, \tag{10}$$

Writing $s = \min\{k, n-k\}$, this is equivalently

$$\kappa^{\text{all}}_{d,k}(n) = \begin{cases} \binom{n}{d}, & d < s, \\ \binom{n}{k}, & s \le d \le k. \end{cases} \tag{11}$$

For a restricted frozen policy family, Equation 9 can be much smaller; if every router selects the same action, it is zero. For a small universal-family example, $n = 4, k = 2$ gives six legal pairs. A first-order model has only four independent singleton effects, so four interventions

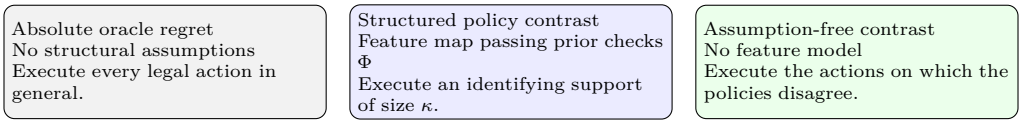


Figure 3: The required intervention support is determined by the estimand, not by the raw size of the action space alone.

suffice to cover all deterministic-route contrasts; once pair interactions are allowed, the requirement reaches all six actions. For the $n = 12, k = 3$ universal family used in the controlled study, the first-, second-, and third-order requirements are 12/66/220.

Identification does not guarantee numerical stability. If $\Phi_{x,O}^{\top}\lambda = g$, noisy action values are multiplied by the coefficients in λ. We summarize the best achievable amplification by

$$\chi_O(g) = \min_{\Phi_{x,O}^{\top}\lambda = g} \|\lambda\|_1. \tag{12}$$

Thus two seven-action supports can both identify the same contrast while one is much more fragile. For example, a certificate $\lambda = [.5, .5]$ has $\|\lambda\|_1 = 1$, whereas $\lambda = [2, -.5]$ has norm 2.5. The second recipe can amplify the same small observation errors by roughly 2.5 times as much, even though both supports are equally small and both identify the target exactly in the noiseless model. With an optimally allocated repeat budget M, the Hoeffding radius scales as $\chi_O(g)\sqrt{\log(2/\delta)/(2M)}$. This is simply a high-probability error bar: larger χ makes it wider, while more repeated executions make it shrink roughly as $1/\sqrt{M}$. Appendix A gives the robust misspecification form.

More explicitly, the optimal exact-model radius is

$$\chi_O(g)\sqrt{\frac{\log(2/\delta)}{2M}}. \tag{13}$$

Finally, structure is optional. Let $\delta_x(a) = \pi_x(a) - \rho_x(a)$. Without any cross-action model,

$$\Delta_x = \sum_a \delta_x(a) y_{x,a}. \tag{14}$$

For example, if $\pi = [.6, .3, .1]$ and $\rho = [.2, .3, .5]$ over actions A, B, C, then $\delta = [.4, 0, -.4]$: action B cancels from the comparison entirely. If A is observed at .8 while C remains only known to lie in $[0, 1]$, then $\Delta = .32 - .4y_C \in [-.08, .32]$; observing C makes the contrast exact.

Corollary 3 (Assumption-free disagreement support). If each unobserved action satisfies $y_a \in [\ell_a, u_a]$, then the exact identified interval is

$$\left[\sum_{\delta(a)\geq 0} \delta(a)\ell_a + \sum_{\delta(a)<0} \delta(a)u_a,\ \sum_{\delta(a)\geq 0} \delta(a)u_a + \sum_{\delta(a)<0} \delta(a)\ell_a\right], \tag{15}$$

with width

$$W_x(\pi, \rho) = \sum_a |\delta_x(a)|(u_a - \ell_a). \tag{16}$$

For unit-cost exact queries from $[0, 1]$, the minimax budget-B design observes the B largest $|\delta_x(a)|$ coordinates, leaving

$$W_x^{\star}(B) = \sum_{j>B} |\delta_x|_{(j)}. \tag{17}$$

This yields three distinct evaluation regimes, summarized in Figure 3.

Algorithm 1: COVER evaluation for a frozen policy pair

1. Declare the legal actions, public information boundary, downstream stack G, scalar endpoint, tolerance, and target policies.
2. Validate a candidate feature map on data that are separate from the fresh evaluation. If no map passes, use the assumption-free disagreement design.
3. From features and policy probabilities only, choose an identifying support; optionally add overidentifying check actions that should be predictable from that support.
4. Execute the frozen support under the same stack G and compute the certificate estimate $\widehat{\Delta} = \lambda^\top y_O$.
5. If a predeclared check fails, correct or fall back according to the frozen rule; otherwise report the structured estimate and its uncertainty.

Figure 4: COVER separates design from reveal. Fresh outcomes are not used to choose the policy, feature map, support, tolerance, or fallback rule.

Table 1: Core evidence. "Matched" means an assumption-free comparator targeting the same tolerance. MAE is mean absolute error; RMSE is root mean squared error. The safeguard stress test is intentionally not an efficiency win.

Setting	Reference	COVER	Main result
Controlled $d = 2$ table	220	66	64/64 tables exact; 9,856 omitted values exact
Controlled contrast-first	220	8	1,792 policy contrasts exact; full-table RMSE .490
Fresh MuSiQue, $\epsilon = .01$	11 matched / 15 full	9	43/43 within tolerance; MAE 4.23×10^{-16}
Fresh MMLU-Pro, $\epsilon = .05$	7.24 matched / 12 full	5	38/42 within; 29.9% fewer hosted tokens than matched
ToolSandbox structural audit	16 full	7	only 1/6 tables exact; omitted RMSE .347
Safeguard stress test	7.38 matched	9	42/42 numerical contrasts within tolerance; not an efficiency result

Practical workflow. Structural compression is used only after a feature hierarchy has been checked on independent or historical complete tables. Passing that calibration check is evidence, not a taskwise guarantee of validity. For coalitions, one can test singleton features, then pair interactions, then richer maps. Support selection itself uses only the frozen features and policy probabilities, not fresh outcomes. Small action spaces permit exact search; in the 15-action MuSiQue study the seven-action identifying support is verified to be minimum. Larger spaces can use rank-revealing greedy selection followed by deletion, preferring lower χ among equally small supports; we do not claim this general heuristic is support-optimal. If no compressive map survives the declared calibration criterion, COVER returns to the assumption-free endpoint. A check action is deliberately redundant: if the support predicts its value as .64 and execution returns .63, the structure looks consistent; if execution returns .20, the discrepancy is evidence that the structural model should not be trusted blindly on that task.

4 Empirical validation

The experiments answer three different questions. First, can complete intervention tables distinguish routing policies? Second, when the declared structural class is correct, do the predicted support savings actually occur? Third, what happens on natural tasks when the feature model is only approximately correct or simply wrong? Table 1 keeps those questions separate.

4.1 Complete tables and controlled mechanism tests

We first check that the evaluation table is capable of distinguishing routing policies at all. On MuSiQue-12, a public balanced-partition selector improves over the same local scores without global completion; a stronger positive control—a deliberately advantaged method used to verify that the benchmark can reveal improvement—does better still but uses latent construction labels unavailable at ordinary routing time. On HotpotQA, a permutation-invariant set scorer, whose score is unchanged when worker order is permuted, also produces substantially lower regret than the direct leave-one-out classifier. Benchmark construction and implementation details are reported in the appendix. These studies establish that the fixed intervention tables are informative; they do not yet establish that the tables can be compressed.

The first structural test makes the second-order assumption true by construction: every 220-action table is generated only from individual and pair effects. The phase law therefore predicts 66 required interventions, and 66 frozen rows recover every omitted value on all 64 tasks to numerical precision. The next experiment is deliberately harder in a different way. Eight randomized policies disagree across all 220 actions, but the directions in which those policies differ can be spanned by only eight action rows. Those eight observations recover all 1,792 pairwise policy contrasts exactly, while reconstruction of the complete table from the same data remains poor (RMSE .490). This is the central empirical distinction: we can know the policy difference exactly without knowing the whole action-value table.

We also test the stability quantity χ. Two supports use the same seven-action budget and equal per-action variance, but have very different certificate norms. The frozen theory predicts a high/low estimation-error standard-deviation ratio of 10.590; over 4,096 trials per condition, the observed ratio is 10.492. Support size and support quality are therefore separate design questions.

4.2 A natural boundary: ToolSandbox

ToolSandbox is useful precisely because the structural story does not simply succeed. A frozen 51-scenario, seven-router panel contains 816 possible task–team coordinates, but the routers collectively select only 124 of them. Those 124 coordinates are enough for the realized frozen-router comparison without any structural model, so we treat this as an assumption-free cached route-union baseline rather than structural compression. The idea is simple: if two frozen routers ever choose only teams A and B on a task, then comparing those two routers needs the cached values of A and B, not every team that neither router chose. The same public-minus-all-workers comparison evaluated under four serving stacks gives $+.180, +.123, +.113, -.046$; the sign reversal shows why the downstream stack is part of the estimand.

We then test a rank-7 structural map on a separate same-stack audit. After seven support cells per task, only 1/6 complete tables is exact, omitted-value RMSE is .347, and one true tie is estimated as $+.375$. Five of the six primary router contrasts remain exact mostly because the selected routes themselves were directly observed; only one relies on a genuinely omitted route that happened to be reconstructed correctly. This does not invalidate the identification theorem. It invalidates the feature model for those tasks, which is exactly why COVER makes structural adequacy an explicit premise.

4.3 Prospective MuSiQue: compression before reveal

The main natural coalition test uses 43 previously untouched MuSiQue source tasks: 19 geographic, 16 relational, and 8 temporal. Each task has 15 candidate actions. Before observing any fresh outcomes, we freeze the previously learned action feature map, two stochastic policies, a seven-action identifying support, two additional check actions, the tolerance $\epsilon = .01$, and the matched assumption-free comparator. The check actions are not needed to compute the contrast; they test values that should be predictable if the structural map is behaving as expected.

Each action receives a fixed graded value

$$V_{\mathrm{EV}} = r(.45q + .45c - .05d) - .03t - .02\ell, \quad r = \{1 + \exp[8(\max(t, \ell) - .8)]\}^{-1}, \tag{18}$$

where $q, c, d, t, \ell \in [0, 1]$: q is answer quality, c verified evidence coverage, d the fraction of non-required workers, and t, ℓ normalized token and latency usage. Since $r \in (0, 1)$, the declared endpoint lies in $[-.10, .90]$; this same interval is supplied to the matched assumption-free certifier. The intent is simple: reward correct, well-supported answers while mildly penalizing unnecessary workers and excessive resources. The rule and range are fixed before policy comparison.

All nine support-plus-check outcomes are completed before the remaining six actions are revealed. COVER estimates all 43 policy contrasts within .01 (MAE 4.23×10^{-16}; maximum 1.04×10^{-15}), and all 43 structural checks pass. The matched assumption-free procedure needs 11 actions per task (MAE .00416), while exhaustive evaluation needs 15. Thus fresh intervention totals are 387, 473, and 645: COVER saves 18.2% versus matched and 40.0% versus exhaustive.

Could that result be an artifact of the resource terms in Equation 18? After full reveal, we remove the reliability gate and token/latency penalties while leaving the frozen policies, feature map, seven-action supports, and certificates unchanged. The content-only value $.45q + .45c - .05d$ remains exact on all 43 tasks. Removing the redundancy term as well, leaving only $.45q + .45c$, still yields 43/43 within .01 (MAE .000244). Resource variation therefore increases numerical rank rather than manufacturing the compression. The limitation is different: the content-only tables exhibit only three distinct patterns, so this ablation does not establish broad natural-content diversity.

We also ask whether an assumption-free method could simply stop early once the decision is clear. After each new action, such a method keeps an interval of all policy differences still compatible with the unseen values. If the interval is $[.03, .15]$, it may safely stop and declare a positive difference; if it is $[-.02, .10]$, either policy could still be better and more actions are required. For strict winner sign, this sequential interval certifier averages 9.07 actions, essentially tying the nine-action operational COVER budget; moreover, an uncertified one-action midpoint guess happens to get all 43 signs right, making raw sign unusually easy here. The predeclared margin-aware state is harder: positive if $\Delta > .01$, negative if $\Delta < -.01$, neutral otherwise. The deployable assumption-free certifier then averages 11.58 actions. Even a post-reveal oracle choosing the best possible certifying subset averages 10.07 and cannot certify more than 21/43 states at a budget of nine, whereas COVER identifies all 43. The claim is therefore about an identified margin-aware comparison, not about guessing the winner.

4.4 MMLU-Pro: non-coalition actions, real hosted cost, and failure safeguards

To test the theorem outside worker teams, we use 42 untouched MMLU-Pro questions, three from each of 14 subjects (Wang et al., 2024b). Here an "action" is a complete hosted-model configuration. The 12 possibilities come from two GPT-OSS model sizes × three reasoning-effort levels × two prompting modes. For example, "120B + high reasoning + self-check" is one legal action. Two frozen stochastic policies trade predicted correctness against measured token cost, and separate calibration data choose the rank-5 action representation before any fresh test outcome is read. "Rank 5" does not mean that five configurations are universally best; it means that, on calibration data, the variation among the 12 configuration values is summarized by five independent structural directions.

On the fresh cohort, COVER executes five configurations per task (210 calls) versus 7.24 for the matched assumption-free design (304) and 12 exhaustive (504). COVER is within $\epsilon = .05$ on 38/42 tasks with MAE .0173; the matched design is 42/42 with MAE .0197. Because each intervention is a real hosted completion, the action reduction translates into cost: 178,489 versus 254,536 hosted tokens (29.9% fewer) and 486.7 versus 818.0 accumulated request-seconds (40.5% lower) relative to matched. Each task–configuration cell is executed once, so these results describe the realized finite intervention table rather than population means under repeated provider/model randomness. When the shared calibration panel is

charged equally to all methods, COVER still saves 11.6% actions, 12.8% tokens, and 14.7% latency on the first cohort.

The four misses are informative because they let us ask why the structural estimate failed. After exhaustive reveal, the errors track how badly the rank-5 representation fits the complete action table, not how large the certificate coefficients are. All four failures also contain a very large residual on the same omitted large-model/high-reasoning/direct configuration. In other words, the five-action support was not merely numerically unstable; the structural model was missing a task-specific model–reasoning–prompt interaction. The detailed correlations are reported in the appendix.

A disjoint stress cohort then tests whether this kind of structural failure can be detected prospectively. Before any fresh outcome is observed, we freeze five structural actions, four additional check actions, a correction rule, and a fallback rule. Each check action has a value predicted from the five-action structural model. If its observed value differs from that prediction, the residual tells us how much the structural model was wrong on an action that matters to the policy comparison. We use those residuals to correct the contrast:

$$\widehat{\Delta}_{\text{corr}} = \widehat{\Delta}_{\text{struct}} + \sum_{a \in C} (\pi(a) - \rho(a))(y_a - \widehat{y}_a). \tag{19}$$

For example, suppose the structural estimate is .10, one checked action has policy-probability difference $\pi(a) - \rho(a) = .3$, and its observed value is .4 higher than the structural prediction. That action contributes a missed contrast of $.3 \times .4 = .12$, so the corrected estimate becomes .22. The check is therefore not merely an alarm: its measured residual can directly repair the target comparison. Equation 19 is an empirical safeguard, not a general misspecification guarantee. Writing $r_a = y_a - \widehat{y}_a$, after correcting the checked set C, the remaining structural bias is exactly $\sum_{a \notin C} (\pi(a) - \rho(a)) r_a$. Without an assumption or bound on those unchecked residuals, a finite check set cannot certify that the remaining bias is small. The uncorrected structural estimate is outside the .05 tolerance on 11/42 tasks. All 11 harmful cases produce nonzero check residuals, and the frozen correction brings all 11 back within tolerance. Across the full cohort, the corrected estimator is therefore 42/42 within .05 (MAE .00779, maximum error .03311). Only one task falls back, and no structurally acceptable task is rejected. The cost is important: this safeguard uses 9.0 actions per task versus 7.38 for the matched assumption-free comparator, so it is evidence that misspecification can be detected and repaired prospectively, not evidence of an efficiency gain. One categorical decision still fails: the true contrast $-.0428$ lies inside the neutral band $[-.05, .05]$, but the corrected estimate $-.0565$ falls just outside it. Numerically the error is only .0137, yet it crosses the declared superiority boundary.

5 Scope and conclusion

COVER reframes routing evaluation as an intervention-identification problem. The raw number of legal actions is not, by itself, the evaluation complexity. If the goal is exact assumption-free oracle regret, an unobserved action may still be the optimum and complete coverage is generally necessary. If the goal is a comparison between frozen policies, the required information can be much smaller: an identifying support under a declared structural map that passed the prior calibration checks, or the policies' disagreement support when no cross-action structure is assumed.

The experiments separate three issues that are easy to conflate. Support complexity asks how many actions are required: the controlled studies realize $220 \to 66$ for full second-order recovery and $220 \to 8$ for contrast-only recovery, while fresh MuSiQue and MMLU-Pro show prospective reductions on natural tasks. Conditioning asks whether an identifying support is stable: the observed 10.492 noise-amplification ratio closely matches the 10.590 prediction from χ. Model adequacy asks whether the declared feature class is true enough for the new task: ToolSandbox and MMLU-Pro show that it can fail, and the safeguard studies show both the promise and cost of detecting or correcting that failure.

The present theory assumes a finite declared action family, frozen target policies, a fixed downstream protocol, and a scalar endpoint. Continuous controls, dynamically expanding

action spaces, and policies that alter the available actions after intermediate observations require additional identification theory. Results are also stack-conditional: changing prompts, models, tools, workers, finalizers, or the scoring procedure changes G and can change the policy ordering. The natural prospective studies here use one frozen policy pair per cohort and action spaces of 12 or 15 actions; broader evidence across policy distances, larger action spaces, and repeated hosted executions remains future work. Finally, passing calibration or a finite set of structural checks is evidence, not a universal proof, that a feature map remains valid on every future task.

COVER does not propose another router. It asks a different question: once the routers are fixed, what must actually be executed in order to compare them? Under valid structure, an exact policy contrast can require far less information than the complete action-value table; when that structure fails, the failure should be exposed rather than hidden behind extrapolation. The key shift is from asking "How large is the action space?" to asking "What information is actually required to identify the policy comparison?"

AI use statement

Generative AI tools were used to aid and polish writing, support retrieval and discovery of related work, draft portions of the manuscript, and assist with checking and proving mathematical claims. The authors reviewed AI-assisted text, citations, and mathematical material against the underlying sources and derivations, and take responsibility for the final manuscript and all scientific claims.

Ethics statement

This study evaluates how routing policies are compared. We use existing benchmark datasets together with automated model and tool executions; we do not recruit, observe, or experiment on human participants. The public release excludes API keys, credentials, and private provider information. Anyone reproducing the experiments must obtain the datasets and models through their normal sources and follow their original licenses and terms.

Reproducibility statement

We provide the information needed to trace each main result from design to output. The appendix contains the formal proofs, benchmark construction, the order in which policies and intervention supports were frozen, cost accounting, failure analyses, and a step-by-step reproduction checklist. The checklist identifies the frozen study records and analyses required to reproduce the central claims. Credentials and licensed benchmark content are not redistributed, so reproducing new model runs requires obtaining those datasets separately.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the Twenty-First International Conference on Machine Learning, 2004. doi: 10.1145/1015330.1015430.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325, 2023.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ba6c657b03fc7c8dd4dff8e45defcd2-Abstract.html.

Simon Foucart. Instances of computational optimal recovery: Refined approximability models. Journal of Complexity, 62:101503, 2021. doi: 10.1016/j.jco.2020.101503.

Heming Fu, Shan Lin, Qianqian Xie, and Guojun Xiong. Not all tokens are equal: Inflation-aware routing for agentic llm systems. arXiv preprint arXiv:2608.13571, 2026. URL https://arxiv.org/abs/2608.13571.

Sadegh Jokar and Marc E. Pfetsch. Exact and approximate sparse solutions of underdetermined linear equations. SIAM Journal on Scientific Computing, 31(1):23–44, 2008. doi: 10.1137/070686676.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 23734–23757, 2024. URL https://proceedings.mlr.press/v235/khan24b.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Proceedings of the 36th International Conference on Machine Learning, pp. 3744–3753, 2019.

Adhyyan Narang, Andrew Wagenmaker, Lillian J. Ratliff, and Kevin Jamieson. Sample complexity reduction via policy difference estimation in tabular reinforcement learning. In Advances in Neural Information Processing Systems, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28795419a644f41ede3fa058b13fc622-Abstract-Conference.html.

B. K. Natarajan. Sparse approximate solutions to linear systems. SIAM Journal on Computing, 24(2):227–234, 1995. doi: 10.1137/S0097539792240406.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. arXiv preprint arXiv:2406.18665, 2024.

Friedrich Pukelsheim. Optimal Design of Experiments. Wiley, 1993.

Aaman Rebello, Shengpu Tang, Jenna Wiens, and Sonali Parbhoo. Leveraging factored action spaces for off-policy evaluation. arXiv preprint arXiv:2307.07014, 2023. URL https://arxiv.org/abs/2307.07014.

Shayle R. Searle. Linear Models. Wiley, 1971.

Tatsuhiro Shimizu, Koichi Tanaka, Ren Kishimoto, Haruka Kiyohara, Masahiro Nomura, and Yuta Saito. Effective off-policy evaluation and learning in contextual combinatorial bandits. In Proceedings of the 18th ACM Conference on Recommender Systems, 2024. doi: 10.1145/3640457.3688099. URL https://arxiv.org/abs/2408.11202.

Joseph F. Traub, G. W. Wasilkowski, and Henryk Woźniakowski. Information-Based Complexity. Academic Press, 1988.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-Agents enhances large language model capabilities. arXiv preprint arXiv:2406.04692, 2024a.

Tianle Wang, Yanghe Zou, Xiang Liu, Ziyao Huang, Chenchen Fu, and Weiwei Wu. Beyond task-only matching: Personalized skill routing with counterfactual evaluation. arXiv preprint arXiv:2608.28241, 2026a. URL https://arxiv.org/abs/2608.28241.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 3589–3597, 2017. URL https://proceedings.mlr.press/v70/wang17a.html.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574, 2024b.

Zifei Wang, Wei Wen, Qiang Ji, Keyu Chen, Ruizhi Qiao, and Xing Sun. Skill is not document: Query-conditioned compatibility for llm agent skill routing. arXiv preprint arXiv:2606.03565, 2026b. URL https://arxiv.org/abs/2606.03565.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Advances in Neural Information Processing Systems, volume 30, 2017.

YanZhao Zheng, ZhenTao Zhang, Chao Ma, YuanQiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and Gang Yu. Skillrouter: Retrieve-and-rerank skill selection for llm agents at scale. arXiv preprint arXiv:2603.22455, 2026. URL https://arxiv.org/abs/2603.22455.

A Formal identification details

A.1 Why moving-pipeline comparisons are not selection comparisons

Write G for the full downstream protocol: public/private timing, worker behavior, communication, finalizer, and utility. For frozen routers r_0, r_1, the selection contrast under a common protocol is

$$\Delta_x^G(r_1, r_0) = V_x^G(r_1(x)) - V_x^G(r_0(x)).$$

Theorem 2 (Moving-pipeline non-identification). Observing $V_x^{G_0}(r_0(x))$ and $V_x^{G_1}(r_1(x))$ does not identify $\Delta_x^{G_0}(r_1, r_0)$ when G_0 and G_1 may differ, even with arbitrarily many repeated observations of those two systems.

Proof. Consider one task, teams A, B, $r_0(x) = A$, $r_1(x) = B$, and observed system outcomes $0, 1$. In one compatible world $G_0 = G_1$ and $V^{G_0}(A) = 0, V^{G_0}(B) = 1$, giving contrast one. In another, both teams have value zero under G_0 but $V^{G_1}(B) = 1$ because the downstream protocol changed, giving contrast zero. The observations are identical. □

Proposition 1 (Within-stack exactness does not imply stack invariance). Even if every team in $\mathcal{F}_x$ is observed under G_0, neither the team ranking nor the regret of a policy under a different stack G_1 is identified without an assumption or observations linking $V_x^{G_0}$ and $V_x^{G_1}$.

Proof. For two teams A, B, a complete G_0 table with $V_x^{G_0}(A) = 1$ and $V_x^{G_0}(B) = 0$ is compatible both with a G_1 table preserving that ordering and with one reversing it. Both worlds agree on every G_0 intervention. Thus exhaustive coverage identifies a table conditional on G_0, not transport to G_1. □

Panel summaries do not create invariance. For a predeclared finite panel of stacks $\mathcal{G}$, sensitivity can be summarized without collapsing the stack index. For example,

$$\overline{R}_{\mathcal{G}}(r) = |\mathcal{G}|^{-1} \sum_{G \in \mathcal{G}} \mathcal{R}^{\text{bench},G}(r), \qquad s^2_{\mathcal{G}}(r) = (|\mathcal{G}| - 1)^{-1} \sum_{G \in \mathcal{G}} \left(\mathcal{R}^{\text{bench},G}(r) - \overline{R}_{\mathcal{G}}(r)\right)^2,$$

together with pairwise policy-rank reversals across G. These quantities describe the observed panel only: neither the panel mean, its variance, nor pairwise agreement identifies transport to an unobserved downstream stack.

Theorem 3 (Exact identified set for a finite action table). Fix G. Let $O_x \subseteq \mathcal{F}_x$ be observed interventions, all values lie in $[0, 1]$, $r(x) \in O_x$, and $m_x^G = \max_{S \in O_x} y_{x,S}^G$. If $O_x = \mathcal{F}_x$, then

$$\mathcal{R}_x^{\text{bench,G}}(r) = m_x^G - y_{x,r(x)}^G.$$

If $O_x \subsetneq \mathcal{F}_x$, the identified set of $\mathcal{R}_x^{\text{bench,G}}(r)$ over all bounded finite tables compatible with the observations is exactly

$$\left[m_x^G - y_{x,r(x)}^G,\ 1 - y_{x,r(x)}^G\right].$$

Thus, with at least one unobserved team, exact nonparametric regret is point-identified if and only if $m_x^G = 1$.

Proof. When all actions are observed, the displayed equality is the definition of finite-table regret under G. Otherwise, the observed maximum supplies the lower endpoint and the range supplies the upper endpoint. Assign every unobserved team a value at most m_x^G to attain the lower endpoint, or assign one unobserved team value one to attain the upper endpoint. Both completions agree on O_x. The endpoints coincide exactly when $m_x^G = 1$. □

A.2 Structural identification, disagreement support, and interaction order

Additional structural consequences. The main text keeps only the consequences needed for the primary argument. We record the remaining exact statements here.

Theorem 4 (General COVER support design is NP-complete). Consider the single-contrast decision problem whose input is a finite family of rational feature rows $\{\phi(a) \in \mathbb{Q}^q : a \in \mathcal{U}\}$, two explicitly represented rational policies π, ρ on $\mathcal{U}$, and an integer budget B. The question is whether there exists an intervention support $O \subseteq \mathcal{U}$ with $|O| \le B$ such that

$$M_\Phi(\pi) - M_\Phi(\rho) \in \operatorname{rowspan}(\Phi_O).$$

This problem is NP-complete. The hardness holds even with one deterministic policy and one policy that is uniform on an explicitly listed subset of actions.

Proof. Membership in NP follows because a proposed support O is a polynomial-size certificate and rational Gaussian elimination verifies the span condition in polynomial time.

For hardness, reduce from the classical sparse exact linear-system decision problem: given rational $A = [a_1 \ \cdots \ a_n] \in \mathbb{Q}^{m\times n}$, nonzero $b \in \mathbb{Q}^m$, and $1 \le k \le n$, decide whether $Ax = b$ has a solution with at most k nonzero entries. This decision problem is NP-complete; sparse linear-system optimization and its approximate variants are classical (Natarajan, 1995; Jokar & Pfetsch, 2008).

Let $C \in \mathbb{Q}^{k\times(k+1)}$ be the first-difference matrix whose ith row has 1 in column i, -1 in column $i+1$, and zeros elsewhere. Thus $\ker C = \operatorname{span}\{\mathbf{1}\}$. Build a COVER instance in dimension $m + k$. For each original column create an action with feature

$$u_j = \begin{bmatrix} a_j \\ 0 \end{bmatrix}.$$

For each $\ell \in \{1, \dots, k+1\}$ create a fallback action with feature

$$f_\ell = \begin{bmatrix} b/(k+1) \\ c_\ell \end{bmatrix},$$

where c_ℓ is column ℓ of C, and add one dummy action with zero feature. Let ρ choose the dummy action deterministically and let π be uniform on the $k+1$ fallback actions. Since $C\mathbf{1} = 0$,

$$g := M_\Phi(\pi) - M_\Phi(\rho) = \frac{1}{k+1}\begin{bmatrix} b \\ 0 \end{bmatrix}.$$

Set the intervention budget to $B = k$.

If $Ax = b$ has a k-sparse solution, the corresponding original actions span $[b; 0]$ and hence span g, so the COVER instance is feasible. Conversely, suppose a support of size at most k spans g. Rescaling the representation by $k+1$ gives a representation of $[b; 0]$. Let $y \in \mathbb{Q}^{k+1}$ be the coefficients assigned to fallback columns. The lower block of the representation gives $Cy = 0$, hence $y = t\mathbf{1}$. If $t \neq 0$, all $k+1$ fallback coefficients are nonzero, so the support contains at least $k+1$ fallback actions, contradicting $|O| \le k$. Therefore $y = 0$, and the upper block expresses b using at most k original columns of A. This is exactly a k-sparse solution of $Ax = b$. The construction has polynomial encoding size, completing the reduction. □

The theorem is a complexity statement about the general finite rational-feature problem, not about the fixed-slice or deterministic special cases below. Its role is to explain why those closed forms are substantive: unrestricted structured-action support optimization already contains a classical sparse-representation problem, whereas coalition structure can collapse it to an affine-dimension or incidence-rank calculation.

Corollary 4 (Strict interaction-order hierarchy). For d-additive coalition values, take features $\phi_T(S) = \mathbf{1}\{T \subseteq S\}$ for all $|T| \le d$. Then

$$M_{\Phi,x}(\pi)_T = \Pr_{S\sim\pi_x}[T \subseteq S].$$

Two routers are value-equivalent for every d-additive coalition value if and only if all subset-inclusion probabilities through order d agree. Whenever the legal family contains the full subset cube on some d workers, the hierarchy is strict: distinct policies can agree through order $d-1$ and differ at order d. At $d=1$, only marginal worker-inclusion probabilities matter; at $d=n$, the inclusion-moment vector is an invertible zeta transform of the full policy distribution.

Proposition 2 (Robust stochastic certificate and conditioning). Let δ_x be the full action-indexed policy-difference vector and let P_O select the executed coordinates, so $P_O\mu = \mu_O$. Suppose the population mean table satisfies $\mu = \Phi_x\theta + r$ with $\|r\|_\infty \le \eta$. For any certificate λ satisfying $\Phi_{x,O}^\top \lambda = g_x = \Phi_x^\top \delta_x$, execute each $S \in O$ with $\lambda_S \neq 0$ independently m_S times, with outcomes in $[0,1]$, population mean μ_S, and sample mean $\widehat{\mu}_S$. Then with probability at least $1-\delta$,

$$|\Delta_x^G(\pi,\rho) - \lambda^\top \widehat{\mu}_O| \le \eta\,\|\delta_x - P_O^\top \lambda\|_1 + \sqrt{\frac{\log(2/\delta)}{2}\sum_{S\in O}\frac{\lambda_S^2}{m_S}}. \tag{20}$$

For a fixed certificate and continuous repeat budget $M = \sum_{S\in O} m_S$, the sampling term is minimized by

$$m_S = M\frac{|\lambda_S|}{\|\lambda\|_1} \quad \text{for } \lambda_S \neq 0, \tag{21}$$

which gives sampling radius $\|\lambda\|_1\sqrt{\log(2/\delta)/(2M)}$. In the exact structural model $\eta = 0$, defining

$$\chi_O(g_x) = \min_{\Phi_{x,O}^\top \lambda = g_x} \|\lambda\|_1$$

therefore yields the optimal radius in Equation 13. Under misspecification, Equation 20 makes the bias–variance tradeoff explicit: a certificate with smaller sampling amplification can have larger structural residual, and vice versa.

Corollary 5 (Dataset and multi-router budget frontier). For a fixed evaluation set $\mathcal{D}$, the sharp interval for the mean nonparametric contrast is obtained by averaging the taskwise endpoints of Equation 15. For a finite router family Π and unit-cost budget B, minimizing the sum of all pairwise identified-set widths is achieved by selecting the B task–team coordinates with largest

$$g_x(S) = \sum_{i<j} |\pi_{i,x}(S) - \pi_{j,x}(S)|.$$

Proof of Theorem 1. For any $f_\theta \in \mathcal{V}_{\Phi,x}$,

$$\mathbb{E}_{a\sim\pi_x} f_\theta(a) = \sum_{a\in\mathcal{U}_x} \pi_x(a)\phi_x(a)^\top\theta = M_{\Phi,x}(\pi)^\top\theta.$$

Therefore two policies have equal value for every θ if and only if their embeddings agree.

For a pair π,ρ, write $g = M_{\Phi,x}(\pi) - M_{\Phi,x}(\rho)$. Exact executed values satisfy $y_{x,O} = \Phi_{x,O}\theta$. If $g \in \mathrm{rowspan}(\Phi_{x,O})$, there is λ with $\Phi_{x,O}^\top\lambda = g$, giving $\lambda^\top y_{x,O} = g^\top\theta = \Delta_x^G(\pi,\rho)$. Conversely, if g is not in the row span, there exists $v \in \ker(\Phi_{x,O})$ with $g^\top v \neq 0$. Then θ and $\theta + tv$ agree on every executed action for every t but yield different router contrasts.

For a finite family, all pairwise embedding differences lie in $\mathcal{H}_{\Phi,x}(\Pi)$, and the differences from π_1 span it. Thus all contrasts are identified exactly when this target subspace is contained in the executed row span. Any $|O|$ rows span dimension at most $|O|$, giving $\dim\mathcal{H} \le \kappa$. A row basis of Φ_x gives $\kappa \le \mathrm{rank}(\Phi_x)$. Executing every coordinate in $D_x(\Pi)$ identifies every pairwise contrast without any structural assumption because omitted coordinates have identical probability under all policies, so $\kappa \le |D_x(\Pi)|$. □

Proof of Corollary 1. For deterministic routers, $M_{\Phi,x}(\pi_k) = \phi_x(a_k)$, so the target subspace in Theorem 1 has dimension exactly r_x. If $r_x = 0$, every selected feature vector is identical and all contrasts vanish throughout the declared value class.

Now let $r_x > 0$. Every action feature row has constant coordinate one, whereas every policy-difference vector $\phi_x(a_k) - \phi_x(a_1)$ has constant coordinate zero. Suppose m action values are queried. Their row span has dimension at most m and contains a vector with nonzero constant coordinate; therefore its intersection with the zero-constant hyperplane has dimension at most $m-1$. To contain the r_x-dimensional target subspace requires $m-1 \geq r_x$, proving $\kappa \geq r_x + 1$.

Conversely, choose $r_x + 1$ selected action feature vectors forming an affine basis of $\{\phi_x(a_k)\}$. Querying their action values makes every difference between basis points observable, and those differences span the full target subspace. Theorem 1 therefore gives $\kappa \leq r_x + 1$, proving equality. □

Proof of Corollary 2. Here Π_{all} is the family of all deterministic policies selecting one action from the full k-subset layer. Index rows by all k-subsets $S \subseteq [n]$. For d-additive values, columns are the inclusion features $\mathbf{1}\{T \subseteq S\}$ for $|T| \leq d$. On the fixed-cardinality layer, every lower-order column is in the span of the order-d columns: for $|T| = s < d$,

$$\sum_{\substack{U \supseteq T \\ |U|=d}} \mathbf{1}\{U \subseteq S\} = \binom{k-s}{d-s} \mathbf{1}\{T \subseteq S\}.$$

Hence the full $\leq d$ feature matrix has the same rank as the classical inclusion matrix whose rows are k-sets and columns are d-sets. Over the reals, Gottlieb's maximal-rank theorem gives

$$\operatorname{rank}(W_{k,d}(n)) = \min\left\{\binom{n}{k}, \binom{n}{d}\right\} =: R.$$

The constant feature belongs to this column span by the same identity with $s = 0$, so the full row set has affine dimension $R - 1$ whenever $d \geq 1$. Since Π_{all} selects every row, Corollary 1 yields $\kappa_{\Phi,x}(\Pi_{\text{all}}) = R$. A restricted frozen policy family can have smaller affine dimension and therefore smaller κ, as Equation 9 states.

Now write $s_\star = \min\{k, n-k\}$. Binomial coefficients are strictly increasing up to $\lfloor n/2 \rfloor$ and symmetric around $n/2$. Hence for $d < s_\star$, $\binom{n}{d} < \binom{n}{k}$, while for $s_\star \leq d \leq k$, $\binom{n}{d} \geq \binom{n}{k}$. This proves Equation 11. At $d = s_\star$, the order-$s_\star$ inclusion matrix has row rank $\binom{n}{k}$ and exactly $\binom{n}{k}$ columns because $\binom{n}{s_\star} = \binom{n}{k}$. It is therefore invertible. The vector of order-$s_\star$ inclusion moments of a randomized policy is the policy-mass vector multiplied by this matrix, so the map is injective and determines the full policy distribution on the k-slice. If $k \in \{0, n\}$ there is only one legal coalition and all pairwise contrasts are vacuous, giving $\kappa = 0$. □

Proof of Corollary 4. For a d-additive set function,

$$f(S) = \sum_{\substack{T \subseteq S \\ |T| \leq d}} \mu(T).$$

Taking expectation under π gives

$$\mathbb{E}_\pi f(S) = \sum_{|T| \leq d} \mu(T) \Pr_{S \sim \pi}[T \subseteq S],$$

so the policy embedding is exactly the vector of subset-inclusion probabilities through order d. Equality of these moments is sufficient for value-equivalence. It is necessary because $f_T(S) = \mathbf{1}\{T \subseteq S\}$ is itself a bounded $|T|$-additive value function.

For strictness, fix d workers U. Let π be uniform on even-cardinality subsets of U and ρ uniform on odd-cardinality subsets. For every proper $T \subsetneq U$, exactly half of its supersets have each parity, so $\Pr_\pi[T \subseteq S] = \Pr_\rho[T \subseteq S] = 2^{-|T|}$. For $T = U$, only the distribution whose parity matches $|U|$ assigns positive probability to the full set. Hence all moments through order $d - 1$ coincide while the order-d moment differs. At $d = n$, the inclusion-moment map is the zeta transform of the policy mass function and is invertible by Möbius inversion. □

Proof of Proposition 2. Write $\mu = \Phi\theta + r$ with $\|r\|_\infty \le \eta$. For a certificate $\Phi_O^\top \lambda = \Phi^\top \delta$,

$$\Delta - \lambda^\top \widehat{\mu}_O = (\delta - P_O^\top \lambda)^\top r + \lambda^\top (\mu_O - \widehat{\mu}_O).$$

Hölder's inequality bounds the first term by $\eta\|\delta - P_O^\top \lambda\|_1$; this deterministic term is sharp over the residual box by choosing each residual coordinate with the sign of $\delta - P_O^\top \lambda$. For the second term, an observation from coalition S enters with coefficient λ_S/m_S. Hoeffding's inequality therefore gives

$$\Pr\big(|\lambda^\top(\widehat{\mu}_O - \mu_O)| \ge t\big) \le 2\exp\left(-\frac{2t^2}{\sum_{S\in O} \lambda_S^2/m_S}\right),$$

which yields Equation 20 after setting the right side to δ.

For the allocation claim, Cauchy–Schwarz gives

$$\|\lambda\|_1^2 \le \left(\sum_{S\in O} \frac{\lambda_S^2}{m_S}\right)\left(\sum_{S\in O} m_S\right) = M \sum_{S\in O} \frac{\lambda_S^2}{m_S},$$

with equality exactly when $m_S \propto |\lambda_S|$ on the active certificate coordinates. Substituting Equation 21 gives radius $\|\lambda\|_1\sqrt{\log(2/\delta)/(2M)}$. Minimizing over all exact certificates gives Equation 13 and the definition of $\chi_O(g)$. □

Proof of Corollary 3. The target $\Delta = \sum_a \delta(a) y_a$ is linear and the uncertainty set is a Cartesian product of intervals, so every coordinate optimizes independently. This gives Equation 15, and every endpoint is attained. Subtracting endpoints gives Equation 16.

For a scalar known only to lie in $[L, U]$, the midpoint minimizes worst-case absolute error with radius $(U - L)/2$. An exact query from an initial $[0, 1]$ interval removes exactly $|\delta(a)|$ from contrast width. After any support O of size at most B, residual width is $\sum_{a\notin O} |\delta(a)|$, minimized by querying the B largest absolute coefficients, proving Equation 17. Every adaptive path also terminates in some support of size at most B, so none can have smaller worst-case residual width than the same top-B tail in this finite-coordinate setting. □

Proposition 3 (Finite-sample coverage–execution decomposition). Fix a pair π, ρ and observed support O_x. Suppose each $S \in O_x$ is executed independently m_S times with outcomes in $[0, 1]$, population mean μ_S, and sample mean $\widehat{\mu}_S$. Let $\widehat{c}_x = \sum_{S\in O_x} \delta(S)\widehat{\mu}_S$ and

$$r_\delta = \sqrt{\frac{\log(2/\delta)}{2} \sum_{S\in O_x} \frac{\delta(S)^2}{m_S}}.$$

Then with probability at least $1 - \delta$ the population contrast belongs to

$$\left[\widehat{c}_x - r_\delta + \sum_{\substack{S\notin O_x \\ \delta(S)<0}} \delta(S), \quad \widehat{c}_x + r_\delta + \sum_{\substack{S\notin O_x \\ \delta(S)>0}} \delta(S)\right].$$

For a fixed total repeat budget $M = \sum_{S\in O_x} m_S$ and continuous $m_S > 0$, the allocation minimizing r_δ is

$$m_S = M \frac{|\delta(S)|}{\sum_{T\in O_x} |\delta(T)|} \tag{22}$$

for coordinates with $\delta(S) \ne 0$.

Proof. Expand $\widehat{c}_x - \sum_{S\in O_x} \delta(S)\mu_S$ into the individual bounded observations. An observation from team S enters with coefficient $\delta(S)/m_S$, so Hoeffding's inequality gives

$$\Pr\left(\left|\widehat{c}_x - \sum_{S\in O_x} \delta(S)\mu_S\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{S\in O_x} \delta(S)^2/m_S}\right).$$

Setting the right side to δ gives r_δ. Combining this simultaneous bound on the observed contribution with the sharp $[0,1]$ extrema for unobserved coordinates proves the interval.

For allocation, minimize $Q = \sum_S \delta(S)^2/m_S$ subject to $\sum_S m_S = M$. Cauchy–Schwarz yields

$$\left(\sum_S |\delta(S)|\right)^2 \le \left(\sum_S \frac{\delta(S)^2}{m_S}\right)\left(\sum_S m_S\right) = QM,$$

with equality exactly when $m_S \propto |\delta(S)|$. This proves Equation 22 and the stated optimal radius. □

Theorem 5 (Interaction-order intervention complexity). Let $f : 2^{[n]} \to [0,1]$ be a coalition-value table with Möbius representation

$$f(S) = \sum_{T \subseteq S} \mu(T).$$

If $\mu(T) = 0$ for every $|T| > d$, then the values $\{f(S) : |S| \le d\}$ uniquely identify f on all 2^n coalitions. This canonical design uses

$$q_d = \sum_{j=0}^{d} \binom{n}{j}$$

interventions. Conversely, no design containing fewer than q_d scalar coalition-value observations can identify every bounded d-additive table in general.

Proof. For $|T| \le d$, Möbius inversion gives

$$\mu(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} f(U),$$

so observing all coalitions through order d identifies every nonzero coefficient and therefore every $f(S)$. For necessity, the real vector space of d-additive set functions has dimension q_d, one free coefficient for each T with $|T| \le d$. Fewer than q_d scalar evaluations define a linear map with a nontrivial nullspace. Choose a nonzero null function h, center at the constant table $f_0 \equiv 1/2$, and scale h by sufficiently small $\epsilon > 0$ so that $f_0 \pm \epsilon h$ both remain in $[0,1]$. They agree on every observed intervention but differ elsewhere, so the full table is not identified. □

For two deterministic frozen routes, define $U_x(r)$ as the observed endpoint for $r(x)$. Route-union regret uses only the two executed values, and its difference cancels the route-union oracle:

$$\big[\max(U_x(r_0), U_x(r_1)) - U_x(r_0)\big] - \big[\max(U_x(r_0), U_x(r_1)) - U_x(r_1)\big] = U_x(r_1) - U_x(r_0).$$

This is a relative contrast over frozen policies, not absolute regret against $\max_{S \in \mathcal{F}_x} y^G_{x,S}$. It is the deterministic special case of Corollary 3.

B Benchmark construction and information controls

B.1 MuSiQue-12

Each held-out MuSiQue task contributes one target multi-hop evidence component. Three non-target components are drawn from source-disjoint donor examples and combined with the target to form a twelve-card pool. Card order and superficial identifiers are randomized so that routing cannot exploit fixed positional or naming cues.

A legal action is any size-three coalition, yielding $\binom{12}{3} = 220$ candidate teams per task. The downstream evaluator is fixed across all actions. It receives the selected workers' evidence in canonical order and assigns value according to the required evidence units recovered by the

coalition. The resulting benchmark therefore provides a controlled setting with substantial combinatorial complementarity while preserving a common execution protocol across routing decisions.

The construction is intentionally synthetic at the coalition level. In particular, it does not assume that natural agent pools are organized into balanced three-worker groups; the fixed-cardinality design is used to obtain a controlled and exhaustively enumerable action space.

The train-only component-membership labels used by the pre-specified Partition-COVER positive control are treated as privileged construction supervision. These labels are available only during fitting of that diagnostic interface and are never exposed to held-out routers at decision time. Because such labels are not naturally available in deployment, results depending on them are interpreted only as a positive control rather than as evidence for a deployable routing interface.

B.2 HotpotQA-4

Each HotpotQA task exposes four public worker cards and requires the router to select exactly three workers. Private worker evidence is constructed so that successful finalization requires combining information from the selected messages. The four legal actions therefore correspond to the four possible single-worker omissions.

The small action space permits exhaustive evaluation of every legal action on held-out tasks and supports direct comparisons between set-based routing and classifiers that can enumerate all four alternatives. Accordingly, HotpotQA-4 is used primarily as a sample-efficiency and inductive-bias benchmark rather than as evidence of a representation-theoretic advantage over exhaustive action scoring.

C Validation-policy implementation details

C.1 Unified-COVER

Unified-COVER uses a frozen MiniLM sentence encoder for task and public-card embeddings. For a task embedding t and card embedding v_i, a learned projection receives

$$[t, v_i, t \odot v_i, |t - v_i|].$$

Self-attention pools card states for both the full available pool and a candidate team. Permutation-invariant sums and maxima are concatenated with these pooled states and passed to a small MLP. No worker position or hidden role label is supplied.

Training uses observed training-coalition values with regression, pairwise-ranking, and best-team listwise objectives. On HotpotQA-4, Unified-COVER scores all four legal teams directly.

C.2 Partition-COVER

Partition-COVER is a structured decoder rather than a second learned set-value model. A logistic text interface fitted on 120 training tasks maps public task/card pair features to a bounded compatibility score q_{ij}. The pre-specified positive-control interface fits this map using latent component-membership targets.

At held-out routing time, the decoder consumes only public task and card information. It applies a fixed monotone logit transform to the clipped q_{ij}, combines these scores with task relevance and path coherence, scores candidate triples, and then applies exact bitmask dynamic programming over balanced partitions.

Each matched graph receives the same interface and relevance terms. Thus, within-interface comparisons isolate the contribution of global completion conditional on shared local information rather than differences in edge prediction. The task-relevance and path-coherence weights $(8, 1)$ were selected on development data and frozen before held-out evaluation.

The decoder runs in $O(2^n n^2)$ time and $O(2^n)$ memory. It is practical for $n = 12$ but is not claimed as a hundred-worker algorithm. If natural coalitions have unequal sizes, the balanced feasible family can exclude the optimum; variable-size partitioning remains future work.

D Core experiment provenance

This appendix reports the provenance needed to verify the paper's main experimental claims. It focuses on frozen protocols, execution chronology, intervention counts, decision freezes, and released artifacts for the primary prospective studies. Exploratory variants, secondary diagnostics, and superseded protocol iterations are excluded from this section and retained in the release artifacts.

D.1 Controlled structural validation

The $d = 2$ compression study freezes 64 tasks with 12 workers and all 220 size-three coalitions. Values are generated from 66 pair-inclusion coordinates, so the declared class is exact by construction. The support is selected from feature rows before outcome generation and contains 66 coalitions/task. After 4,224 support executions, all 9,856 omitted predictions are frozen before the omitted values are revealed; every prediction is exact at tolerance 10^{-10} (omitted RMSE 2.40×10^{-16}, maximum 1.11×10^{-15}). Across 1,541,760 deterministic-router pairwise contrasts, the maximum error is 1.78×10^{-15}, with no sign or argmax mismatch. The anonymous release contains the frozen protocol and aggregate summary for this study.

The contrast-first study uses the same 220-action slice but freezes eight randomized routers and an eight-row common identifying support before any task outcomes. The routers differ on all 220 action probabilities; the target embedding-difference span has dimension seven, and the affine constraint on legal pair-feature rows makes eight rows necessary and sufficient. Across 64 hidden-seed tasks, all 1,792 frozen router contrasts are exact (MAE 7.22×10^{-17}; maximum 2.90×10^{-16}). As a secondary post-reveal diagnostic, the minimum-norm full-table completion from those same eight observed values has omitted-cell RMSE .490. Thus the policy contrasts are identified exactly even though the full action-value table is not accurately reconstructed. The frozen support and aggregate results are included in the anonymous release.

D.2 Natural ToolSandbox boundary

For the nonparametric natural panel, the frozen cohort contains 51 scenarios in 34 family clusters, seven routers, 816 possible task–team coordinates, and a 124-coordinate selected-route union. The union is frozen before outcomes and exactly supports all selected-route means and pairwise contrasts. Because this support is simply the cached union of selected routes, it is a nonparametric evaluation baseline rather than structural compression.

Across the four serving stacks, the public-minus-all-workers safe-evidence differences are $+.180$, $+.123$, $+.113$, and $-.046$. The sign reversal on GLM-4.7 provides direct evidence that the estimand is stack-conditional. Infrastructure retries and invalid tool calls remain in the released execution logs.

D.3 Fresh direct-Groq MuSiQue contrast study

The prospective protocol is frozen before any outcomes are observed and contains 43 untouched MuSiQue sources: 19 geographic, 16 relational, and 8 temporal. The cohort has zero source overlap with the earlier 18-task prospective cohort, and the two tasks that received any partial execution in a stopped predecessor are excluded entirely. The direct serving stack uses GPT-OSS-120B through Groq, with no intermediate model-serving layer.

The rank-8 feature map and stochastic policy pair are inherited unchanged from the earlier frozen design. For each task, support selection uses only the frozen action features and policy probabilities. It identifies a minimum seven-action support whose row span contains

the target policy-difference embedding, then adds two overidentifying actions whose feature rows are predictable from that support and have large $|\pi - \rho|$. Phase 1 therefore contains exactly nine actions/task. The matched assumption-free comparator is frozen at the same $\epsilon = .01$ and requires eleven actions/task; exhaustive evaluation has fifteen.

The execution chronology is fixed and auditable. All 387 phase-1 rows are completed first. The 43 structural estimates and accept/fallback decisions are then frozen before any remaining outcomes are revealed. Only afterward are the remaining 258 rows executed. All 43 tasks accept the structural estimate. COVER contrast MAE is 4.23×10^{-16} and maximum error is 1.04×10^{-15}; all 43 are within .01, with zero false superiority claims, zero missed check failures, and zero fallbacks. The matched assumption-free comparator is also 43/43 within .01 with MAE .00416. Fresh intervention totals are 387/473/645 for COVER, matched assumption-free, and exhaustive evaluation, corresponding to reductions of 18.2% and 40.0%, respectively. Including the shared 1,200 historical calibration actions gives totals 1,587/1,673/1,845, or 5.14% and 13.98% reductions. Because the proof-short-circuited runtime gives all three counterfactual paths the same 43 hosted rows, 21,922 tokens, and 58,912 ms, this study provides evidence for intervention-count savings rather than hosted-compute savings. The release includes the frozen pre-reveal decisions needed to audit this chronology.

Post-reveal utility ablation. To test whether the apparent structure is created by token or latency terms, the fully revealed tables are rescored while keeping the frozen policies, rank-8 feature map, seven-action supports, and contrast certificates unchanged. Removing the reliability gate and all token/latency terms gives the content endpoint $.45q + .45c - .05d$; the same certificates remain exact on all 43 tasks (MAE 4.22×10^{-16}). Removing the redundancy term as well, leaving $.45q + .45c$, still yields 43/43 contrasts within .01 (MAE .000244, maximum .000417). The full composite has raw rank 6 whereas the content-only table has raw rank 3, so the resource terms add algebraic variation rather than creating the low-rank content pattern. The content-only tables exhibit only three exact patterns, which remains a breadth limitation.

Adaptive assumption-free decision audit. Using the already revealed complete tables, an assumption-free sequential baseline queries actions in descending $|\pi(a) - \rho(a)|$ and stops only when the exact remaining interval certifies the decision. For strict sign it stops after six actions on 21 tasks and twelve on 22, for a mean of 9.07. For the predeclared $\epsilon = .01$ three-state decision it stops after eight actions on 21 tasks and all fifteen on 22, for a mean of 11.58. A post-reveal oracle that may choose the best certifying subset for each task requires seven actions on 21 tasks and thirteen on 22, for a mean of 10.07. Thus at a nine-action budget even this oracle can certify only 21/43 margin-aware states, whereas the frozen structural design identifies all 43. This audit is retrospective and is used only to characterize the decision-budget comparison; it does not become a new prospective experiment.

D.4 Prospective MMLU-Pro structured-configuration study

To test the finite-action theorem outside coalitions, the prospective configuration study uses 42 untouched MMLU-Pro test questions, with three questions from each of 14 categories (Wang et al., 2024b). The legal action set is a 12-cell Cartesian grid: GPT-OSS-120B or GPT-OSS-20B, low/medium/high reasoning effort, and direct/self-check prompting. Every action is one direct Groq completion, so intervention count, hosted tokens, and recorded request latency are all observable marginal costs.

A separate validation set is executed before the fresh test cohort. Public question text is used to train per-action ridge correctness models, which are used only to define two frozen stochastic policies: a quality policy and a token-cost-aware policy. A calibration-SVD feature hierarchy is evaluated on 14 held-out validation questions. The frozen rule selects the smallest rank with at least 90% of validation contrasts within $\epsilon = .05$, validation MAE at most .025, and mean structural support below the matched assumption-free support. Rank 5 is selected: 13/14 validation contrasts are within tolerance, MAE is .0158, and maximum

validation error is .0828. The fact that the maximum already exceeds ϵ is retained: the design intentionally tests approximate rather than exact structural transport.

The fresh protocol is frozen before test labels or outcomes are materialized. COVER then executes exactly five identifying actions/task (210 rows total), after which all 42 estimates are frozen before any exhaustive reveal. The matched assumption-free design is completed next and its decisions are also frozen. Only then are the remaining 138 cells executed, producing all 504 task–action outcomes. The anonymous release contains the corresponding pre-reveal commitments.

Prospectively, COVER uses 5.00 actions/task versus 7.24 matched assumption-free and 12 exhaustive. It achieves MAE .0173, maximum error .1406, and 38/42 contrasts within .05; the matched comparator is 42/42 within tolerance with MAE .0197 and maximum error .0493. COVER records 178,489 hosted tokens and 486,699 ms, versus 254,536 tokens and 817,958 ms for the matched design and 425,366 tokens and 1,336,463 ms exhaustive. Thus COVER uses 30.9% fewer interventions, 29.9% fewer tokens, and 40.5% lower accumulated request latency than the matched comparator. The four structural misses are reported unchanged rather than removed by post-hoc tuning.

Calibration-inclusive accounting. The policies and structural geometry reuse the completed calibration panel: 28 fit and 14 validation questions, all 12 actions each, for 504 hosted calls, 340,825 tokens, and 1,431,670 ms. Because the matched comparator evaluates the same learned policy pair, this calibration cost is shared by both methods. Charging this calibration cost equally gives first-cohort totals of 714/808/1008 actions, 519,314/595,361/766,191 tokens, and 1,918,369/2,249,628/2,768,133 ms. COVER therefore still saves 11.63% actions, 12.77% tokens, and 14.73% latency versus the matched comparator. With repeated reuse, these savings approach the fresh-only 30.9%/29.9%/40.5% rates. As a conservative sensitivity analysis, if the entire calibration panel is charged only to COVER while the comparator is assumed to receive an already-frozen policy pair at no calibration cost, break-even occurs after approximately 225 tasks by action count, 188 by tokens, and 182 by latency.

Prospective single-check safeguard. A disjoint confirmation cohort freezes 42 new test questions with zero overlap with the earlier fresh cohorts, together with the inherited rank-5 map and policy pair, a five-action structural support, one reconstructible check, the matched nonparametric support, and accept/fallback rules. All decisions are frozen before exhaustive reveal. Exhaustive evaluation later identifies seven tasks with structural contrast error above .05; the frozen check rejects all 7/7. It also rejects 26/35 structurally acceptable tasks, so the screen is conservative: nine tasks accept structure and 33 fall back. The final estimator is 42/42 within .05 (MAE .01534, max .04506) at 7.05 actions/task versus 7.36 for the matched comparator, saving 4.21% actions, 6.79% tokens, and 5.77% latency. One false-superiority claim remains, equal to the matched comparator's count, so the predeclared zero-false-superiority criterion is not satisfied. This therefore provides prospective evidence that one frozen check is sensitive to harmful structural failures, with an explicit efficiency cost, not a claim that a single check universally certifies structure.

Prospective multi-check stress safeguard. To evaluate the safeguard under a higher incidence of harmful misspecification while reducing the single-check specificity cost, a disjoint 42-task stress cohort is frozen before any fresh outcome is observed: 10 unseen questions each from biology, chemistry, and psychology, plus six each from computer science and physics as negative controls. The inherited rank-5 policies/geometry, five-action identifying support, four reconstructible checks with largest $|\pi - \rho|$, correction formula, development-envelope fallback, and success criteria are all fixed before execution. For a check action a, the observed structural residual $r_a = y_a - \widehat{y}_a$ contributes exactly $\delta_a r_a$ to the policy contrast, so the corrected estimator is

$$\widehat{\Delta}_{\text{corr}} = \widehat{\Delta}_{\text{struct}} + \sum_{a \in C} \delta_a r_a.$$

Exhaustive reveal later shows 11/42 raw structural errors above .05 (31/42 within tolerance; MAE .03592, max .16629). All 11 harmful cases produce nonzero residuals on the frozen checks, and the four-check correction brings all 11 back within tolerance. Across all 42 tasks,

the corrected estimator is 42/42 within .05, with MAE .00779 and maximum error .03311. The stress stratum contributes 9/30 failures and the control stratum 2/12. The frozen development-envelope rule triggers one fallback and no fallbacks on structurally acceptable tasks. Operationally, the result remains 42/42 within tolerance (MAE .00876, max .04531), but it uses 9.0 actions/task versus 7.38 matched assumption-free and therefore serves as a safeguard stress test rather than a compression result. The predeclared overall success criterion nevertheless remains failed because one corrected estimate classifies a true contrast of $-.0428$ as below the $-.05$ superiority boundary (estimate $-.0565$), despite an absolute contrast error of only .0137; the matched comparator has zero such boundary crossings.

D.5 Earlier 18-task MuSiQue matched-tolerance study

The earlier 18-task protocol is frozen before any fresh outcome is observed. Eighteen previously unused MuSiQue source IDs are selected outcome-blind: six geographic, six relational, and six temporal. The anonymous release contains the corresponding frozen cohort and pre-reveal commitment.

The value is Equation 18. Let R be the required evidence roles and S the selected team. A role is recovered when its expected citation or opaque verification fragment appears in that role's emitted message; legacy masked traces also recognize the role-specific verification marker. Thus $c = |R_{\text{recovered}}|/|R|$ when $R \neq \emptyset$ (otherwise $c = q$), and $d = |S \setminus R|/\max(1, |S|)$. Resource fractions are $t = \min(1, \text{token_cost}/4000)$ and $\ell = \min(1, \text{latency_ms}/60000)$; the budgets are protocol constants.

The historical 80-task training table fixes a rank-8 right-singular action space and an eight-action identifying support. The target stochastic policy pair is selected from the pre-existing ridge/DyLAN/static policy grid using only frozen probability divergence on public task text. Before execution, the protocol also freezes $\epsilon = .01$, two sentinels/task chosen by largest omitted $|\pi - \rho|$, the matched assumption-free comparator, and success criteria.

Execution is staged. The 180 support+sentinel rows are completed first; all 18 estimates and accept/fallback decisions are then frozen before any remaining outcomes are revealed. Only afterward are the remaining 90 rows executed. All 18 tasks pass the sentinel screen. COVER uses 180 fresh interventions versus 198 for the matched-tolerance assumption-free comparator and 270 for exact cached evaluation. Contrast MAE is 1.13×10^{-17}, maximum error 3.99×10^{-17}, with 18/18 within tolerance, zero false superiority claims, and zero missed sentinel failures. Including the shared 1,200-intervention calibration phase gives totals 1,380/1,398/1,470. Hosted finalizer calls, tokens, and recorded finalizer latency are identical across the three counterfactual paths because incomplete proofs short-circuit before synthesis.

The anonymous release contains the aggregate accuracy and cost summaries for this cohort.

D.6 Hosted-cost translation on genuine heldout executions

The train-learned four-cell supports were fixed before heldout scores were read. On the 36 HotpotQA heldout tasks, those four cells reconstruct every 15-action score table exactly. The complete heldout table contains 540 hosted interventions and records 422,591 tokens and 3,940,008 ms; the frozen support contains 144 interventions, 157,852 tokens, and 1,444,304 ms. Thus support reduction is 73.3% in interventions, 62.6% in recorded tokens, and 63.3% in recorded latency. The MuSiQue counterpart does not reduce hosted cost because its proof gate short-circuits all but the same 40 hosted rows, showing that intervention reduction does not necessarily translate into hosted-compute savings under every downstream execution stack. The corresponding aggregate cost summary is included in the anonymous release.

D.7 Conditioning validation

A natural equal-support repeat study compares two seven-cell certificates with $\chi = 1.7393$ and 18.4191 using 140 executions per arm. The empirical-SE and bootstrap-width ratios are 7.31 and 6.02, but coalition variances differ across supports. The controlled homoskedastic

study reuses the frozen supports and certificates, with every coalition variance fixed at .01. With 4,096 independent trials per arm, the frozen integer-allocation formula predicts a high/low error-SD ratio of 10.5901, and the observed ratio is 10.4924 (0.92% relative deviation). The frozen supports, certificates, and aggregate simulation results are included in the anonymous release.

E Reproduction checklist

The anonymous release contains the frozen study specifications, route files, per-task outputs, aggregate summaries, pre-reveal commitments, and analysis code needed for the central claims. A compact reproduction path is:

1. run the released identification tests to verify the sharp bounds, affine-dimension law, fixed-cardinality saturation, interaction hierarchy, and certificate identities;
2. verify the two controlled-study specifications, then reproduce the 66-of-220 exact table recovery and 8-of-220 exact contrast results from their frozen summaries;
3. verify the 43-task MuSiQue specification, its 387-row support-plus-check phase, the frozen pre-reveal decisions, and the 258-row reveal; then recompute the 9/11/15 intervention totals and the adaptive assumption-free stopping analysis;
4. verify the initial MMLU-Pro configuration specification and 210-row structural phase, reproduce the 5/7.24/12 intervention totals and hosted token/latency accounting, reproduce the post-reveal four-failure diagnosis, and verify both disjoint safeguard cohorts and their pre-reveal decisions;
5. verify the frozen 51-scenario ToolSandbox route specification and reproduce the 124-of-816 cached route-union summaries, followed by the rank-7 six-task omitted-cell and contrast-error audit;
6. reproduce the controlled conditioning experiment and its predicted/observed SD ratios of 10.590/10.492; and
7. compile the manuscript with the included portable Tectonic configuration.

Source-bearing questions, passages, model transcripts, provider metadata, and credentials are excluded from the public archive; licensed source datasets must be reconstructed locally for new routing runs.