UNCERTAINTY IS NOT ACTIONABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Two questions can carry the same perfectly calibrated error probability yet favor opposite interventions: reflection repairs one, whereas retrieval repairs the other. Uncertainty omits the oriented state–action effect that determines whether an intervention crosses a decision boundary. We prove that this information gap forces tight worst-case regret $\Delta(1 - 1/K)$ for every action-blind policy over K interventions, even under perfect calibration and randomized choice. Under decision-preserving transformations, we then characterize which effects are identifiable and derive minimal-dimensional coordinates for the resulting maximal invariants: anchored margins for answer scores and anchored cycles for assignment scores. When these coordinates are hidden, black-box logs reveal only the repair, harm, and cost of the executed action. We therefore introduce the Factorized Action-Value Estimator (FAVE), which estimates conditional action value by sharing need, response, and direction across actions. Across 30,918 natural matching cycles, conditional repair rises from 16.8% in the positive-but-insufficient region to 72.7% after boundary crossing. On a 2,000-item, five-intervention portfolio, FAVE exceeds equal-information pooled prediction by .0096 (two-way 95% CI $[.0030, .0163]$) and by .0203 on held-out intervention types; its advantage persists across all predeclared token prices. At an eightfold harm price, validation-selected pooling attains .0768 utility while every fixed intervention has nonpositive population value. Uncertainty locates the incumbent. Actionability predicts where each intervention will move it and what that movement is worth.

1 INTRODUCTION

Consider a worked derivation with a sign error and a factual question whose answer lies outside the model's parametric knowledge. Suppose the system assigns the two incumbent answers the same calibrated error probability. A second reasoning pass can repair the sign but cannot supply the missing fact. Retrieval can supply the fact but may distract the otherwise sound derivation. A gate that observes only error probability must treat the items alike, even though the two interventions rank in opposite orders. Calibration locates incumbent risk; it does not orient the effect of an action. Need asks whether the incumbent should change. Value asks what a particular intervention will change and whether that change is beneficial.

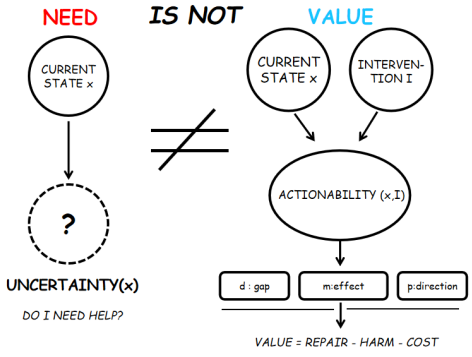


Figure 1: **Need is not value.** Equal incumbent uncertainty can coexist with opposite intervention orderings.

Choosing an intervention therefore requires two quantities: how far the incumbent is from changing, and how a particular action will move it. The second quantity has a simple geometry. In a binary slice, let d denote the target's margin deficit against its current competitor, and write the intervention effect as orientation $Z_I \in \{-1, +1\}$ times magnitude $m_I \geq 0$. The post-intervention target margin is

$$\phi_I = -d + Z_I m_I, \qquad \text{repair} \iff \phi_I > 0.$$

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.

The red coordinate belongs to the incumbent; the blue coordinate belongs to the state–action pair. Their relation, rather than either coordinate alone, determines whether the intervention crosses the boundary.

Existing calibration, selective-prediction, retrieval-gating, routing, and test-time-compute methods primarily use properties of the incumbent or query to determine whether further computation is warranted (Guo et al., 2017; Geifman & El-Yaniv, 2017; Moskvoretskii et al., 2025; Chuang et al., 2025; Chen et al., 2026; Jeong et al., 2024; Asai et al., 2024; Yao et al., 2025; Snell et al., 2025; Damani et al., 2025; Zuo & Zhu, 2026). Action-conditioned routers instead learn preferences among decision makers or model–budget configurations from preference data, contextual-bandit feedback, or difficulty–ability structure (Madras et al., 2018; Mozannar & Sontag, 2020; Ong et al., 2025; Panda et al., 2025; Fernandez et al., 2026; Xue et al., 2026); universal routing represents new models through predictions on probe prompts (Jitkrittum et al., 2026). A heterogeneous menu of retrieval, reflection, verification, tree search, and model escalation requires a joint ordering of interventions whose repair, harm, and cost profiles differ on the same item. A one-action-per-item log hides every unchosen consequence; a new intervention type has no outcome history at all.

This raises an identifiability question: which intervention effects can each observation interface recover? Complete answer scores expose boundary-relative displacement through margins. Structured assignment scores recover it through cycle contrasts after row–column nuisance is removed. A black-box log exposes neither representation; it retains only the consequence of one chosen action. The decision object remains the same across these interfaces, but its observable form changes from exact contrast to conditional action value. This progression leads to FAVE for the black-box regime. FAVE factors that value into incumbent need, output response, directional reliability, and cost, then uses logged validation to determine which components should share evidence across actions.

Contributions.

- A boundary-relative geometry separates incumbent deficit from oriented action displacement. It yields a universal crossing law and a tight $\Delta(1 - 1/K)$ minimax obstruction for every action-blind policy.
- A maximal invariant completely characterizes the intervention effects identifiable under decision-preserving transformations. Anchored margins and assignment cycles provide minimal-dimensional coordinate representations of these maximal invariants in finite-choice and assignment interfaces.
- In the black-box regime, FAVE estimates conditional action value when exact effect coordinates are hidden and only one action consequence is logged. It improves on pooled prediction under sparse feedback and held-out intervention types, while a decision-weighted defect bound and logged validation characterize when cross-action sharing is beneficial.

2 Decision Geometry and the Limits of Uncertainty

2.1 Local boundary geometry

Suppose a finite decision uses score vectors $F^0, F^I \in \mathbb{R}^K$ before and after intervention, and y is the audit target. For every competitor $a \neq y$, define the incumbent deficit and signed action effect

$$d_{y,a} = F^0_a - F^0_y, \qquad s_{I;y,a} = (F^I_y - F^I_a) - (F^0_y - F^0_a). \tag{1}$$

Theorem 1 (Universal boundary geometry)**.** *Away from ties, the post-intervention decision equals y if and only if*

$$s_{I;y,a} > d_{y,a} \qquad \textit{for every } a \neq y. \tag{2}$$

In the binary case, $s_I = Z_I m_I$ and $\phi_I = -d + s_I$, so repair occurs exactly when $s_I > d$. The deficit locates the incumbent relative to the decision boundary; the signed effect describes the displacement caused by a particular intervention. Incumbent statistics may estimate the first coordinate, but they do not determine the second. Actionability resides in their comparison.

2.2 ACTION-BLIND MINIMAX OBSTRUCTION

Call a policy action-blind when its observation remains unchanged across environments that differ only in the orientation of candidate interventions. Such a policy may observe arbitrary incumbent information, including perfectly calibrated uncertainty, and may randomize over K interventions and abstention. Write $V_e(I)$ for the net value of action I in environment e.

Theorem 2 (K-action minimax obstruction)**.** *For every $K \geq 2$, $\Delta > 0$, and action-blind policy π, there is a family of environments with identical observed information and calibrated incumbent risk such that*

$$\sup_e \mathbb{E}_e[V_e(I_e^\star) - V_e(I_\pi)] \geq \Delta \left(1 - \frac{1}{K}\right). \tag{3}$$

The bound is tight; every deterministic policy has worst-case regret at least Δ. If nonoptimal interventions are harmful, a policy with average intervention coverage ρ incurs harm probability at least $\rho(1 - 1/K)$ in some environment.

The construction holds every incumbent observation fixed and assigns value Δ to a different intervention in each environment. Some action receives probability at most $1/K$ under any randomized policy; the environment in which that action alone is optimal realizes the bound. Uniform randomization attains it. Perfect calibration refines the shared incumbent state but leaves the action permutation unresolved.

3 EXACT AND STRUCTURED OBSERVATION REGIMES

Many numerical representations induce the same decision. A decision-identifiable effect must remain constant along the corresponding nuisance orbits while retaining every contrast capable of altering an action ordering.

3.1 MAXIMAL-INVARIANT CHARACTERIZATION

Let a declared decision-preserving group G act on a representation space $\mathcal{X}$. Observation identifies the orbit $[x]_G$. A statistic $T : \mathcal{X} \to \mathcal{T}$ is maximal invariant if $T(x) = T(x')$ exactly when x' lies in the orbit of x. A maximal invariant induces the classical orbit factorization of invariant decision theory (Eaton, 1989).

Theorem 3 (Complete identifiable-effect characterization)**.** *A pre/post functional $E(x_0, x_1)$ is identifiable under independent $G \times G$ reparameterizations if and only if a unique $\bar{E}$ on the range of $T \times T$ satisfies*

$$E(x_0, x_1) = \bar{E}(T(x_0), T(x_1)). \tag{4}$$

Consequently $(T(x_0), T(x_1))$ is complete, and any strict coarsening loses some identifiable intervention effect.

State-only observation merges environments with different action orderings into one equivalence class. A maximal invariant retains exactly the information by which a richer interface separates those orderings.

We use independent pre- and post-intervention gauges; a shared gauge would instead induce a diagonal action and a different quotient.

3.2 ANSWER SCORES: ANCHORED MARGINS

For $F \in \mathbb{R}^K$, global translation $F \mapsto F + c\mathbf{1}$ preserves every answer comparison. For anchor y,

$$M_y(F) = (F_y - F_a)_{a \neq y}. \tag{5}$$

These margins separate translation orbits and are therefore maximal. Their pre/post values characterize every translation-identifiable intervention effect. For each competitor a,

$$M_y(F^0)_a = -d_{y,a}, \qquad M_y(F^I)_a - M_y(F^0)_a = s_{I;y,a}.$$

Table 1: **Exact observation regimes and maximal invariants.**

Regime	Nuisance group	Maximal invariant	Exact audit
Finite-choice scores	Global translation	$K-1$ anchored margins	2,000 items; 0 boundary failures
Assignment scores	Diagonal row/column scaling	$(n-1)^2$ anchored cycles	878 targets; 0 ranking/boundary failures

3.3 Assignment scores: anchored cycles

For a positive square assignment plan P, diagonal row–column scaling $P \mapsto \operatorname{diag}(u)P\operatorname{diag}(v)$ preserves every permutation-score ordering. For $i, j \geq 2$, define

$$C_{ij}(P) = \log \frac{P_{ij}P_{11}}{P_{i1}P_{1j}}. \tag{6}$$

Theorem 4 (Assignment-cycle maximality)**.** *The $(n-1)^2$ anchored cycles $C(P)$ form a maximal invariant of positive $n \times n$ matrices under diagonal row–column scaling. Every alternating cycle circulation and assignment-score contrast is linear in $C(P)$; hence $(C(P), C(P^+) - C(P))$ characterizes every scaling-identifiable effect.*

Sinkhorn normalization adds row–column potentials in log space, which $C(P)$ cancels (Sinkhorn, 1964; Cuturi, 2013; Shen et al., 2024). Although their nuisance groups differ, margins and cycles play the same decision-geometric role: the incumbent coordinate locates the boundary, and its pre/post difference records intervention-induced displacement.

From exact effects to black-box value. Let $Z^0, Z^I \in \mathcal{X}$ denote the pre/post representations and let $U_I = rR_I - hH_I - \kappa(I)$ be their realized repair–harm–cost consequence. Because repair and harm depend only on the induced decisions, U_I is $G \times G$ invariant. Theorem 3 therefore gives a unique $\bar{U}_I$ such that

$$U_I = \bar{U}_I\big(T(Z^0), T(Z^I)\big), \qquad V(I \mid x) = \mathbb{E}[U_I \mid X = x]. \tag{7}$$

Margins and cycles expose the arguments of $\bar{U}_I$ in the two exact interfaces. A black-box log hides those coordinates and records one realization of U_I; FAVE estimates the conditional mean in equation 7.

4 FAVE: A Constructive Black-Box Instance

4.1 Repair–harm–cost value

When invariant coordinates are hidden, a boundary crossing remains observable through its consequence. Let $\mathcal{I}(x)$ contain candidate interventions and abstention $I_\varnothing$, with $V(I_\varnothing \mid x) = 0$. Let $R_I, H_I \in \{0, 1\}$ denote repair of an incorrect incumbent and harm to a correct incumbent. For repair reward r, harm penalty h, and intervention cost κ, define

$$V(I \mid x) = r \Pr(R_I = 1 \mid x, I) - h \Pr(H_I = 1 \mid x, I) - \kappa(I). \tag{8}$$

The item-level need $q(x) = \Pr(\text{incumbent is wrong} \mid x)$ is common to the action menu. Intervention rankings additionally depend on conditional repair, harm, and cost.

4.2 Black-box projection of the geometry

Let $N = 1$ indicate incumbent error and $C = 1$ that action I produces an output different from the incumbent. Define

$$a_-(x, I) = \Pr(C = 1 \mid N = 1, x, I), \qquad p_-(x, I) = \Pr(R_I = 1 \mid C = 1, N = 1, x, I),$$
$$a_+(x, I) = \Pr(C = 1 \mid N = 0, x, I), \qquad p_+(x, I) = \Pr(H_I = 1 \mid C = 1, N = 0, x, I).$$

The boundary event now appears through three population coordinates. The need $q(x) = \Pr(N = 1 \mid x)$ places incumbent mass on the wrong side; $a_\pm$ measure whether an action changes the output on either side; $p_\pm$ measure whether that change repairs or harms. Their chain-rule value is

$$V(I \mid x) = rqa_-p_- - h(1-q)a_+p_+ - \kappa(I). \tag{9}$$

Equation 9 is the conditional mean in equation 7 after the latent deficit and displacement are integrated out. Appendix A gives the explicit local-to-population map.

Logged supervision. One action log observes $Y_I = rR_I - hH_I - \kappa(I)$ only for the executed action. Uniform randomized logging makes its action-conditional mean equal to $V(I \mid x)$ and supplies balanced overlap. Appendix B states the standard identification conditions and proof.

4.3 Sharing, approximation, and adaptation

The logged conditional mean identifies the population target, but each item still reveals no outcome for its unchosen actions. FAVE shares statistical strength across those missing outcomes by imposing

$$a_- = a_+ = a(x, I), \qquad p_- = p(x, I), \qquad p_+ = 1 - p(x, I), \tag{10}$$

and estimates

$$\widehat{V}_{\mathrm{FAVE}}(I \mid x) = r\widehat{q}\,\widehat{a}\,\widehat{p} - h(1-\widehat{q})\widehat{a}(1-\widehat{p}) - \widehat{\kappa}(I). \tag{11}$$

The first equality pools output response across incorrect and correct incumbents. The second pools directional reliability: conditional on an output change, FAVE parameterizes the repair and harm products with one directional score and its complement. Together they define a partial-pooling surrogate for the unrestricted mechanisms.

For any shared population surrogate a, p, let $V^{\mathrm{sh}}_{a,p} = rqap - h(1-q)a(1-p) - \kappa$ and define its decision defect

$$\mathcal{D}_{a,p} = rq|a_-p_- - ap| + h(1-q)|a_+p_+ - a(1-p)|. \tag{12}$$

Theorem 5 (Approximate sharing controls policy regret)**.** *Let* $\epsilon_{a,p}(x, I) = |\widehat{V}_{\mathrm{FAVE}}(I \mid x) - V^{\mathrm{sh}}_{a,p}(I \mid x)|$. *If* $I^\star$ *and* $\widehat{I}$ *maximize* V *and* $\widehat{V}_{\mathrm{FAVE}}$*, including abstention, then*

$$V(I^\star \mid x) - V(\widehat{I} \mid x) \le 2 \sup_I \{\mathcal{D}_{a,p}(x, I) + \epsilon_{a,p}(x, I)\}. \tag{13}$$

The approximation terms receive weights rq and $h(1-q)$; the harm price therefore changes which departures matter. A factorwise estimation bound and proof appear in Appendix B.

The degree of sharing can itself be selected from logged data. Separate response and direction coefficients interpolate between shared FAVE and the split-head mechanism estimator. A disjoint inner log selects the pair by inverse-propensity-weighted value loss. The construction and its finite-grid risk bound appear in Appendices A and B.

A frozen action representation can serve as a covariate in $\widehat{a}$, $\widehat{p}$, and $\widehat{\kappa}$, while item features determine $\widehat{q}$. The fitted heads can then score a represented action type with no outcome history.

The pooled baseline predicts repair and harm directly and removes the mechanism decomposition. The split-head ablation retains the population factorization but estimates (q, a_-, p_-, a_+, p_+) separately; its neural implementation shares the feature trunk but not the response or direction heads. Within each audit, the estimators use the same inputs, logged actions, folds, backbone width, optimizer, and training schedule.

5 Experiments

5.1 Empirical boundary geometry

The invariant coordinates preserve every audited decision in both exact interfaces. On 2,000 finite-choice items, anchored margins reproduce every pre/post decision and every competitor-wise crossing predicate. Across 878 assignment targets of sizes 2–10, anchored cycles preserve every pre/post

ranking under Sinkhorn normalization and reconstruct assignment-score contrasts with maximum absolute error 1.11×10^{-16}. Neither audit records a ranking or boundary failure.

Anchored-cycle coordinates recover the incumbent deficit and signed effect for 30,918 initially incorrect cycles from natural SPair-71k correspondences (Min et al., 2019). Matching repairs 9.5% when the signed effect is nonpositive, 16.8% when it is positive but smaller than the deficit, and 72.7% once it crosses the deficit.

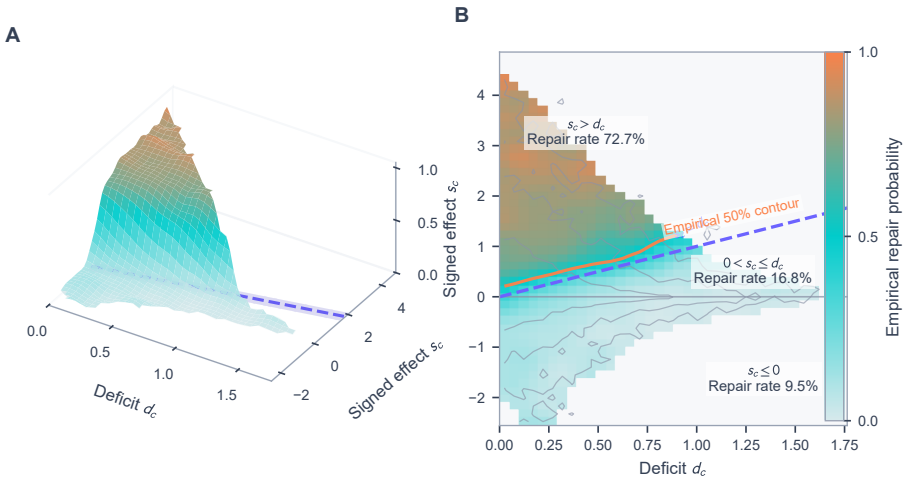


Figure 2: **Actionability landscape.** The incumbent deficit determines the distance to the boundary; the signed intervention effect determines whether the action crosses it. Annotations report conditional repair rates among 30,918 initially incorrect cycles; gray contours show density.

The 55.8-point difference between the last two regions has category-bootstrap 95% CI $[48.4, 61.0]$. Comparable deficits coexist with effects of opposite sign and magnitude, while repair rises sharply only after displacement crosses the boundary.

The coordinates also separate recoverability from incumbent ambiguity. On SPair-71k (Min et al., 2019), signed-cycle recoverability attains .851 repair AUC, compared with .589 for ambiguity. Across SPair-71k and the PF-WILLOW and PF-PASCAL benchmarks (Ham et al., 2016), all 12 benchmark–estimator comparisons favor recoverability, with AUC gains from .148 to .362. The answer-score audit shows the same distinction in a different interface: across Qwen2.5-3B (Qwen Team, 2024) and Phi-3.5-mini (Abdin et al., 2024) on two benchmarks, a fixed CoT intervention (Wei et al., 2022) crosses the margin boundary for only 11.9–24.9% of incumbent errors.

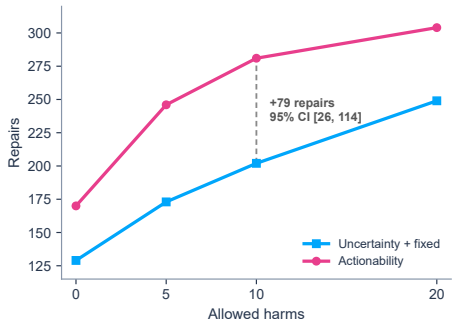


Figure 3: **Full-feedback audit.** Action-conditioned selection yields more repairs at comparable harm.

In a full-feedback cross-fitted audit, action-conditioned selection repairs 281 errors at a budget of ten harms, compared with 202 for uncertainty followed by one fixed action. The gain is 79 repairs, with item-bootstrap 95% CI $[26, 114]$ (Figure 3).

The cost of a state-only ordering grows with the intervention menu. On 514 HotpotQA items, we enumerate every K-action subset of a ten-action portfolio. Relative to a fold-selected fixed action, actionability gains .0109, .0192, .0290, .0403, and .0525 utility per item at $K = 2, 4, 6, 8, 10$; every interval for $K \geq 4$ excludes zero. At $K = 10$, its regret to the hindsight oracle is .1498, compared with .2957 for uncertainty. Additional actions introduce additional orientations that no state-only score can encode.

5.2 SPARSE-LOG ACTION-VALUE ESTIMATION

The 2,000-item benchmark spans BBH, GSM8K, HotpotQA, and MATH (Suzgun et al., 2023; Cobbe et al., 2021; Yang et al., 2018; Hendrycks et al., 2021). Candidate outputs use DeepSeek chat and reasoning models (DeepSeek-AI, 2024; Guo et al., 2025). Its menu contains five interventions and abstention. Each outer-fold training item reveals one uniformly logged action. All estimators share features, logs, folds, menus, 64–64 ReLU backbone width, optimizer, and training schedule; cached counterfactuals evaluate frozen held-out choices only. Across 20 logging seeds, FAVE attains .1395 utility per item versus .1299 for pooled prediction, a paired gain of .0096 with two-way 95% CI $[.0030, .0163]$. Split-head reaches .1363 and also exceeds pooled by .0064 $[.0009, .0117]$; FAVE–split-head is .0032 $[-.0019, .0085]$ (Table 2).

A separate random-forest fit reproduces the FAVE–pooled result: .1367 versus .1278, a gain of .0089 $[.0033, .0147]$. Its split-head estimator obtains .1303; FAVE exceeds it by .0064 $[.0021, .0112]$. The common FAVE–pooled gain holds under both function classes; the FAVE–split contrast depends on how the component heads are estimated.

Across 40 fresh logging seeds, shared response lowers held-out Brier loss from .1461 to .1413 (split minus shared: .00484, $[.00069, .00899]$). Directional Brier loss does not distinguish the endpoints (split minus shared: $-.00364$, $[-.00990, .00257]$). Independent logged validation selects mean sharing weights of .810 for response and .553 for direction at $h = 1$; at $h = 8$, the weights rise to .918 and .878. Adaptive pooling then reaches .07683 utility, .04008 $[.01717, .06196]$ above pooled and .01868 $[.00580, .03215]$ above the matched split endpoint.

The value of conditioning on the item becomes sharp when harm is expensive. Following an exploratory sweep, we fixed $h = 8$ before evaluation on 40 fresh logging seeds; outcome heads remained frozen, and each training fold selected one fixed intervention. Every non-abstaining intervention has negative population utility. Abstention is therefore the population-best fixed policy, whereas itemwise selection retains an oracle value of .2145:

$$\underbrace{\max\{0, \max_I \mathbb{E}[V_h(I \mid X)]\} = 0}_{\text{best fixed}} \; < \; \underbrace{\mathbb{E}[\max\{0, \max_I V_h(I \mid X)\}] = .2145}_{\text{adaptive oracle}}.$$

FAVE attains .08143 utility, exceeding split-head by .01098 $[.00178, .02101]$, pooled by .01630 $[.00346, .02933]$, and the fold-selected non-abstaining action by .10670 $[.08829, .12511]$. Against the population-best fixed policy, abstention, its gain is .08143 $[.06344, .09869]$. Its harm rate is .00236, less than half the pooled rate (.00545) and the split-head rate (.00479).

5.3 ROBUSTNESS TO TOKEN COST

Token price enters equation 8 through $\kappa(I)$, so frozen outcome and cost heads can re-rank the menu without relearning repair or harm. As the predeclared price rises from 0 to .2 utility per 1K tokens, FAVE reduces its mean token use from 23.6 to 16.4; pooled remains substantially heavier, falling from 66.5 to 38.4. The paired FAVE–pooled advantage remains between .0090 and .0096 utility per item, and every pointwise 95% interval excludes zero (Figure 4). The advantage persists as both the action mix and resource consumption adapt to price.

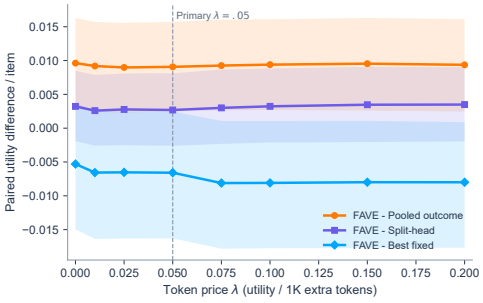


Figure 4: **Cost robustness.** Paired FAVE differences across predeclared token prices with two-way 95% CIs.

At the primary price .05, FAVE obtains .1381 costed utility with 20.2 extra tokens per item, while pooled obtains .1291 with 52.6 tokens. The paired gain is .00907 $[.00249, .01572]$ despite 62% lower token use; before cost, their utilities are .1392 and .1317. Split-head lies between them at .1355 utility and 29.0 tokens. Retrieval also

Table 2: **Sparse feedback and structural sharing.** All panels use matched 64–64 ReLU backbones. Panel D fits the shared and split endpoints on three folds and selects the pooling weights on a disjoint fourth fold. Intervals use a two-way item/logging-seed bootstrap.

A. Policy utility across 20 logging seeds

Estimator	Utility/item	Difference from pooled	95% CI
Pooled	.1299	—	—
Split-head	.1363	+.0064	$[.0009, .0117]$
FAVE	**.1395**	**+.0096**	$[.0030, .0163]$

B. Held-out change Brier loss across 40 fresh logging seeds

Incumbent state	Shared a	Split a_-, a_+	Split–shared (95% CI)
All	**.1413**	.1461	+.00484 $[.00069, .00899]$
Wrong	**.1659**	.1774	+.01142 $[.00167, .02117]$
Correct	.1311	.1332	+.00211 $[-.00229, .00650]$

C. Held-out directional Brier loss across 40 fresh logging seeds

Incumbent state	Shared p	Split $p_-, 1-p_+$	Split–shared (95% CI)
All	.1988	.1952	$-.00364$ $[-.00990, .00257]$
Wrong	.2369	.2340	$-.00297$ $[-.00825, .00229]$
Correct	.1577	.1533	$-.00437$ $[-.01566, .00722]$

D. Validation-selected pooling across 20 new logging seeds

Harm h	Adaptive utility	Adaptive–FAVE endpoint	Adaptive–split-head
1	.13435	$-.00183$ $[-.00443, .00068]$	$+.00268$ $[-.00118, .00635]$
8	**.07683**	$+.00083$ $[-.00143, .00368]$	$+.01868$ $[.00580, .03215]$

provides a strong inexpensive default: best fixed obtains .1447 utility at 16.5 tokens, and FAVE–best-fixed is $-.00659$ $[-.01630, .00247]$.

5.4 Transfer to held-out intervention types

For each transfer task, we remove every training outcome associated with one action type and suppress its action identity. Every estimator instead receives the same frozen five-dimensional representation: use of external context, reasoning depth, branching, verification intensity, and compute scale. The estimators share backbone width, optimization, items, and 20 training seeds.

Aggregated over held-out actions, FAVE reaches .09082 utility, compared with .07055 for pooled; the paired gain is .02027 $[.01415, .02727]$. It exceeds split-head by .00599 $[.00223, .01015]$. The FAVE–pooled intervals exclude zero for retrieval, strong-model escalation, tree-of-thought (Yao et al., 2023), and verification; the chain-of-thought interval contains zero. FAVE scores each omitted type from item features and its frozen representation alone.

5.5 Evaluation from one-action logs

A separate random-forest audit freezes each learned policy, draws a new uniform action log with known propensity $1/5$, and reserves full outcomes for audit truth. For FAVE, direct plug-in evaluation estimates .0901 against truth .1367, whereas IPS and DR estimate .1384 and .1392; DR bias is +.0025 and RMSE .0203 (Dudík et al., 2011). For pooled, direct bias reverses sign at +.0317; DR reduces it to +.0042 with RMSE .0113. Under known randomization, IPS and DR recover frozen-policy value from the same one-action log; direct plug-in estimates retain policy-specific bias.

5.6 Log size and benchmark shift

At 10% of the logged training data, FAVE trails pooled by .00687 $[-.01370, -.00003]$; at 25%, the difference is .00076 $[-.00585, .00687]$. Its advantage appears at 50% (+.01036 $[.00406, .01648]$)

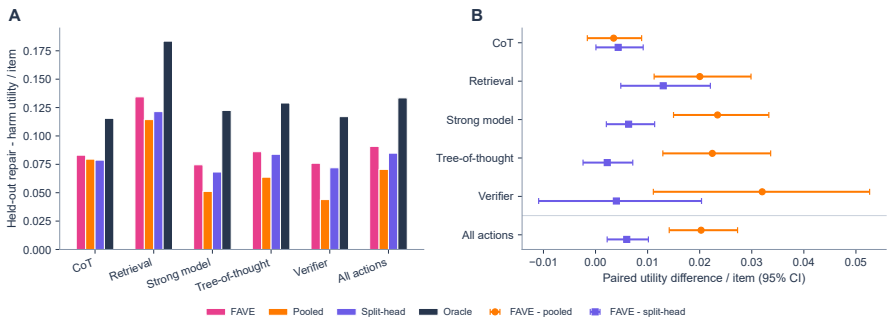


Figure 5: **Held-out-action transfer from a frozen representation.** Outcomes of the held-out intervention type are absent from training. (A) Held-out utilities. (B) Paired FAVE differences with two-way item/seed 95% CIs.

and persists at 100% (+.00907 $[.00249, .01572]$). The structural advantage emerges once at least half of the log is available.

Leave-one-benchmark-out evaluation changes the item distribution rather than the action type. The aggregate FAVE–pooled difference is $-.00750$ $[-.01730, .00100]$: GSM8K and MATH favor FAVE, BBH is indistinguishable, and HotpotQA favors pooled. FAVE transfers more consistently across represented actions than across benchmark distributions in these audits.

6 Discussion

Across observation interfaces, the boundary does not change; only access to action-induced displacement does. Complete scores expose it directly, structured scores recover it modulo a gauge, and black-box logs replace exact displacement with a realized repair–harm–cost consequence. FAVE estimates conditional action value in this last regime.

Approximate sharing trades mechanism fidelity for statistical information. The product defect prices each departure by the prevalence of its incumbent state and the cost of an error in that state. Logged validation chooses this tradeoff: response remains strongly shared at $h = 1$, while the higher harm price draws both components closer to the shared endpoint. The best-fixed comparison isolates state–action heterogeneity. Inexpensive retrieval is strong at $h = 1$; at $h = 8$, every fixed intervention has nonpositive population value while itemwise selection remains useful. This is the gap between $\max\{0, \max_I \mathbb{E}[V_I(X)]\}$ and $\mathbb{E}[\max\{0, \max_I V_I(X)\}]$.

A frozen representation supports extrapolation from four observed action types to a fifth under the same item distribution; benchmark shift changes the map from representation to consequence. Text-derived action embeddings and explicit support diagnostics could replace hand-specified coordinates. Under contextual logging, unequal propensities would change action-specific effective sample sizes and the preferred degree of pooling, but not the repair–harm–cost target. Sequential composition poses a different problem because each intervention changes the state encountered by the next.

7 Conclusion

Uncertainty locates incumbent risk; actionability evaluates the consequence of a particular intervention. We separated incumbent deficit from signed action displacement, proved that calibrated uncertainty cannot rank interventions, and characterized the effect coordinates identifiable across exact, structured, and black-box interfaces. When those coordinates are hidden, FAVE estimates conditional action value from sparse feedback and transfers to unseen intervention types. Its advantage matters most when harm is costly: adaptive intervention retains positive value even when abstention dominates every fixed action. Uncertainty can show that change may be needed; only action-conditioned value can determine which change is worthwhile.

AI USE STATEMENT

Generative AI was used to assist with writing and polishing to improve readability, and with code review and correction. The authors reviewed all AI-assisted content and retain full responsibility for the final content of this work, including its text, claims, code, and other artifacts.

REPRODUCIBILITY STATEMENT

Appendices A–H provide the notation, assumptions, proofs, exact-observation audits, sparse-feedback protocol, off-policy evaluation, transfer analyses, and empirical scope needed to reproduce the reported claims. The experiments use item-grouped training and test splits, stated model architectures and optimization schedules, fixed logging seeds and randomized propensities, predeclared intervention costs, and paired item/seed bootstrap procedures. The logging draws, cached outcomes, token accounting, action representations, and exact-audit records are retained as committed artifacts; Appendix H specifies the interpretation and limitations of each evaluation regime.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.09669.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to route LLMs with confidence tokens. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10859–10878. PMLR, 2025. URL https://proceedings.mlr.press/v267/chuang25b.html.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.04707.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.

Olaf Dünkel, Basavaraj Sunagad, Haoran Wang, David T. Hoffmann, Christian Theobalt, and Adam Kortylewski. SOCO: Benchmarking semantic object correspondence in vision foundation models. *arXiv preprint arXiv:2605.31597*, 2026. URL https://arxiv.org/abs/2605.31597.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1989.

Nigel Fernandez, Branislav Kveton, Ryan A. Rossi, Andrew S. Lan, and Zichao Wang. RADAR: Reasoning-ability and difficulty-aware routing for reasoning LLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25426.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3475–3484, 2016.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, and Sanjiv Kumar. Universal model routing for efficient LLM inference. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ka82fvJ5f1.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. URL https://arxiv.org/abs/1908.10543.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 6355–6384. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.319.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7076–7087. PMLR, 2020.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive LLM routing under budget constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23934–23949. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1301.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Binrui Shen, Qiang Niu, and Shengxin Zhu. Adaptive softassign via hadamard-equipped sinkhorn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17638–17647, 2024.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. doi: 10.1214/aoms/1177703591.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

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*, 2024. URL https://arxiv.org/abs/2406.01574.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Jiaqi Xue, Qian Lou, Jiarong Xing, and Heng Huang. R2-Router: A new paradigm for LLM routing with reasoning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02823.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 27022–27043. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1312.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.12721.

A Notation and Population Semantics

Need, deficit, repair, and harm. $q(x)$ is an incumbent error probability; d is a realized score deficit. Repair is the event that an incorrect incumbent becomes correct, while harm is the event that a correct incumbent becomes incorrect. Thus local (d, s) geometry and population probability factorization are related but not interchangeable.

Unrestricted chain rule. With N and C as in Section 4,

$$\Pr(R_I = 1 \mid x, I) = q(x)a_-(x, I)p_-(x, I), \tag{14}$$
$$\Pr(H_I = 1 \mid x, I) = [1 - q(x)]a_+(x, I)p_+(x, I). \tag{15}$$

Equation 9 follows immediately. FAVE adds equation 10; it is not part of the chain rule.

Local-to-population map. Let D and S_I be the random deficit and signed action effect conditional on x. Away from ties,

$$N = \mathbb{I}\{D > 0\}, \quad R_I = \mathbb{I}\{D > 0, S_I > D\}, \quad H_I = \mathbb{I}\{D < 0, S_I < D\}.$$

Consequently,

$$\begin{aligned} q(x) &= \Pr(D > 0 \mid x), \\ qa_-p_- &= \Pr(D > 0, S_I > D \mid x, I), \\ (1 - q)a_+p_+ &= \Pr(D < 0, S_I < D \mid x, I). \end{aligned} \tag{16}$$

The multiclass events are the conjunctions in Theorem 1.

Factorwise sharing bound. The product defect in equation 12 obeys

$$\begin{aligned} \mathcal{D}_{a,p} \le \mathcal{B}_{a,p} = rq\{p_-|a_- - a| + a|p_- - p|\} \\ + h(1 - q)\{p_+|a_+ - a| + a|p_+ - (1 - p)|\}. \end{aligned} \tag{17}$$

Intrinsic sharing departures. The distances

$$\Delta_a = |a_- - a_+|, \qquad \Delta_p = |p_- + p_+ - 1|$$

measure departure from the two sharing restrictions without choosing a surrogate. Taking $a^\circ = (a_- + a_+)/2$ and $p^\circ = (p_- + 1 - p_+)/2$ in equation 17 gives

$$\begin{aligned} |V - V^{\mathrm{sh}}_{a^\circ,p^\circ}| \le \frac{1}{2}\big[\{rqp_- + h(1 - q)p_+\}\Delta_a \\ + a^\circ\{rq + h(1 - q)\}\Delta_p\big]. \end{aligned} \tag{18}$$

Thus a shared value approximation exists whenever the response and directional departures are small; exact equality is not required.

Adaptive pooling path. Let superscripts sh and sp denote shared and split fits trained independently of the validation log. For $\gamma = (\gamma_a, \gamma_p) \in [0, 1]^2$,

$$\begin{aligned} \widehat{a}^\gamma_\pm &= (1 - \gamma_a)\widehat{a}^{\mathrm{sh}} + \gamma_a\widehat{a}^{\mathrm{sp}}_\pm, \\ \widehat{p}^\gamma_- &= (1 - \gamma_p)\widehat{p}^{\mathrm{sh}} + \gamma_p\widehat{p}^{\mathrm{sp}}_-, \\ \widehat{p}^\gamma_+ &= (1 - \gamma_p)(1 - \widehat{p}^{\mathrm{sh}}) + \gamma_p\widehat{p}^{\mathrm{sp}}_+. \end{aligned} \tag{19}$$

These heads define $\widehat{V}_\gamma = r\widehat{q}\widehat{a}^\gamma_-\widehat{p}^\gamma_- - h(1-\widehat{q})\widehat{a}^\gamma_+\widehat{p}^\gamma_+ - \widehat{\kappa}$. The shared and split endpoints are $\gamma = (0, 0)$ and $(1, 1)$.

Local utility. For magnitude $m = |s|$ and direction probability $p = \Pr(s > 0 \mid \text{action evidence})$, a diagnosed-error local utility is

$$U(\text{apply} \mid d, m, p) = pr\mathbb{I}\{m > d\} - (1 - p)h - \kappa. \tag{20}$$

It acts only if $m > d$ and $p > (h + \kappa)/(r + h)$. This conditional state is not a tuple of population probabilities.

B PROOFS

B.1 PROOF OF THEOREM 1

For every $a \neq y$,

$$F_y^I - F_a^I = (F_y^0 - F_a^0) + s_{I;y,a} = -d_{y,a} + s_{I;y,a}.$$

The target is the unique post-intervention argmax exactly when every displayed margin is positive. A deterministic tie rule replaces strict inequality by the corresponding weak/strict convention; all exact audits have zero ties.

B.2 PROOF OF THEOREM 2

At the common observation let the policy probabilities for the K interventions be $p_1, \dots, p_K$, with remaining mass on abstention. In environment $e = j$, action j has value Δ and all other choices have value at most zero. The incumbent distribution and calibrated uncertainty are identical across environments; only the hidden action orientation is permuted. Regret in $e = j$ is at least $\Delta(1 - p_j)$. Since $\sum_j p_j \leq 1$, some $p_j \leq 1/K$, proving equation 3. Uniform randomization attains equality when nonoptimal values are zero. A deterministic rule assigns zero probability to some action. For harm, make every nonoptimal action harmful and average uniformly over environments: conditional on acting, the unique nonharmful choice is selected with probability at most $1/K$, so some environment has harm at least $\rho(1 - 1/K)$.

B.3 PROOF OF THEOREM 3

If E is invariant under $G \times G$, choose representatives with $T(x_b) = t_b$ and define $\bar{E}(t_0, t_1) = E(x_0, x_1)$. Maximality implies that any other representatives are $g_b x_b$, so invariance makes the definition independent of the choice. Uniqueness holds on the range of $T \times T$. Conversely, any factorization through T is invariant. If a statistic strictly coarsens T, it merges two distinct orbits; an invariant function separating those orbits factors through T but not the coarsening.

B.4 ANSWER-MARGIN ORBIT SEPARATION

Translation cancels from M_y. Conversely, equality $M_y(F) = M_y(F')$ implies $F'_a - F_a = F'_y - F_y$ for every a, hence $F' = F + c\mathbf{1}$. Therefore M_y separates translation orbits and is maximal. The translation orbit has dimension one in $\mathbb{R}^K$, so the quotient has dimension $K-1$; the $K-1$ anchored margins therefore give a minimal-dimensional representation. Applying Theorem 3 to the pre/post pair gives the claim.

B.5 PROOF OF THEOREM 4

Under $Q = \operatorname{diag}(u) P \operatorname{diag}(v)$, every log entry gains $\log u_i + \log v_j$, which cancels in equation 6. Conversely, if $C(P) = C(Q)$ and $D = \log Q - \log P$, then

$$D_{ij} - D_{i1} - D_{1j} + D_{11} = 0.$$

Setting $r_i = D_{i1}$ and $c_j = D_{1j} - D_{11}$ gives $D_{ij} = r_i + c_j$, so $Q = \operatorname{diag}(e^r) P \operatorname{diag}(e^c)$. Hence C is maximal. The row–column nuisance space has dimension $2n - 1$, so the quotient has dimension $n^2 - (2n-1) = (n-1)^2$ and is represented without redundancy by the anchored four-cycle contrasts. Alternating cycles and differences of permutation matrices have zero row and column sums, so every assignment-score contrast is a linear combination of these coordinates.

B.6 SINKHORN CYCLE LAW AND ASSIGNMENT COMPRESSION

For a log-intervention H,

$$P^+_{ij} = u_i P_{ij} e^{H_{ij}} v_j.$$

On every alternating cycle, each row and column potential occurs once with each sign and cancels, leaving

$$\Gamma_c(P^+) = \Gamma_c(P) + \Delta_c(H).$$

Likewise, for permutation matrices X_π, X_ρ,

$$F_\pi(P^+) - F_\rho(P^+) = F_\pi(P) - F_\rho(P) + \langle X_\pi - X_\rho, H\rangle.$$

The difference $X_\pi - X_\rho$ decomposes into disjoint alternating cycles.

B.7 Proof of Theorem 5

Suppress (x, I). The triangle inequality first gives $|V - V^{\mathrm{sh}}_{a,p}| \le \mathcal{D}_{a,p}$. The exact decompositions

$$a_-p_- - ap = p_-(a_- - a) + a(p_- - p),$$
$$a_+p_+ - a(1-p) = p_+(a_+ - a) + a\{p_+ - (1-p)\}$$

show that $\mathcal{D}_{a,p} \le \mathcal{B}_{a,p}$. For $f(q,a,p) = rqap - h(1-q)a(1-p)$, every coordinate derivative on $[0,1]^3$ has magnitude at most $L = \max\{r,h\}$. The mean-value inequality therefore bounds the estimation term $\epsilon_{a,p}$ by

$$L(|\widehat{q} - q| + |\widehat{a} - a| + |\widehat{p} - p|) + |\widehat{\kappa} - \kappa|.$$

The triangle inequality gives $|\widehat{V}_{\mathrm{FAVE}} - V| \le \mathcal{D}_{a,p} + \epsilon_{a,p}$. Optimality of $\widehat{I}$ under $\widehat{V}_{\mathrm{FAVE}}$ then gives

$$\begin{aligned} V(I^\star) - V(\widehat{I}) &\le \{\mathcal{D}_{a,p} + \epsilon_{a,p}\}(I^\star) + \widehat{V}_{\mathrm{FAVE}}(I^\star) - \widehat{V}_{\mathrm{FAVE}}(\widehat{I}) + \{\mathcal{D}_{a,p} + \epsilon_{a,p}\}(\widehat{I}) \\ &\le 2\sup_I\{\mathcal{D}_{a,p}(I) + \epsilon_{a,p}(I)\}. \end{aligned}$$

Finite-grid validation bound. Condition on a training sample and let a finite family $\{\widehat{V}_\gamma : \gamma \in \Gamma\}$ be frozen independently of an i.i.d. validation log of size n_{val}. Let $\mathcal{A}(X)$ contain the logged non-abstention actions, and assume $\mathbb{E}[Y \mid X, I_{\log} = I] = V(I \mid X)$, $|Y|, |\widehat{V}_\gamma(I \mid X)| \le B$, and $e_I(X) \ge e_{\min} > 0$ for $I \in \mathcal{A}(X)$. If $\widehat{\gamma}$ minimizes

$$\widehat{L}(\gamma) = \frac{1}{n_{\mathrm{val}}} \sum_j \frac{\{Y_j - \widehat{V}_\gamma(I_j \mid X_j)\}^2}{e_{I_j}(X_j)},$$

then, with probability at least $1 - \delta$,

$$\mathcal{R}(\widehat{\gamma}) \le \min_{\gamma\in\Gamma} \mathcal{R}(\gamma) + \frac{8B^2}{e_{\min}} \sqrt{\frac{\log(2|\Gamma|/\delta)}{2n_{\mathrm{val}}}}, \tag{21}$$

where $\mathcal{R}(\gamma) = \mathbb{E}_X \sum_{I\in\mathcal{A}(X)} \{\widehat{V}_\gamma(I \mid X) - V(I \mid X)\}^2$. The expected propensity-weighted loss equals $\mathcal{R}(\gamma)$ plus a conditional-variance term common to every candidate. Each observed loss lies in $[0, 4B^2/e_{\min}]$. Hoeffding's inequality, a union bound over Γ, and empirical optimality prove equation 21. Write $J(\pi) = \mathbb{E}[V(\pi(X) \mid X)]$. If π_γ maximizes $\widehat{V}_\gamma$ and $\pi^\star$ maximizes V, both including zero-valued abstention, Jensen's inequality gives

$$J(\pi^\star) - J(\pi_\gamma) \le 2\sqrt{\mathcal{R}(\gamma)}.$$

Logged-feedback identity. Consistency and conditional exchangeability give

$$\mathbb{E}[Y \mid X = x, I_{\log} = I] = \mathbb{E}[Y_I \mid X = x, I_{\log} = I] = \mathbb{E}[Y_I \mid X = x],$$

and overlap makes the conditional mean observable.

B.8 Soft–discrete bridge and companion symmetry

For a 2×2 block a, b, c, d, set $x = (a+d)/2$, $y = (b+c)/2$, $\alpha = (a-d)/2$, and $\beta = (b-c)/2$. Then

$$ad - bc = (x - y)(x + y) + (\beta^2 - \alpha^2).$$

If $|\beta^2 - \alpha^2| < |x - y|(x + y)$, multiplicative circulation and additive swap gain share sign. Hungarian output additionally requires top-two isolation. Separately, if an identity exchange τ leaves x invariant and F is equivariant, then $F(x) = F(\tau x) = \tau F(x)$; equivariance cannot create a missing symmetry-breaking signal.

C Exact-Observation Audits

C.1 Finite-choice answer scores

We score every answer option in FP32 before and after a fixed Direct→CoT intervention for Qwen2.5-3B and Phi-3.5-mini (Qwen Team, 2024; Abdin et al., 2024). Fixed 500-item samples from BBH and MMLU-Pro (Wang et al., 2024) yield 2,000 audits. The all-competitor criterion equation 2 has zero ties, boundary failures, or region mismatches.

C.2 Assignment interface

Semantic plans use frozen DINOv2 features (Oquab et al., 2024). Among 900 scanned SPair-71k replication pairs (Min et al., 2019), 878 admit valid target assignments of size 2–10. There are zero universal-boundary failures and zero ranking changes from Sinkhorn scaling. The maximum discrepancy between an assignment contrast and its cycle decomposition is 1.11×10^{-16}. On 718 SOCO cycles (Dünkel et al., 2026), multiplicative circulation and additive swap-gain signs agree on 95.82%, and on 100% in the lowest imbalance quartile. These checks support the stated conditional interface, not unconditional Hungarian equivalence.

D Sparse-Feedback Protocols

Portfolio and folds. The exact overlap contains 522 BBH, 484 GSM8K, 514 HotpotQA, and 480 MATH items. All five non-abstention actions for an item remain in one fold. Features exclude correctness, repair, harm, target answers, and all post-outcome quantities. Candidate outputs are cached before learning.

Randomized training log. Within every outer fold, one action is sampled uniformly for each training item. FAVE learns need from item features, capability from logged change, and direction conditional on change. Pooled learns repair and harm directly. The split-head ablation retains the mechanism factorization but relaxes its response and direction restrictions. The primary comparison uses the same 64–64 ReLU backbone width, optimizer, 100-epoch schedule, and 20 logging seeds. No threshold or hyperparameter is selected on test outcomes.

Cross-learner replication. A separate random-forest fit (100 trees, depth 8, minimum leaf 6) yields .1278 for pooled, .1303 for split-head, and .1367 for FAVE. Relative to pooled, the FAVE gain is +.0089 $[.0033, .0147]$, while the split-head gain is +.0025 $[-.0030, .0083]$. FAVE exceeds split-head by +.0064 $[.0021, .0112]$. This fit also supplies the frozen policies in the off-policy audit of Appendix F.

Evaluation and uncertainty. The frozen policy chooses among five actions and abstention; only then do cached outcomes score its utility. Item-cluster intervals average over seeds before resampling. Two-way intervals resample both item IDs and logging seeds, preserving all within-item action dependence.

Nested restriction audit. To isolate response sharing from the other FAVE restrictions, we fit four nested models: FAVE shares response and direction; share-a shares response but splits direction; share-direction splits response but shares direction; and split-head relaxes both restrictions. Every model receives the same features, uniformly logged rows, folds, backbone width, optimizer, training schedule, and 40 fresh logging seeds. Cached counterfactual output-change labels enter only the held-out audit. Table 2 reports item-averaged Brier loss; its intervals resample items and logging seeds. The policy-utility point estimates favor full FAVE, but their nested contrasts contain zero. Held-out response prediction provides direct evidence for a-sharing. Directional Brier is evaluated only among changed outputs; its shared-versus-split contrasts contain zero in both incumbent states. The policy ablation supports the complete restriction jointly but leaves the contribution of direction sharing unresolved.

Adaptive pooling audit. For each outer test fold, the cyclically next predefined fold supplies an independent validation log; the other three train every shared and split component. These frozen

Table 3: **Finite-choice Direct→CoT audit.**

Benchmark	Model	Direct	CoT	Repairs	Harms
BBH-500	Qwen2.5-3B	29.4	37.2	88	49
BBH-500	Phi-3.5-mini	53.0	51.8	44	50
MMLU-Pro-500	Qwen2.5-3B	19.8	29.8	89	39
MMLU-Pro-500	Phi-3.5-mini	37.6	39.0	37	30

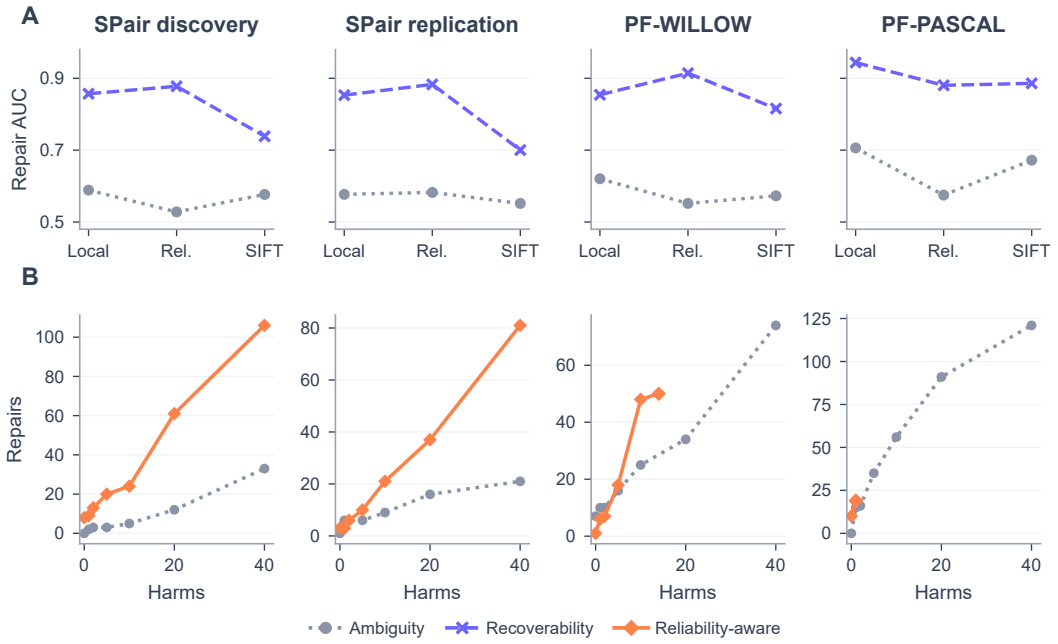


Figure 6: **Matching evidence atlas.** Repair AUC and complete low-harm frontiers for SPair-71k (Min et al., 2019), and for PF-WILLOW and PF-PASCAL (Ham et al., 2016).

fits score both validation and outer-test items and are not refit after selection. Squared repair-minus-h harm error on one logged action per validation item selects (γ_a, γ_p) from $\{0, .25, .5, .75, 1\}^2$, separately for $h = 1$ and $h = 8$; ties favor more sharing. Uniform five-action logging makes inverse-propensity weighting a common factor of five, so it leaves the minimizing pair unchanged. Outer-test counterfactuals enter only after the selected policy is frozen. The audit uses 100 AdamW epochs and 20 new logging seeds (20261801–20261820) on CUDA. All 200 outer selections have a unique validation-loss minimum. At $h = 1$, mean response and direction sharing weights are .810 and .553; at $h = 8$ they are .918 and .878. The resulting utilities are .13435 and .07683. Paired differences at $h = 8$ are +.04008 $[.01717, .06196]$ from pooled and +.01868 $[.00580, .03215]$ from split-head. The complete joint weight distribution is stored with the frozen validation losses and outer-test decisions.

Intermediate-algebra stress test. The stratum was found on the original 20 logging seeds, then fixed with $\lambda = 0$ before running 20 fresh seeds on the same 63 held-out items. It is therefore a held-out logging-seed confirmation on a post-discovery task stratum, not an item-independent replication; Holm correction covers the three baseline contrasts, not the preceding subgroup search. For each comparison, we first average the paired difference over the fresh seeds within each item, remove exact zeros, apply an exact sign test to the remaining item signs, and then Holm-correct across pooled, split-head, and best fixed. The positive/negative item counts are 20/0, 8/1, and 13/1, with adjusted p-values at most .039. Fresh-seed two-way CIs for the three mean gains are $[.0238, .0802]$, $[0, .0222]$, and $[0, .0310]$, respectively. The latter intervals touch zero because most item–seed differences are exact ties; the sign test addresses this discrete, zero-inflated contrast without treating

1,260 pairs as independent observations. The incumbent error rate is 74.6%; its out-of-fold need score has mean .776 and median .908, compared with mean .426 on the remaining MATH items.

The mechanism audit distinguishes learned best fixed from a stratum-oracle fixed action. Retrieval obtains .6508, exactly the itemwise oracle utility, and is oracle-equivalent on all 63 items. FAVE and learned best fixed choose different actions on 188 of 1,260 item–seed pairs; realized utilities favor FAVE on 17, tie on 170, and favor best fixed on one. The sole FAVE–oracle mismatch is item 165 under seed 20260937, where FAVE chooses CoT and retrieval repairs the incumbent. A canonical tie-break labels the oracle decisions as 22 abstentions, 21 CoT, and 20 retrieval choices (normalized entropy .999), but these labels are utility ties. Because retrieval equals itemwise oracle on all 63 items, the audit measures sparse-log stabilization around a dominant fixed action.

Action-ordering disagreement audit. For each baseline, the audit freezes the paired decisions and restricts to item–seed pairs on which both policies intervene but select different actions. It then bootstraps the conditional utility contrast by resampling items and seeds jointly; selection into the set depends only on the frozen policies, not on realized test utility. The committed audit also decomposes the overall gain into different-action, FAVE-only, baseline-only, and identical decision regions.

Cost heads. The GPU cost replication uses the same 64–64 ReLU backbone width and 100-epoch AdamW schedule for FAVE, pooled, and split-head. All share a log-token regression head trained on identical logged rows. Changing λ only recombines frozen outcome and token predictions; best fixed is reselected on training folds for every price.

Harm-price audit. We set repair reward to one and token price to zero, train each outcome model once per fold, and recombine its frozen repair and harm predictions at $h \in \{1, 2, 4, 8\}$. Best fixed is reselected from the same logged training rows at every h. An exploratory run identified $h = 8$ as the primary high-stakes point before evaluation on 40 fresh logging seeds. At this point the full-population utilities of verifier, retrieval, CoT, strong-model, and tree-of-thought fixed actions are $-.0830$, $-.1045$, $-.1285$, $-.2295$, and $-.2990$; the population-best fixed policy therefore abstains with value zero, while itemwise oracle value is .2145. FAVE obtains .08143 and its two-way paired differences from split-head, pooled, and the fold-selected non-abstaining action are +.01098 $[.00178, .02101]$, +.01630 $[.00346, .02933]$, and +.10670 $[.08829, .12511]$. Its difference from the population-best fixed policy, abstention, is +.08143 $[.06344, .09869]$. These are pointwise intervals for the fixed primary harm price. The complete harm-price curve and all frozen decisions accompany the reproducibility archive.

E Additional Empirical Results

E.1 Log-size and cost curves

E.2 Matching robustness and action-set scaling

Across SPair discovery/replication, PF-WILLOW, and PF-PASCAL, recoverability improves repair AUC over ambiguity for all benchmark–estimator pairs; the full atlas and backbone/subgroup analyses are artifact results. A separate 514-item HotpotQA intersection enumerates intervention subsets from a ten-action menu. Candidate values are five-fold cross-fitted by item; for $K = 2, 4, 6, 8, 10$, the audit averages over all 45, 210, 210, 45, and 1 subsets, respectively. Item bootstrap intervals resample each item's gain after averaging over subsets. The committed records contain portfolio-level utility, item-averaged gains, endpoint decisions, and regret to the hindsight oracle.

F Off-Policy Evaluation

Each held-out item independently logs one of five actions with propensity $1/5$, separate from the training log. Direct, IPS, and DR evaluate the same frozen policy (Dudík et al., 2011). Cached full outcomes supply audit truth only. For outcome model μ_I, the doubly robust score is

$$\psi_I = \mu_I(X) + \frac{\mathbb{I}\{I_{\log} = I\}}{e_I(X)}\{Y - \mu_I(X)\}. \tag{22}$$

Table 4: **Logged-data boundary at $\lambda = .05$.**

Logs	FAVE	Pooled	Difference (95% CI)
10%	.1012	.1081	$-.00687$ $[-.01370, -.00003]$
25%	.1152	.1145	$+.00076$ $[-.00585, .00687]$
50%	.1295	.1192	$+.01036$ $[.00406, .01648]$
100%	.1381	.1291	$+.00907$ $[.00249, .01572]$

Table 5: **Harm-price curve on 40 fresh logging seeds.** Outcome models are frozen across prices; best fixed is reselected from each training fold.

h	FAVE	Pooled	Split-head	Fold-selected fixed
1	.1402	.1282	.1351	**.1424**
2	**.1209**	.1098	.1159	.1007
4	**.0973**	.0890	.0931	.0225
8	**.0814**	.0651	.0705	$-.0253$

Conditioning on $X = x$ shows that its mean equals $V(I \mid x)$ when the propensity is correct. If the outcome model is correct, the augmentation has mean zero for any positive fitted propensity. For FAVE, the seed-bootstrap bias interval for DR is $[-.0060, .0113]$; for pooled it is $[-.0006, .0086]$. The empirical regime uses known randomized propensities with potentially misspecified outcome models; interval coverage and the complementary nuisance configuration remain outside this audit.

G Transfer

G.1 Representation-conditional leave-one-action-out

All outcomes of the target action are removed. Action IDs are replaced with five frozen protocol coordinates. The vectors (external context, reasoning depth, branching, verification, compute scale) are CoT $(0, .45, 0, .15, .25)$, retrieval $(1, .10, 0, .10, .08)$, verifier $(0, .25, 0, 1, .35)$, tree-of-thought $(0, .90, 1, .25, .80)$, and strong model $(0, .55, 0, .20, 1)$. These values encode the intervention protocol and use no target outcomes. They define the experimental support for extrapolation; the audit does not attribute transfer to any coordinate in isolation.

G.2 Leave-one-benchmark-out

Benchmark and subset indicators are removed, while known action identities remain. The overall FAVE–pooled difference is $-.00750$ $[-.01730, .00100]$. HotpotQA accounts for the negative aggregate shift, whereas GSM8K and MATH retain positive contrasts. Held-out-action evidence concerns extrapolation under the stated frozen representation; leave-one-benchmark-out evaluates a different distribution shift.

H Empirical Scope and Reproducibility

The theoretical scope follows the declared observation interface. Finite-choice scores use global translation, assignment scores use row–column scaling, and their maximal invariants belong to distinct quotient spaces. Black-box FAVE targets repair and harm relative to observed audit outcomes; latent reasoning states and exact score displacement remain unobserved. The soft-to-discrete bridge retains the conditions stated in Appendix B.

The empirical scope follows the recorded intervention menu. Sparse-feedback results use one randomized action per item and compare FAVE with pooled, split-head, and fold-selected fixed policies on the stated portfolio. Log-size effects are reported at 10, 25, 50, and 100%; transfer uses a fixed five-dimensional action representation and varies across benchmark distributions. Cost intervals refer to the eight predeclared prices. Training and off-policy audits use uniform randomized propensi-

Table 6: **Separate random-forest evaluation of frozen policies.**

Policy	Estimator	Estimate	Bias	RMSE
FAVE (truth .1367)	Direct	.0901	$-.0466$	.0468
	IPS	.1384	+.0017	.0232
	DR	.1392	+.0025	.0203
Pooled (truth .1278)	Direct	.1595	+.0317	.0335
	IPS	.1345	+.0067	.0150
	DR	.1320	+.0042	.0113

Table 7: **Representation-conditional unseen-action transfer.**

Held-out action	FAVE–pooled (95% CI)	FAVE–split-head (95% CI)
CoT	+.0035 $[-.0016, .0089]$	+.0044 $[.0001, .0092]$
Retrieval	+.0200 $[.0113, .0299]$	+.0130 $[.0048, .0221]$
Strong model	+.0234 $[.0150, .0333]$	+.0064 $[.0021, .0114]$
Tree-of-thought	+.0224 $[.0129, .0336]$	+.0023 $[-.0024, .0072]$
Verifier	+.0320 $[.0111, .0526]$	+.0040 $[-.0109, .0204]$
All actions	**+.0203** $[.0141, .0273]$	**+.0060** $[.0022, .0102]$

ties, hence balanced overlap. Cached counterfactuals evaluate frozen decisions rather than an online deployment.

All training/test splits group candidate actions by item. Logging draws, 20-seed outputs, paired item/seed bootstrap inputs, token accounting, action representations, and exact-audit records are stored as committed artifacts. Rejection/abstention has value zero and returns the incumbent output exactly.

Table 8: **Cross-benchmark transfer boundary.**

Held-out benchmark	FAVE	Pooled	Difference (95% CI)
BBH	.1131	.1151	$-.0020$ $[-.0116, .0079]$
GSM8K	.0041	$-.0099$	$+.0140$ $[.0041, .0255]$
HotpotQA	.0900	.1407	$-.0507$ $[-.0841, -.0226]$
MATH	.3029	.2919	$+.0110$ $[.0043, .0192]$
All	.1264	.1339	$-.0075$ $[-.0173, .0010]$