COUNTING EVIDENCE, NOT FEEDBACK: PROVENANCE-AWARE AGENT SELF-IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-improving language-model agents can reduce manual workflow design, but deciding which updates to retain remains difficult when repeated interpretations of the same observations are counted as additional evidence. We introduce Provenance-Adjusted Evidence (PAE), a fixed-budget acceptance rule that traces feedback to its originating observations, removes replayed records, and pools judgments sharing the same sources. PAE assigns influence according to source overlap independently of feedback volume, reducing to the mean paired task gain with complete direct scores and adjusting overlapping joint assessments when per-task scores are incomplete. Across GEPA, TextGrad, and MIPROv2-No-Demos with Qwen3-8B, PAE improves five-run mean accuracy and code pass@1 by 0.72–1.33 percentage points on GSM8K, MATH-500, MBPP+, and IFBench under matched budget ceilings. PAE provides a common acceptance framework for direct and joint evaluations, allowing richer interpretations while keeping their influence tied to the underlying observations.

1 INTRODUCTION

Self-improving language-model agents can adapt reasoning and software workflows from execution feedback, reducing manual redesign. Reflective prompt search has demonstrated this potential across diverse tasks (Agrawal et al., 2026), while textual feedback supports optimization of interconnected modules (Yuksekgonul et al., 2025) and multi-agent coordination (Ma et al., 2026). Sustaining these benefits requires deciding which proposed changes merit retention when evaluation is costly and evidence is limited.

Generating feedback is often easier than obtaining new observations. Iterative self-feedback can improve outputs (Madaan et al., 2023), but successive critiques and judgments may revisit the same execution. These interpretations can be useful without supplying independent empirical evidence. When each interpretation receives equal weight, elaborating on selected observations can change an acceptance decision without evaluating another task. This creates a risk of *evidence pseudoreplication*: the amount of generated feedback can be mistaken for the amount of empirical support.

Existing approaches address complementary reliability problems. Dependence-aware aggregation models interactions among judges (Balasubramanian et al., 2026), and CARE separates quality from shared latent confounders (Zhao et al., 2026). PACE provides per-candidate false-acceptance control under optional stopping, subject to its conditional null (Shawn, 2026). These methods account for statistical dependence or control stopping. A separate question is *how to use richer feedback while keeping support for an update tied to the observations behind it*.

Our goal is to preserve the value of interpretation while preventing its volume from determining evidential strength. We introduce *Provenance-Adjusted Evidence* (PAE), a fixed-budget acceptance rule that complements existing candidate-generation methods. PAE uses recorded evidence origins to separate feedback content from evidential influence: additional interpretations can refine an assessment without automatically giving its underlying observations more weight. Figure 1 illustrates this distinction. With complete direct scores, the rule reduces exactly to averaging paired task gains; for joint assessments with incomplete per-task feedback, it adjusts their contributions for shared observations.

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.

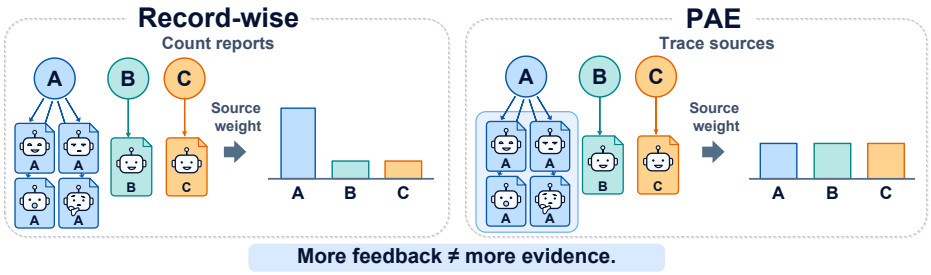


Figure 1: **More feedback does not mean more evidence.** Robot faces represent feedback about sources A–C. Counting reports gives A more weight (left); PAE traces feedback to its sources (right). Equal weights illustrate the complete direct-score case.

We make three contributions:

- **We reveal decision sensitivity to evidence reuse.** With candidates and task observations fixed, 15.0% of 240 candidate pairs change their acceptance decision as the number and allocation of interpretations vary (Section 3), motivating a distinction between additional feedback and additional observations.
- **We propose provenance-adjusted acceptance.** PAE provides a common rule for direct task scores and overlapping joint feedback. Under the stated independent-task model with complete direct scores, its task-averaging rule achieves the optimal worst-case error exponent (Section 4.5).
- **PAE improves task performance across optimizers.** Across GEPA, TextGrad, and MIPROv2-No-Demos, replacing the native selector with PAE improves mean scores by 0.72–1.33 percentage points on four benchmarks under matched budget ceilings (Table 1).

2 Related Work

Feedback-driven self-improvement. The focus has expanded from output refinement and reflective memory (Madaan et al., 2023; Shinn et al., 2023) to persistent prompt and code optimization. MIPROv2 searches over instructions and demonstrations (Opsahl-Ong et al., 2024), whereas TextGrad and ANN propagate textual feedback across modules (Yuksekgonul et al., 2025; Ma et al., 2026). GEPA and Darwin Gödel Machine combine execution feedback with evolutionary search (Agrawal et al., 2026; Zhang et al., 2026). These approaches improve candidate generation and selection. PAE complements them with an acceptance rule that allocates evidential weight among interpretations sharing the same observations.

Dependence-aware judgment aggregation. Dawid–Skene models annotator reliability under conditional independence (Dawid & Skene, 1979), an assumption challenged by correlated LLM errors (Kim et al., 2025). Ising aggregation models interactions among judgments (Balasubramanian et al., 2026), while CARE separates quality from shared latent confounders without ground-truth labels (Zhao et al., 2026). These methods infer statistical dependence from judgment data. PAE uses recorded execution ancestry: it deduplicates invocations, pools identical ancestry, and determines weights from source overlap. The resulting correction captures structural overlap among observations.

Acceptance tests and evidence provenance. PACE replaces score-based commitment with an anytime-valid paired test, controlling per-candidate false acceptance under its conditional null (Shawn, 2026). This provides stopping-time validity under the stated assumptions. Provenance research supplies traceable evidence–execution relations for auditing (Wang et al., 2026). PAE turns these relations into a fixed-budget acceptance rule: direct task scores reduce to root averaging, while overlapping joint feedback permits source-overlap correction when singleton coverage is incomplete. PAE makes a single decision on each fixed evidence block.

3 PRELIMINARY STUDY

Can additional evaluations change a decision without new external evidence? We freeze 240 incumbent–candidate pairs, each with 16 task observations (*roots*). We select fresh interpretations from fixed pools under nested per-root caps $L \in \{1, 2, 4, 8, 16\}$, keeping candidates, observations, evaluator, and rubric unchanged. At $L = 1$, each root contributes one judgment; larger layouts use unequal, pre-specified quotas. Acceptance requires a positive candidate-minus-incumbent mean, without provenance adjustment (Appendix B.6).

Observed instability. In Figure 2(a), decisions differing from $L = 1$ increase from 4.6% (11/240) at $L = 2$ to 12.9% (31/240) at $L = 16$; 15.0% (36/240) change at least once along the path. Across 9,600 counterfactual layouts (40 per pair), flip frequency is 0% (0/287) in the low-concentration, low-conflict region and 39.4% (448/1,137) in the high-concentration, high-conflict region (Figure 2(b); thresholds in Table 6). These frequencies describe repeated layouts clustered within candidate pairs.

Implication. Audit agreement varies non-monotonically with L (Table 5); added interpretations change both judgment content and relative root weights. This motivates our method's central distinction: interpretations may refine a source's judgment, while source influence is assigned independently of interpretation count.

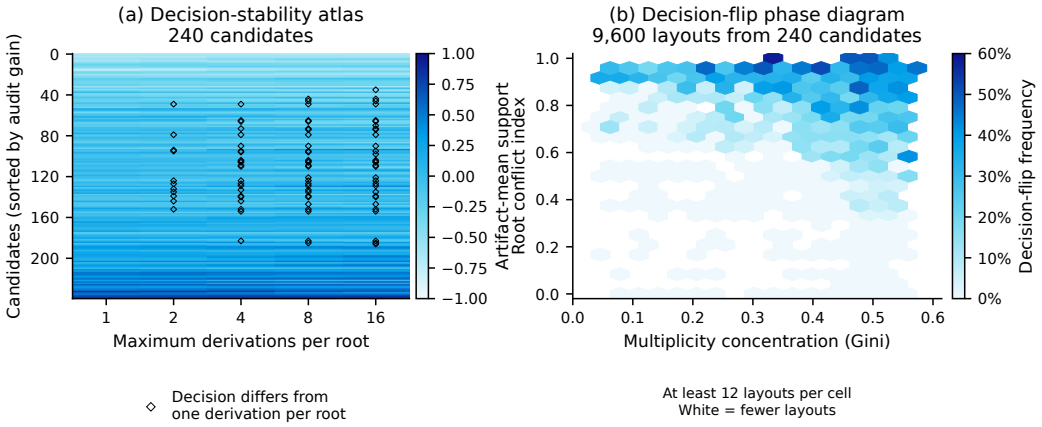


Figure 2: **Decision instability under fixed external evidence. (a)** Mean support for 240 candidate pairs, sorted by audit gain; diamonds mark decisions differing from $L = 1$. **(b)** Flip frequency across 9,600 layouts by multiplicity concentration and root conflict. Cells contain at least 12 layouts; white regions fall below this threshold. See Appendix B.6.

4 METHOD

4.1 OVERVIEW

Provenance-Adjusted Evidence (PAE) weights comparisons by their sources. It removes replayed records, excludes abstentions, and averages comparable judgments that use the same task roots. With direct paired task scores, PAE reduces exactly to root averaging. When an evaluation interface provides overlapping bundle-level comparisons without complete singleton feedback, PAE accounts for source overlap through a minimum-norm projection. A maximum-entropy tie-break makes the weights unique, and a fixed numerical guard determines the decision. Figure 3 summarizes the procedure.

4.2 COMPARABLE JUDGMENTS AND SOURCE GROUPS

Fix an incumbent S_t and candidate S'_t before evaluating them on a fresh block of K tasks. The target is the expected gain $\Delta_t = \mathbb{E}_{x\sim\mathcal{D}}[R(S'_t, x) - R(S_t, x)]$, where $\mathcal{D}$ is the task distribution and

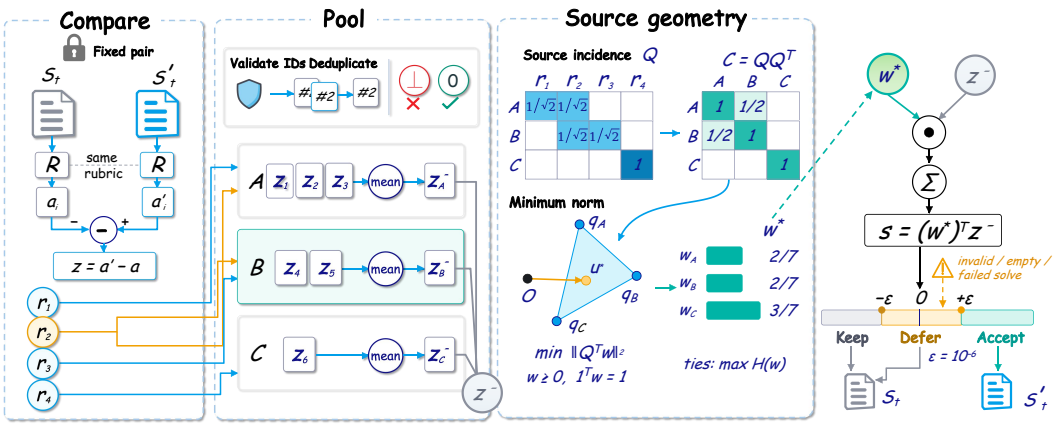


Figure 3: **PAE with overlapping joint feedback.** For a fixed incumbent–candidate pair and shared rubric, PAE deduplicates invocation IDs, excludes abstentions, retains valid zero differences, and averages comparisons with identical root ancestry. The normalized incidence matrix Q defines source overlap $C = QQ^\top$. Minimum-norm projection onto the convex hull of its rows, followed by maximum-entropy tie-breaking, gives weights $w^\star$ independently of group means $\bar{z}$. The support $s_{\mathrm{PAE}} = (w^\star)^\top \bar{z}$ determines the decision under Equation (7); both KEEP and DEFER retain S_t. This extension has incomplete singleton coverage; the primary direct-score protocol reduces to root averaging.

$R \in [0,1]$ is the fixed task metric. Each task defines a *root* containing its paired executions and evaluator feedback; repeated runs share that root. The primary protocol retains one direct score difference per root, preserving continuous score magnitudes, and makes no additional judgment calls. Its aggregate score, or *support*, is the empirical paired gain. Any unavailable paired score returns DEFER.

The bundle extension handles interfaces that judge several tasks jointly, such as a requirement spanning multiple interaction episodes. Evaluation units and their evidence inputs are declared before execution. Each unit produces one comparison under a common normalized rubric. Every available direct singleton comparison is retained alongside the declared joint comparisons. Appendix A.1.1 specifies the direct-score and fixed language-judge adapters and the terminal-record schema.

Distinguish equal scores from abstention. For terminal record i, let $a_i, a_i' \in [0,1]$ denote the incumbent and candidate scores under the same rubric. The comparison value is

$$z_i = \begin{cases} a_i' - a_i, & \text{a comparable score pair is available,} \\ \perp, & \text{the evaluator abstains.} \end{cases} \tag{1}$$

The symbol $\perp$ denotes a missing comparison, which is excluded before grouping; a valid tie gives $z_i = 0$ and is retained. Malformed responses are logged separately and excluded without replacement. Adding valid terminal abstentions therefore leaves comparisons and weights unchanged. Bundle scores target differences under the declared rubric; direct task scores target the population gain Δ_t.

Deduplicate calls, then group by ancestry. An external logger records the evidence DAG $G = (V, E)$. Root r_k has ancestry $A(r_k) = \{k\}$; each non-root node v inherits $A(v) = \bigcup_{u \in \mathrm{Pa}(v)} A(u)$ from the evidence nodes it actually reads, $\mathrm{Pa}(v)$. Set union counts each root once, regardless of repeated paths. Every executed comparison receives an immutable invocation ID ι_i: cached copies keep that ID, whereas separate calls receive distinct IDs. Conflicting records with one ID invalidate the decision. Thus replay is distinguished from additional interpretation (Appendix A.1.2).

After deduplication and exclusions, group records by identical nonempty ancestry. Let $\mathcal{U}_g$ contain the records with ancestry A_g, for $g = 1, \ldots, M$, and define

$$\bar{z}_g = \frac{1}{|\mathcal{U}_g|} \sum_{i \in \mathcal{U}_g} z_i, \qquad q_g = \frac{\mathbf{1}_{A_g}}{\sqrt{|A_g|}} \in \mathbb{R}^K, \qquad Q = [q_1, \ldots, q_M]^\top. \tag{2}$$

Here $\bar{z}_g$ is the group mean, $\mathbf{1}_{A_g}$ is the binary indicator of its roots, and $Q \in \mathbb{R}^{M\times K}$ stacks the normalized incidence vectors. Additional judgments can refine the group mean; its weight depends only on ancestry. Interpretations of the same observations may be statistically dependent.

4.3 A UNIQUE AGGREGATE FROM SOURCE OVERLAP

Distinct source groups can still reuse the same roots. Their normalized overlap is the Gram matrix

$$C = QQ^\top, \qquad C_{gh} = \frac{|A_g \cap A_h|}{\sqrt{|A_g||A_h|}}. \tag{3}$$

The matrix $C \in \mathbb{R}^{M\times M}$ is positive semidefinite with unit diagonal. Square-root normalization gives all groups equal self-overlap: two disjoint groups containing one and four roots receive equal weight. This normalization balances exposure to shared observations and treats disjoint groups equally regardless of size.

Let $\mathcal{W}_M = \{w \in \mathbb{R}^M : w \geq 0, \mathbf{1}^\top w = 1\}$ be the probability simplex. We first compute

$$\widetilde{w} \in \operatorname*{arg\,min}_{w\in\mathcal{W}_M} \|Q^\top w\|_2^2, \qquad u^\star = Q^\top \widetilde{w}. \tag{4}$$

The vector $u^\star \in \mathbb{R}^K$ is the unique projection of the origin onto $\operatorname{conv}\{q_1, \ldots, q_M\}$.

Several weight vectors may represent $u^\star$. We resolve this ambiguity by maximizing entropy, then compute the support:

$$\begin{aligned} w^\star &= \operatorname*{arg\,max}_{w\in\mathcal{W}_M,\, Q^\top w = u^\star} H(w), \qquad H(w) = -\sum_{g=1}^{M} w_g \log w_g, \\ s_{\mathrm{PAE}} &= (w^\star)^\top \bar{z}, \qquad \bar{z} = (\bar{z}_1, \ldots, \bar{z}_M)^\top. \end{aligned} \tag{5}$$

With $0 \log 0 = 0$, strict concavity of entropy ensures a unique $w^\star$ and hence a support $s_{\mathrm{PAE}} \in [-1, 1]$. Entropy only breaks ties among primary minimizers, requiring no regularization coefficient. Both optimizations depend only on provenance; $\bar{z}$ enters only the final support calculation. Appendices A.2 and D.1 give the numerical procedure and uniqueness proof.

Why minimize the source norm? Each coordinate of $Q^\top w$ measures one root's total exposure across weighted judgments. For a source perturbation $e \in \mathbb{R}^K$ with radius $\rho \geq 0$,

$$\sup_{\|e\|_2 \leq \rho} |w^\top Q e| = \rho \|Q^\top w\|_2. \tag{6}$$

Equation (4) therefore minimizes worst-case support sensitivity to additive perturbations Qe; ρ is used only for this interpretation. Here C describes structural overlap. Statistical error dependence is a separate estimation problem addressed by fitted models (Balasubramanian et al., 2026; Zhao et al., 2026).

Closed forms and scope. Pairwise disjoint groups give $C = I_M$ and $w^\star = \mathbf{1}/M$. If every represented root has a singleton group, the unique optimum assigns $1/K_\cup$ to each singleton and zero to all multi-root groups, where $K_\cup = |\bigcup_g A_g|$. The primary protocol thus averages roots without a solver or additional multi-root judgments. Multi-root weighting matters only when singleton coverage is incomplete. Appendix D.2 proves these results and defines the structural diagnostic $N_{\mathrm{prov}} = 1/\|u^\star\|_2^2$.

4.4 ACCEPTANCE AT A FIXED OPERATING POINT

PAE makes one decision per completed evidence block, using a fixed threshold of zero. With exact support, it accepts when $s_{\mathrm{PAE}} > 0$ and retains the incumbent when $s_{\mathrm{PAE}} \leq 0$. Closed forms use exact arithmetic on recorded finite-decimal scores. For numerically computed support $\widehat{s}$, the rule is

$$\mathrm{status}(\widehat{s}) = \begin{cases} \textsc{Accept}, & \widehat{s} > \epsilon_{\mathrm{num}}, \\ \textsc{Keep}, & \widehat{s} < -\epsilon_{\mathrm{num}}, \\ \textsc{Defer}, & |\widehat{s}| \leq \epsilon_{\mathrm{num}}, \end{cases} \qquad \epsilon_{\mathrm{num}} = 10^{-6}. \tag{7}$$

The fixed tolerance ϵ_{num} specifies numerical resolution. Only ACCEPT sets $S_{t+1} = S'_t$; both other statuses retain S_t. Invalid provenance, no retained comparisons, or failed numerical checks return DEFER. The guard prevents acceptance based on roundoff-scale signs. Certifying support error requires an additional stability bound beyond the objective residual. Appendix A.2 details the solver checks and a four-cycle boundary case.

Candidate generation and acceptance use separate task families. Each candidate is fixed before accessing a fresh preallocated acceptance block; final audit tasks remain inaccessible until the run ends. PAE requires no threshold calibration. Appendices A.1.3 and A.3 give the full algorithm and its statistical and computational scope.

4.5 INFORMATION LIMITS OF REPEATED EVIDENCE

Fix the incumbent, candidate, and evaluator before drawing K i.i.d. complete roots $X_{1:K}$ with valid direct-score differences $D_k = R(S', x_k) - R(S, x_k) \in [-1, 1]$. Derived artifacts may process these roots adaptively through a common channel under all hypotheses, using only the sampled observations. For $0 < \gamma < 1$, let $\mathcal{R}_K(\phi; \gamma)$ denote the worst-case probability of accepting a worse candidate or rejecting a better one over score laws with $|\mathbb{E}D_k| \geq \gamma$; deferral counts as retention.

Theorem 1 (Independent-root information limit)**.** *Every artifact-based decision can be matched by a randomized root-based decision with identical error probabilities. Let $J(\gamma) = -\frac{1}{2}\log(1-\gamma^2)$. For some constant $c_\gamma > 0$ independent of K, the primary PAE rule, which accepts exactly when $\sum_k D_k > 0$, satisfies*

$$\frac{c_\gamma}{\sqrt{K+2}} e^{-KJ(\gamma)} \leq \mathcal{R}_K(\phi_{\text{PAE}}; \gamma) \leq e^{-KJ(\gamma)}. \tag{8}$$

On a pair of admissible root experiments with gains $\pm\gamma$, every artifact-based decision has maximum error at least this lower bound. Thus PAE attains the optimal worst-case error exponent $J(\gamma)$, regardless of artifact count.

Proof sketch. Averaging an artifact decision over the channel's randomness yields a root-based decision with the same errors. A binary root experiment with gains $\pm\gamma$ gives the lower bound via its central binomial mass. For PAE, bounding the exponential by the line joining its endpoint values and optimizing at $\lambda = \operatorname{arctanh}\gamma$ gives the upper bound. Taking logarithms yields the matching exponent. Appendix D.3 provides full proofs, the exact finite-sample benchmark, and root-budget bounds. □

The guarantee applies to the error exponent under complete direct scores. Appendix D.3 distinguishes this optimal exponent from exact finite-sample minimax risk. For the bundle extension, N_{prov} measures source geometry and the guard specifies numerical resolution.

5 EXPERIMENTS

We evaluate task performance (RQ1, Section 5.2), acceptance behavior (RQ2, Section 5.3), decision shifts under repeated feedback (RQ3, Section 5.4), and source-exposure redistribution (RQ4, Section 5.5).

5.1 EXPERIMENTAL SETUP

Benchmarks and metrics. We use GSM8K, SVAMP, MATH-500, and AIME 2025 for mathematics (Cobbe et al., 2021; Patel et al., 2021; Hendrycks et al., 2021; Agrawal et al., 2026); HumanEval+, MBPP+, and LiveCodeBench for code (Liu et al., 2023; Jain et al., 2024); and ARC-Challenge and IFBench for generalization (Clark et al., 2018; Pyatkin et al., 2025). Metrics are answer accuracy, pass@1, and IFBench prompt-level loose accuracy, supplemented by false/harmful-commit rates and beneficial-update recall. Each coding problem or instruction-following prompt counts as one task, regardless of its number of tests or constraints. Final scores are measured on held-out audit sets. Commit metrics complement these scores by quantifying update risk and how often beneficial candidates are accepted.

Table 1: **Matched optimizer comparison.** Qwen3-8B scores (%, ↑), five-run means; paired initialization, task blocks, and budget ceilings. †: fresh-block native selector. Blue marks PAE; bold/underline rank selectors within each pair. Green subscripts are percentage-point differences from GEPA (prompt-only) or MIPROv2 (demonstration-enabled), the strongest baseline in each block. Demonstrations count toward cost.

Setting	Method	GSM8K	MATH-500	MBPP+	IFBench
Prompt-only, single-agent interface	GEPA† (Agrawal et al., 2026)	89.31	70.52	71.43	38.40
	GEPA + PAE (ours)	**90.14** $_{+0.83}$	**71.48** $_{+0.96}$	**72.43** $_{+1.01}$	**39.73** $_{+1.33}$
	TextGrad† (Yuksekgonul et al., 2025)	87.81	69.04	69.89	36.73
	TextGrad + PAE (ours)	**88.69** $_{-0.62}$	**69.76** $_{-0.76}$	**70.95** $_{-0.48}$	**37.87** $_{-0.53}$
	MIPROv2-No-Demos† (Opsahl-Ong et al., 2024)	88.01	68.56	70.53	37.27
	MIPROv2-No-Demos + PAE (ours)	**88.90** $_{-0.41}$	**69.28** $_{-1.24}$	**71.27** $_{-0.16}$	**38.20** $_{-0.20}$
Demonstration-enabled sensitivity	MIPROv2† (Opsahl-Ong et al., 2024)	89.96	71.00	72.12	37.60
	MIPROv2 + PAE (ours)	**90.55** $_{+0.59}$	**71.84** $_{+0.84}$	**72.80** $_{+0.69}$	**38.93** $_{+1.33}$

Comparison protocol. Acceptance baselines are greedy/root averaging, a paired test, and PACE (Shawn, 2026). Aggregation baselines are mean/vote, Dawid–Skene (Dawid & Skene, 1979), Ising (Balasubramanian et al., 2026), and CARE (Zhao et al., 2026); system baselines are GEPA (Agrawal et al., 2026), MIPROv2 (Opsahl-Ong et al., 2024), TextGrad (Yuksekgonul et al., 2025), and ANN (Ma et al., 2026). Comparisons match backbones, data access, and budgets within each track.

Backbones and implementation. We use Qwen2.5-3B-Instruct, Qwen3-8B, and Llama-3.1-8B-Instruct with frozen weights (Qwen et al., 2025; Yang et al., 2025; Grattafiori et al., 2024). Appendix B details splits, comparison tracks, implementation, budgets, and uncertainty estimation.

5.2 RQ1: Task performance across settings

Replacing native selectors with PAE improves GEPA, TextGrad, and MIPROv2-No-Demos by 0.72–1.33 percentage points on the four primary benchmarks (Table 1). Under complete direct feedback, PAE equals root averaging, so these gains measure selector replacement. GEPA + PAE achieves the highest mean scores on all nine benchmarks (Table 2) and gains 0.80–1.27 points on two additional backbones (Figure 4); ANN gains 0.67–1.96 points (Appendix C.5). PAE improves each optimizer's native selector, although TextGrad and MIPROv2-No-Demos with PAE still trail native GEPA on all four benchmarks.

All four primary GEPA paired 95% intervals exclude zero, with 4–8 extra task calls and 0.06–0.08 extra optimization GPU hours per run (Table 10). Root-budget sensitivity and realized costs appear in Table 13.

Table 2: **Final task performance.** Qwen3-8B scores (%, ↑), five-run means; greedy decoding, thinking disabled. †: fresh-block adaptation. Bold/underline mark best/second-best scores. Evaluation details: Appendix B.7.

	Mathematics				Code			General	
Method	GSM8K	SVAMP	MATH-500	AIME 2025	HumanEval+	MBPP+	LiveCodeBench	ARC-Challenge	IFBench
Initial	86.05	90.60	66.80	20.00	68.90	67.72	22.29	81.57	35.33
TextGrad† (Yuksekgonul et al., 2025)	87.81	92.18	69.04	20.67	70.85	69.89	22.97	83.05	36.73
MIPROv2-No-Demos† (Opsahl-Ong et al., 2024)	88.01	92.02	68.56	20.00	70.49	70.53	23.31	83.29	37.27
GEPA† (Agrawal et al., 2026)	89.31	93.30	70.52	22.67	72.20	71.43	24.69	83.94	38.40
GEPA + PAE (ours)	**90.14**	**93.88**	**71.48**	**24.67**	**73.54**	**72.43**	**25.83**	**84.76**	**39.73**

Takeaway (RQ1). Changing the acceptance rule improves task performance across optimizers without changing their proposal mechanisms.

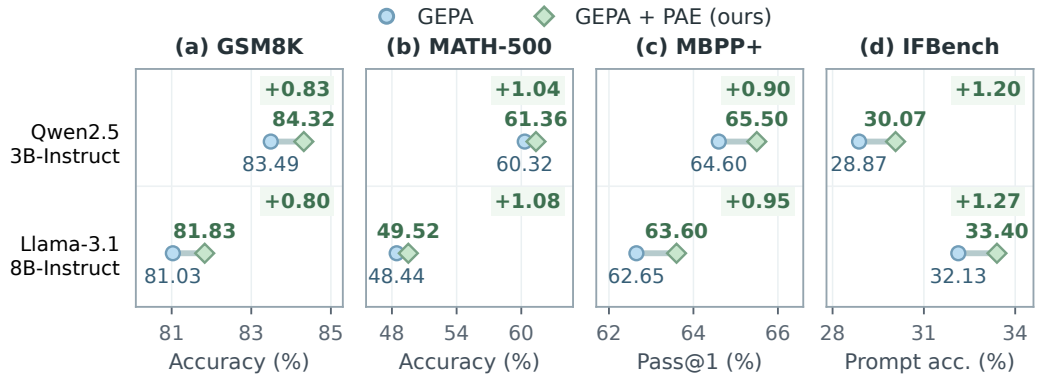


Figure 4: **Cross-backbone gains.** Five-run means for GEPA (Agrawal et al., 2026) (blue circles) and GEPA + PAE (green diamonds); rows identify backbones. Labels show percentage-point gains; axes differ by benchmark. Full results: Table 8.

5.3 RQ2: Acceptance behavior and judge aggregation

PAE's gains accompany higher beneficial-update recall and nonzero commit risk. It accepts 6.2–7.2 updates per run, versus 0.4–0.8 for the paired test and none for PACE (Table 7). On a common candidate bank, PAE recovers 54.76%/58.54% of beneficial GSM8K/MBPP+ updates (Table 11), separating gate behavior from trajectory differences.

With identical, balanced singleton judge evidence, continuous PAE equals the mean and binary PAE equals majority vote (Table 3). CARE-SVD has the lowest continuous-score commit risk and CARE-Tensor the lowest binary risk, both at lower recall than PAE. Source accounting and judge-dependence estimation thus address distinct aspects of aggregation (Appendix C.2).

Table 3: **Shared-evidence aggregation.** Pooled percentages over 100 frozen pairs per benchmark, 16 singleton roots per pair, and three judgments per root. Binary PAE averages $2b - 1$. Blue marks PAE; bold/underline mark best/second-best values within evidence groups, including ties. Green subscripts show percentage-point improvements over each metric's strongest same-group baseline; positive is better.

		GSM8K			MBPP+		
Evidence	Aggregator	FCR↓	HCR_0 ↓	$Recall_0$ ↑	FCR↓	HCR_0 ↓	$Recall_0$ ↑
Continuous score differences	Mean (Zhao et al., 2026)	27.27	18.18	**68.09**	32.50	22.50	**62.79**
	CARE-SVD (Zhao et al., 2026)	**21.05**	**13.16**	63.83	**26.47**	**17.65**	58.14
	PAE (ours)	27.27$_{-6.22}$	18.18$_{-5.02}$	**68.09**$_{+0.00}$	32.50$_{-6.03}$	22.50$_{-4.85}$	**62.79**$_{+0.00}$
Binary preferences $b = \mathbf{1}[z > 0]$	Majority vote (Balasubramanian et al., 2026)	23.08	15.38	**63.83**	34.29	22.86	**53.49**
	Dawid–Skene (Dawid & Skene, 1979)	22.86	14.29	57.45	29.03	19.35	51.16
	Ising (class-indep.) (Balasubramanian et al., 2026)	22.22	11.11	59.57	28.12	15.62	**53.49**
	Ising (class-dep.) (Balasubramanian et al., 2026)	20.00	11.43	59.57	26.67	16.67	51.16
	CARE-SVD (Zhao et al., 2026)	19.44	11.11	61.70	25.81	16.13	**53.49**
	CARE-Tensor (Zhao et al., 2026)	**18.18**	**9.09**	57.45	**24.14**	**13.79**	51.16
	PAE (binary adapter)	23.08$_{-4.90}$	15.38$_{-6.29}$	**63.83**$_{+0.00}$	34.29$_{-10.15}$	22.86$_{-9.06}$	**53.49**$_{+0.00}$

Takeaway (RQ2). Evaluate acceptance by task gains, harmful commits, and missed beneficial updates together.

5.4 RQ3: Why repeated feedback changes decisions

Using 9,600 layouts from 240 incumbent–candidate pairs with $K = 16$ fixed roots, we decompose score shifts into root reweighting and within-root judgment changes, normalized by the

single-interpretation baseline score (Appendix C.7). Of 4,905 reversals (51.1%), 71.4% have a larger reweighting component, with similar proportions for audit-correcting and audit-harming flips (70.6%/71.7%; Figure 5).

PAE remains unchanged under cache aliases but can respond to fresh interpretations or observations (Table 12); complete singleton coverage makes it identical to root averaging.

Takeaway (RQ3). Repeated feedback can reverse decisions by changing source influence, even with observations fixed. Separate reweighting from judgment refinement when evaluating such changes.

5.5 RQ4: How PAE redistributes source influence

Across 1,600 controlled layouts with $K = 16$ roots, PAE shifts normalized root exposure toward sources underrepresented by equal-group averaging (Figure 6; definitions in Appendix C.8). From the lowest to highest overlap decile, mean transfer increases monotonically from $+0.02$ to $+0.51$ for the least-exposed root and decreases from -0.47 to -2.17 for the most exposed. PAE reduces $\|Q^\top w\|_2^2$ by a median 40.2% and lowers exposure Gini in 95.7% of layouts. Structural and numerical controls appear in Table 14.

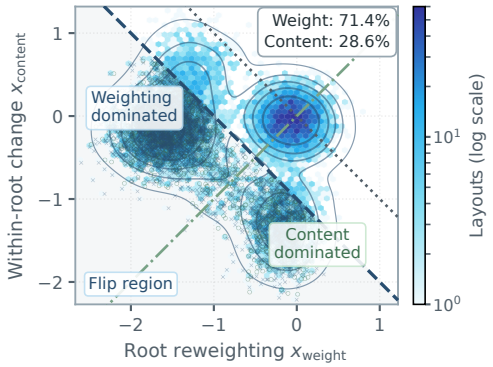


Figure 5: **Root reweighting dominates decision reversals.** Hexagons count layouts; contours show density. Dashed, dotted, and dash-dotted lines mark the flip boundary, zero shift, and equal components. Green circles correct audit mismatches; blue crosses introduce them. Inset percentages condition on flips.

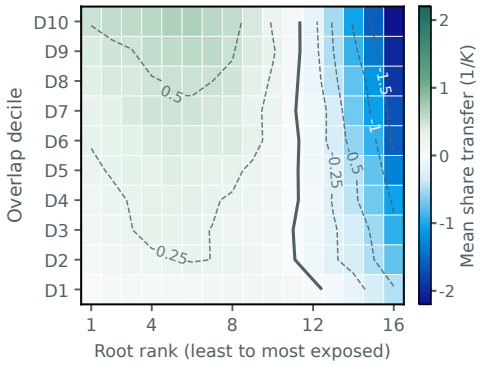


Figure 6: **PAE shifts exposure toward underrepresented roots.** Rows are overlap deciles (160 layouts each); columns rank roots from least to most exposed under equal-group averaging. Colors show mean share transfer in units of $1/K$: green indicates gain and blue loss. The solid contour marks zero.

Takeaway (RQ4). Equal report weights can leave source influence unbalanced; PAE accounts for overlap when allocating that influence.

6 Conclusion

We studied how evidence reuse can change which agent updates are retained, even when the underlying observations remain fixed. PAE makes source attribution part of acceptance, allowing assessments to evolve while preventing their repetition from automatically amplifying evidential weight. Improvements across multiple optimizers demonstrate the practical value of revisiting acceptance design, while our independent-task analysis establishes the optimal error exponent of PAE with complete direct scores. Together, these findings motivate evaluation protocols that distinguish new observations, revised judgments, and shifts in source influence. An open question is how provenance can guide adaptive evaluation budgets, allocating effort between refining existing assessments and acquiring new observations.

AI USE STATEMENT

We used generative AI tools to assist with literature review, checking mathematical proofs, interpreting experimental results, and polishing the language of the manuscript. The authors reviewed and verified the AI-assisted content and take full responsibility for the accuracy, originality, and integrity of the final manuscript, including its claims, proofs, results, references, and conclusions.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.19457.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Krishnakumar Balasubramanian, Aleksandr Podkopaev, and Shiva Prasad Kasiviswanathan. Dependence-aware label aggregation for LLM-as-a-judge via ising models, 2026. URL https://arxiv.org/abs/2601.22336. arXiv:2601.22336.

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

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952. doi: 10.1214/aoms/1177729330. URL https://doi.org/10.1214/aoms/1177729330.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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, 2021. URL https://arxiv.org/abs/2110.14168.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979. doi: 10.2307/2346806. URL https://doi.org/10.2307/2346806.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2103.03874.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.01210.

Xiaowen Ma, Yunpu Ma, Chenyang Lin, Sikuan Yan, Jinhe Bi, Zixuan Cao, Yijun Tian, Volker Tresp, and Hinrich Schuetze. Self-evolving multi-agent systems via textual backpropagation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9918–9951. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.483. URL https://aclanthology.org/2026.findings-acl.483/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9340–9366. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.525. URL https://aclanthology.org/2024.emnlp-main.525/.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, 2021. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2507.02833.

Qwen et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zayx Shawn. PACE: Anytime-valid acceptance tests for self-evolving agents, 2026. URL https://arxiv.org/abs/2606.08106. arXiv:2606.08106.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Yiqi Wang, Jiaqi Zhang, Zhangkai Wu, Taotao Cai, Zirui Liu, Zequn Sun, Manqing Dong, Mingkai Zheng, Yiqun Duan, Xuefei Yin, and Yanming Zhu. From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents, 2026. URL https://arxiv.org/abs/2606.04990v5. Version 5.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639:609–616, 2025. doi: 10.1038/s41586-025-08661-4. URL https://www.nature.com/articles/s41586-025-08661-4.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin gödel machine: Open-ended evolution of self-improving agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pUpzQZTvGY.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-aware aggregation for reliable LLM evaluation, 2026. URL https://arxiv.org/abs/2603.00039v2. arXiv:2603.00039v2.

Appendix

Contents

A Technical details . . . 14
- A.1 Evaluation and provenance protocol . . . 14
- A.2 Optimization and finite-precision decisions . . . 15
- A.3 Computational cost and interpretation . . . 16

B Experimental details . . . 17
- B.1 Benchmarks, scoring, and data separation . . . 17
- B.2 Comparable methods and evaluation tracks . . . 18
- B.3 Acceptance metrics and uncertainty . . . 19
- B.4 Evidence interventions and overlap diagnostics . . . 19
- B.5 Backbones, implementation, and budgets . . . 20
- B.6 Preliminary-study protocol and visualization . . . 20
- B.7 Reporting conventions and literature alignment . . . 22
- B.8 Decision-shift protocol and diagnostics . . . 23
- B.9 Source-exposure protocol and diagnostics . . . 23

C Additional experimental results . . . 25
- C.1 Preliminary-study results . . . 25
- C.2 Shared-evidence risk and recall . . . 25
- C.3 Direct-feedback acceptance control . . . 25
- C.4 Additional backbones . . . 26
- C.5 Multi-agent comparison . . . 27
- C.6 Supplementary controls and diagnostics . . . 28
- C.7 Decision-shift mechanism analysis . . . 33
- C.8 Source-exposure redistribution . . . 33

D Proofs . . . 34
- D.1 Projection, linear fibers, and entropy . . . 34
- D.2 Structural properties . . . 34
- D.3 Evidence information limit . . . 35

A TECHNICAL DETAILS

A.1 EVALUATION AND PROVENANCE PROTOCOL

A.1.1 ACQUISITION AND TERMINAL RECORDS

Primary direct-score protocol. Before a run, the controller fixes the task-family split, task metric R, task blocks, execution settings, and number of rounds. At round t, it freezes (S_t, S'_t), executes both systems once on each task in the next block, and applies the same evaluator to both outputs. The terminal set contains exactly one paired score difference per task. The metric's prescribed failure score is a valid score; an unavailable evaluation is an acquisition failure and defers the entire primary decision. The terminal set consists exclusively of these direct comparisons. The threshold remains zero and task selection follows the frozen blocks. Cached aliases retain the original computation identity. With K completed pairs, $s_{\mathrm{PAE}} = K^{-1} \sum_{k=1}^{K} [R(S'_t, x_k) - R(S_t, x_k)]$. Binary correctness is the special case $R \in \{0, 1\}$; continuous scores retain their magnitudes.

Extension to native joint evaluation. An interface that provides joint feedback declares an ordered list of evaluation units before reading system outputs. Each unit specifies its task roots, evidence inputs, normalized rubric, and evaluator. The extension executes each declared unit once and retains its paired comparison; every direct task comparison provided by the interface is retained as well. Units providing numerical scores use them directly; language judgments are reserved for the remaining units. The native task interface fixes the evaluation units and their source sets. For example, a consistency requirement across several interaction episodes can supply a joint score without supplying a score for each episode separately. Overlapping requirements then produce overlapping ancestry with incomplete singleton coverage.

For units requiring language judgment, a frozen evaluator J reads the paired evidence packet and the task rubric. Its checkpoint revision, tokenizer, prompt, and execution settings are recorded in the run manifest. Decoding is greedy with a maximum of 128 output tokens. Units are sorted by their predeclared identifiers. The incumbent is presented as System A on odd-indexed units and System B on even-indexed units; the parser restores candidate-minus-incumbent orientation. The instruction is fixed:

> Compare System A and System B using the supplied rubric and only the evidence packet. Assign each system a score between 0 and 1 under that same rubric. For a separable task rubric, use the mean score across the tasks in the packet. For a joint rubric, score the complete packet. Return one JSON object with keys status, score_A, score_B. Use status "compared" with two numeric scores when the packet supports a comparison. Equal scores denote a completed comparison with no difference. Use status "abstain" with two null scores when the evidence cannot support a comparison. Return no additional text.

The parser accepts finite-decimal scores in $[0, 1]$, computes their difference without sign quantization, and excludes abstentions. Incorrect keys, out-of-range values, nonfinite values, and inconsistent status fields are malformed responses; each is excluded without a retry. The log records separate counts for comparisons, abstentions, and malformed responses. These counts describe output coverage. Group weights depend only on ancestry. Selective abstention changes the observed comparisons and can therefore affect support.

A.1.2 ROOT BOUNDARIES AND INVOCATION IDENTITY

A root contains the original task identifier, paired system outputs, evaluator feedback, and execution metadata. Multiple tests and repeated executions of the same task remain attached to that root. Task variants share a task family for data splitting. The evaluation DAG uses only roots from the current acceptance block. Historical task-bearing memories are excluded from the evaluator's inputs. Fixed model checkpoints and generic templates are conditioning metadata; evidence-bearing retrieved content requires a current-block ancestor. All parent edges are logged by an external controller inaccessible to the candidate. Ancestry is propagated through every intermediate node, including unscored nodes.

The controller assigns a fresh invocation identifier before executing each comparison. A separate payload digest covers its inputs, actual parents, model revision, decoding configuration, and complete response. A cache hit retains the identifier and payload of the original computation. A new

execution receives a fresh identifier even when its inputs and output match an earlier execution. Repeated identifiers with different payloads cause DEFER; invocation identifiers establish identity even when texts match. Payload consistency is checked before any parsing exclusion. The DAG must be acyclic, all parents must resolve, and each terminal comparison must have nonempty ancestry inside the declared root block. Failure of any check causes DEFER.

Proposal tasks, acceptance tasks, and final audit tasks belong to disjoint families. Every descendant inherits its roots' split. The proposer can use completed acceptance feedback to form the next candidate, but that candidate is evaluated on a fresh block. The final audit evaluates the checkpoint retained after the last decision. The number of rounds is fixed before the run; no acceptance decision is made from a partial block.

A.1.3 COMPLETE ACCEPTANCE ALGORITHM

Algorithm 1 applies to both protocols. The acquisition schedule determines the terminal set; the aggregation itself generates no model calls. Invalid provenance takes precedence over exclusions. An empty retained set returns DEFER.

Algorithm 1 Provenance-Adjusted Evidence for a fixed candidate

Require: Fixed pair (S, S'), fresh task block, frozen evaluation manifest
Ensure: Next system, status, support, weights, and provenance log

Complete the declared paired evaluations and log their actual evidence parents.
Validate acquisition, root boundaries, DAG consistency, and invocation payloads.
On a validation failure, **return** $(S, \text{DEFER}, \perp, \perp, \log)$.
Propagate ancestry; retain one record per invocation identifier.
Parse and orient score differences; exclude abstentions and malformed responses.
Group comparable records by identical ancestry and compute Equation (2).
If $M = 0$, **return** $(S, \text{DEFER}, \perp, \perp, \log)$.
if every represented root has a singleton group **then**
    Set singleton weights to $1/K_{\cup}$ and all remaining weights to zero.
    Compute s_{PAE} exactly from the recorded scores.
else if the group ancestries are pairwise disjoint **then**
    Set $w^{\star} \leftarrow \mathbf{1}/M$ and compute s_{PAE} exactly.
else
    Compute numerical weights and $\widehat{s}$ by Appendix A.2.
    On a failed solver check, **return** $(S, \text{DEFER}, \perp, \perp, \log)$.
end if
For exact support: ACCEPT if positive, otherwise KEEP.
For numerical support: apply Equation (7).
Return S' on ACCEPT; return S for both remaining statuses, with the log.

A.2 OPTIMIZATION AND FINITE-PRECISION DECISIONS

The reference implementation uses float64 arithmetic with NumPy 2.0.2 and SciPy 1.13.1. The closed forms in Algorithm 1 operate on rational representations of the recorded finite decimals, so an exact zero returns KEEP. General overlap uses SLSQP for both convex problems, with analytic gradients, `ftol`$= 10^{-15}$, and at most 3,000 iterations per solve. These precision settings are fixed before task evaluation.

The primary solve starts from $\mathbf{1}/M$ and returns $\widetilde{w}$. Negative roundoff is clipped and the vector is normalized only when this changes it by at most 10^{-10} in ℓ_{∞}; a larger correction causes DEFER. Set $\widehat{u} = Q^{\top}\widetilde{w}$. To express the secondary equalities, form $B = [\mathbf{1}, Q]^{\top}$ and $b = (1, \widehat{u}^{\top})^{\top}$, so $B \in \mathbb{R}^{(K+1)\times M}$. Pivoted QR of $B^{\top}$ selects independent rows using a relative diagonal threshold 10^{-12}. If the resulting rank is M, the fiber has at most one weight vector and the secondary solve is unnecessary. Otherwise SLSQP minimizes negative entropy on $Bw = b$, starting at $\widetilde{w}$, with bounds $0 \le w_g \le 1$. The entropy gradient evaluates $\log w_g$ using the smallest positive normal float as its floor. This floor protects arithmetic at zero while leaving the entropy objective unchanged.

The final weights must satisfy all original equalities, including the rows omitted by QR, to 10^{-10} in ℓ_∞. Each required solver must return successful termination. All values must be finite. For a simplex-feasible numerical weight $\widehat{w}$, the primary check uses

$$\widehat{c} = \widehat{w}^\top C\widehat{w}, \qquad \gamma(\widehat{w}) = 2\big[\widehat{c} - \min_g (C\widehat{w})_g\big], \qquad 0 \le \widehat{c} - c^\star \le \gamma(\widehat{w}). \tag{9}$$

The bound follows from the first-order convexity inequality minimized over the simplex. The implementation requires $\gamma \le 10^{-8}$ for both returned primary and final weights; computed gaps below -10^{-10} fail the check. Failed feasibility, rank-related residuals, solver termination, and objective checks each return DEFER.

The equality target $\widehat{u}$ approximates $u^\star$. The primary gap bounds objective error, while an ill-conditioned fiber can amplify a small source displacement into a larger weight or support error. Certifying the error in $\widehat{s}$ therefore requires a separate stability bound. Equation (7) sets a fixed numerical resolution for decisions.

A nonunique boundary case. For ancestry sets $\{1,2\}$, $\{2,3\}$, $\{3,4\}$, and $\{4,1\}$, every primary minimizer has the form $w(a) = (a, 1/2 - a, a, 1/2 - a)^\top$, with $0 \le a \le 1/2$. All map to $u^\star = \mathbf{1}/(2\sqrt{2})$ and satisfy $c^\star = 1/2$. For $\bar{z} = (1, -1, 1, -1)^\top$, their support is $4a - 1$. Maximum entropy uniquely gives $a = 1/4$ and exact support zero. Numerical scores of magnitude 10^{-11} all return DEFER under the fixed guard, regardless of sign; the incumbent is retained. The entropy tie-break is essential here because primary minimizers have different supports.

A.3 COMPUTATIONAL COST AND INTERPRETATION

Computation. For a DAG with $|V|$ nodes and $|E|$ edges, dense ancestry propagation costs $O((|V| + |E|)K)$. Constructing C costs $O(M^2K)$ time and $O(M^2)$ storage. A source-space product uses $O(\mathrm{nnz}(Q))$ operations. General solver cost depends on M, K, and precision. The primary protocol needs only identity checks, grouping, and root averaging. Paired execution and native joint evaluation remain part of the total cost; aggregation adds no model calls. A complete-singleton input can discard multi-root groups after validation without evaluating their weights, but already incurred generation costs remain costs.

Comparison operators. The raw artifact mean averages all comparable records with their multiplicities. The deduplicated mean averages unique comparable invocations. Group averaging assigns $1/M$ to every distinct ancestry group. Root averaging first pools singleton records within each represented root, then averages those root means; it uses only singleton feedback. PAE agrees with group averaging on disjoint groups and with root averaging under complete singleton coverage. These are distinct operators outside those regimes.

Interpretation of acceptance. For direct scores on an independently sampled task block, support is the empirical paired task gain; its relation to population gain follows from the sampling model in Appendix D.3. For bundle feedback, support is a provenance-weighted rubric difference. The guard sets numerical resolution, and N_{prov} describes source geometry. Acceptance quality also depends on missing comparisons, shared evaluator biases, and accurate parent logging. Sequential error control requires an additional evaluation stream with explicit null assumptions, as in PACE (Shawn, 2026).

B EXPERIMENTAL DETAILS

We evaluate task performance, acceptance quality, and controlled provenance diagnostics, with supplementary controls in Appendix C.6.

B.1 BENCHMARKS, SCORING, AND DATA SEPARATION

Table 4 separates development sources from held-out evaluation. The suite covers mathematics, code, and instruction following. ARC-Challenge supports comparison with PACE; harder mathematics, recent code problems, and IFBench extend coverage beyond its small-model setting. These tasks supply verifiable task-level scores; overlapping joint feedback is examined in controlled diagnostics.

Table 4: Benchmark suite. Audit sets are inaccessible to proposal generation, calibration, acceptance, and checkpoint selection. Counts are checked against the frozen dataset revision before execution.

Development source	Final audit	Primary metric
GSM8K training split	GSM8K test (1,319); SVAMP (1,000), evaluated as transfer	Normalized final-answer accuracy
MATH training split	MATH-500 (500); AIME 2025 I/II (30), evaluated as transfer	Mathematical answer accuracy
Full MBPP excluding MBPP+ audit IDs	HumanEval+ (164); MBPP+ (378); LiveCodeBench temporal slice	Fraction of programs passing all tests (pass@1)
ARC-Challenge train/validation	ARC-Challenge test (1,172)	Exact answer-option accuracy
IF-RLVR training collection	Official IFBench test split	Prompt-level loose accuracy

Scoring and versions. GSM8K and SVAMP use the benchmark answer normalization (Cobbe et al., 2021; Patel et al., 2021). MATH-500 uses the released 500-problem subset of MATH and a frozen symbolic-equivalence evaluator; AIME 2025 uses exact integer answers, following its use in GEPA (Hendrycks et al., 2021; Agrawal et al., 2026). AIME serves exclusively as a small descriptive transfer probe. We use the official EvalPlus harness for HumanEval+ and MBPP+, reporting augmented-test pass@1 (Liu et al., 2023). Search uses a development pool drawn from full MBPP, excluding every MBPP+ v0.2.0 audit task ID (Austin et al., 2021; Liu et al., 2023). Hidden audit tests and their feedback remain inaccessible during optimization. Each coding problem is one root, regardless of its number of unit tests. Syntax errors, wrong answers, and execution timeouts receive the harness's failure score; an unavailable evaluator triggers Defer as an acquisition failure.

For LiveCodeBench (Jain et al., 2024), we pin `release_v6` and use problems dated January–April 2025, with the exact IDs, full test cases, and harness commit recorded. This evaluates temporal transfer. For IFBench (Pyatkin et al., 2025), the metric is the official prompt-level *loose* accuracy. A prompt, including all its constraints, is one root. The IF-RLVR collection supplies development prompts for optimization with frozen model weights. Official entry points are the MATH-500 dataset card, EvalPlus repository, LiveCodeBench repository, and IFBench repository.

Splits and fresh-task feasibility. We construct family-disjoint sets D_{prop}, D_{cal}, D_{dev}, and D_{audit}. Near-duplicate questions, transformed variants, and all evidence descendants inherit one family assignment. Audit families are excluded from every development split. In non-code domains, eligible development families are split 20/20/60% for proposal, calibration, and acceptance, using a public split seed. For code, the development pool comprises the full MBPP collection after excluding all 378 MBPP+ v0.2.0 audit IDs, leaving 596 eligible tasks. The primary code allocation is 64/80/320 tasks for proposal, calibration, and acceptance, respectively, with family-disjoint assignments. With $K = 16$, the 320 acceptance roots support $T = 20$ rounds and $2KT = 640$ paired task completions per run. Each acceptance block contains previously unused tasks. In every domain, $T = \min\{20, \lfloor |D_{\text{dev}}|/K \rfloor\}$ is fixed before execution; sensitivity experiments with different K use the same feasibility rule and report their actual T and total unique roots.

Calibration is confined to D_{cal}, which is further partitioned by family into signal and quality-label tasks. Learned aggregators are fitted on a fixed bank of 256 calibration candidate pairs, split 192/64 by candidate lineage for fitting and tuning. Their inputs come from signal tasks; quality labels come from the disjoint label tasks. The calibration pairs may share tasks, so task-level dependence is retained within the bank. The same bank and access to its labels are available to all methods requiring calibration. Replay and closed-loop evaluation use separate candidate banks. The proposer may use completed acceptance feedback for the next candidate; future blocks remain inaccessible. Every visited incumbent and candidate can be audited retrospectively after the run. The final checkpoint is the one retained by the last acceptance decision.

B.2 Comparable methods and evaluation tracks

We distinguish three questions: whether a gate makes better decisions on the same evidence, whether an optimizer produces better systems under a common budget, and whether source overlap changes aggregation.

Direct-score acceptance. Every gate receives the same frozen candidate pair and one direct paired score per task, with no auxiliary critique or judge calls. The controls are the unchanged initial system and greedy acceptance on the mean paired task gain, as used in the acceptance comparisons of Shawn (2026). PAE and root averaging share a single row in this regime. For binary task scores we additionally compare the fixed-sample one-sided exact test on discordant pairs and PACE (Shawn, 2026), with $\alpha = 0.05$ and PACE's fixed betting parameter $\lambda = 0.5$. We use the same maximum root budget for every gate and report actual costs for any early stopping in the PACE implementation. PACE uses its own test and conditional null. A separate continuous-score track retains score magnitudes.

Shared-evidence aggregation. In a separate replay study, each frozen pair is evaluated with the same predeclared singleton judge channels. All methods receive an identical response tensor, missingness mask, and calibration bank. Channels are indexed by task position and evaluator checkpoint; their mapping is fixed across calibration and replay. The panel comprises the three backbones below with greedy decoding and the fixed comparison prompt from Appendix A.1.1. Auxiliary judge acquisition is confined to the replay study and charged in full. Mean and majority aggregation are standard controls in the tables of Ising and CARE (Balasubramanian et al., 2026; Zhao et al., 2026); we add Dawid–Skene (Dawid & Skene, 1979), the class-independent and class-dependent Ising variants, and CARE-SVD/CARE-Tensor (ICML 2026). We adapt these judge aggregators to candidate acceptance.

Continuous comparisons use CARE and mean/PAE on normalized score differences. A separate binary track supplies every method the same $b = \mathbf{1}[z > 0]$ representation (mean and PAE aggregate $2b-1$), with completed nonpositive comparisons retained and missing values explicitly masked. This binary adapter is confined to the aggregation comparison; the primary PAE protocol retains continuous scores. Dawid–Skene and Ising are evaluated in that track. Latent fitting, class orientation, and hyperparameter selection use only the calibration bank. CARE's published grids are retained: $\gamma \in \{0.1, 0.2, 0.25, 0.5, 0.75, 1, 2, 3, 5, 7, 10\}$ for SVD and $\gamma, \tau \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$ for Tensor (Zhao et al., 2026). Other estimator options follow a pinned author implementation. Calibration records identify label access, fitted hyperparameters, and fit cost. Commit metrics include every candidate; the records retain output coverage and deferrals. Missing outputs remain masked. An estimator that cannot process a case returns Defer.

Complete optimization systems. GEPA (ICLR 2026) is the recent reflective prompt-search reference (Agrawal et al., 2026). We include MIPROv2 and its no-demonstration variant, and TextGrad, which appear in GEPA's experimental comparisons (Opsahl-Ong et al., 2024; Yuksekgonul et al., 2025). MIPROv2 is the implementation variant associated with the MIPRO paper; its exact DSPy revision is frozen. ANN (Findings of ACL 2026) supplies an additional multi-agent reference on mathematics and code (Ma et al., 2026). We keep its within-instance network construction and textual-update rules, then compare its native cross-instance update procedure with a PAE commit wrapper under identical architecture settings.

The primary system table uses prompt-only optimization with a common single-agent task interface. Demonstrations are disabled in the matched MIPROv2-No-Demos comparison; the demonstration-enabled variant is labeled separately and all demonstration tokens count toward inference cost.

ANN is reported in a separate table because a multi-agent forward pass changes inference capacity. For each optimizer, we compare its native selection with the same proposer plus PAE where an incumbent–candidate commit interface exists. Optimizer-internal search uses D_{prop} and designated calibration data; outer commit evaluations use fresh D_{dev} blocks. Native selectors rerun with this fresh-block interface are labeled protocol adaptations.

Replay versus closed loop. Replay fixes candidates, paired task executions, evidence, and proposal traces before applying any gate. Candidate banks are generated without using a tested gate's decisions, with proposal families and seeds balanced across banks. This identifies differences in acceptance on a common distribution. Closed-loop runs share initial systems, task-block schedules, seeds, and proposal budgets; their later candidates are allowed to diverge after different commits. The gate replay compares the declared operating points on a common candidate bank; PAE's threshold remains zero.

B.3 Acceptance metrics and uncertainty

Let a_j indicate acceptance of frozen pair j, and let $\widehat{\Delta}_j^{\text{audit}}$ be its paired mean gain on the hidden audit tasks, evaluated only after decisions are complete. The false-commit rate (FCR), harmful-commit rate (HCR), and beneficial-update recall are

$$\text{FCR} = \frac{\sum_j a_j \mathbf{1}[\widehat{\Delta}_j^{\text{audit}} \le 0]}{\sum_j a_j}, \qquad \text{HCR}_\delta = \frac{\sum_j a_j \mathbf{1}[\widehat{\Delta}_j^{\text{audit}} < -\delta]}{\sum_j a_j}, \tag{10}$$

$$\text{Recall}_\delta = \frac{\sum_j a_j \mathbf{1}[\widehat{\Delta}_j^{\text{audit}} > \delta]}{\sum_j \mathbf{1}[\widehat{\Delta}_j^{\text{audit}} > \delta]}. \tag{11}$$

The reported commit metrics use $\delta = 0$. Numerator/denominator records accompany the supplementary replay rates; a zero denominator is NA. These metrics are defined by finite-audit gains. Acceptance and deferral are recorded separately. Final audit scores remain separate by benchmark, and gains are measured relative to the same initialization.

The primary inferential contrast is GEPA + PAE versus GEPA on Qwen3-8B for GSM8K, MATH-500, MBPP+, and IFBench. Its effect sizes and paired 95% intervals appear in Table 10; the other task tables report descriptive means. Evaluation uses seeds $\{0, 1, 2, 3, 4\}$. Intervals use 2,000 paired bootstrap replicates over independent runs or candidate-generation sessions and audit task families, retaining method pairing and repeated audit IDs across candidates. Evidence interventions on one candidate stay in the same cluster. The five runs determine between-run precision; repeated artifacts remain clustered, and N_{prov} measures source geometry. The run records retain the individual-seed scores underlying each mean. AIME remains a descriptive 30-problem transfer probe.

Resource reporting focuses on the same primary contrast: task/proposer/judge calls, input and output tokens, and GPU hours per development run. Calibration and all optimization work enter these totals; audit costs are recorded separately. The records also retain unique tasks and failed acquisitions. CPU solver time is reported only for the overlap diagnostic. These measurements report realized resource use under the common budget ceilings.

B.4 Evidence interventions and overlap diagnostics

The task-root sensitivity uses $K \in \{8, 16, 32\}$ on GSM8K with Qwen3-8B and a fixed GEPA proposer, comparing its native selector with PAE under the feasibility rule above. Table 13 reports scores alongside actual rounds, distinct acceptance roots, and task-model calls as K varies.

The fixed-candidate intervention study uses GSM8K and MBPP+ audit gains with Qwen3-8B task outputs. Its banks contain 100 pairs per benchmark, generated independently of the tested gates. Table 11 applies the direct-score gates to the same $K = 16$ root block. Table 12 isolates three changes: cache aliases retain invocation IDs; new interpretations receive new IDs on the same roots; fresh task observations add roots. The multiplicity comparison fixes $K = 16$ and uses the endpoints $L = 1$ and $L = 16$, with uniform and predeclared concentrated quotas. New-root controls compare nested $K = 8 \to 16$ and $16 \to 32$ blocks at $L = 1$. The evaluator, rubric, and acquisition settings stay fixed. Raw artifact, invocation-deduplicated, equal ancestry-group, and singleton root averaging are compared with PAE using Appendix A.3. Only cache replay is required to preserve

PAE exactly; new interpretations may change within-root means. Repeated invocations may be statistically dependent.

Controlled overlap layouts in Table 14 test singleton reduction, source exposure, entropy tie-breaking, and numerical stability. All layouts are constructed, and every supplied singleton is retained. Shuffled or omitted ancestry edges test sensitivity to corrupted inputs; structurally valid but incorrect ancestry can escape consistency checks. The numerical report includes all solver failures and near-zero deferrals.

B.5 Backbones, implementation, and budgets

Qwen2.5-3B-Instruct matches PACE's largest open backbone; Qwen3-8B appears in GEPA and CARE; Llama-3.1-8B-Instruct provides a second model family represented in CARE's judge analysis (Shawn, 2026; Agrawal et al., 2026; Zhao et al., 2026; Qwen et al., 2025; Yang et al., 2025; Grattafiori et al., 2024). The checkpoint identifiers are `Qwen/Qwen2.5-3B-Instruct`, `Qwen/Qwen3-8B`, and `meta-llama/Llama-3.1-8B-Instruct`. All model weights are frozen. The common proposal model is Qwen3-8B, fixed across task backbones and compared methods. Candidate edits modify a persistent template containing reusable instructions only.

Decoding and execution. The primary task protocol uses one greedy completion per task, with Qwen3 thinking disabled, a 16,384-token total context limit, and maximum outputs of 4,096 tokens for mathematics/code and 2,048 for ARC/IFBench. All compared systems use the same task-specific limits and frozen chat templates. Proposals use temperature 0.7, top-$p = 0.95$, and at most 2,048 output tokens. Language-judge diagnostics use the greedy 128-token adapter in Appendix A.1.1. We follow the frozen official code harness's sandbox and timeout policy for every method.

Hardware and resource accounting. The reference hardware configuration comprises two A100 80GB GPUs, at least 16 CPU cores, and 128GB host memory. Inference uses BF16 with one model replica per GPU and a pinned inference-engine build; actual GPU models, driver/CUDA versions, engine version, checkpoint hashes, seeds, and concurrency are recorded before execution. The gate experiment uses at most $T = 20$ proposals and $K = 16$ new paired tasks per decision, hence at most $2TK = 640$ direct task completions, plus proposal generation. We cap proposal work at 64 task completions per run. Auxiliary judge acquisition and fitted-estimator calibration are charged separately and included in total end-to-end cost.

Complete-system comparisons use a common ceiling of 2,000 task-model completions and four million generated tokens per run, stopping when either budget is exhausted; all candidate evaluation, reflection, proposal-model output, demonstrations, and multi-agent calls are included in their respective token/call ledgers. Task-model completions are distinguished from proposal/judge calls. Table 10 gives their actual counts and token costs for the primary contrast, with identical per-instance context caps. All comparisons use these budget definitions, with audits performed after optimization under identical evaluation budgets.

PAE parameters and run records. The acceptance threshold is exactly zero. Closed-form branches use exact arithmetic on recorded finite decimals; the overlap solver uses float64, $\epsilon_{\text{num}} = 10^{-6}$, SLSQP with `ftol`$= 10^{-15}$ and at most 3,000 iterations, and the feasibility/gap checks already specified in Appendix A.2. The reference environment is NumPy 2.0.2 and SciPy 1.13.1. Numerical tolerances are fixed before task evaluation. The entropy tie-break requires no regularization coefficient, and ρ appears only in the perturbation analysis. The run manifest records task-family splits, ordered blocks, all prompts, benchmark/evaluator commits, calibration partitions, candidate lineage, provenance DAGs, immutable invocation IDs, raw outputs, acceptance decisions, audit scores, and cost ledgers.

B.6 Preliminary-study protocol and visualization

B.6.1 Frozen observations and multiplicity layouts

For each of the 240 incumbent–candidate pairs, the candidate, $K = 16$ task executions and root observations, evaluator checkpoint, and rubric remain fixed. Root k has a pre-generated pool of up to 16 fresh interpretations, with candidate-minus-incumbent judgments $z_{k\ell} \in [-1, 1]$. These calls provide potentially dependent interpretations of the same observation.

The atlas uses nested caps $L \in \{1, 2, 4, 8, 16\}$. Each layout selects the first $m_k(L)$ interpretations from each root's pool, where $1 \leq m_k(L) \leq L$, $m_k(1) = 1$, and quotas do not decrease with L. Quotas are sampled before inspecting judgment values. Thus L is a per-root cap; the total number of judgments is $\sum_k m_k(L)$. One layout per pair and cap gives $240 \times 5 = 1{,}200$ condition records. Separately, the phase diagram uses 40 counterfactual layouts per pair at $L = 16$ from the same pools, giving 9,600 layout records. The two collections are analyzed separately.

A separate exact-copy replay check verifies the algebraic sensitivity of an artifact mean to selective duplication. All reported rates and figure panels use fresh interpretations.

B.6.2 Scores, decisions, and descriptive variables

For each layout, the unadjusted score and decision are

$$s_{\mathrm{raw}}(L) = \frac{\sum_{k=1}^{K} \sum_{\ell=1}^{m_k(L)} z_{k\ell}}{\sum_{k=1}^{K} m_k(L)}, \qquad d(L) = \mathbf{1}\{s_{\mathrm{raw}}(L) > 0\}. \tag{12}$$

Thus a zero score retains the incumbent. A flip compares each cap with the baseline: $d(L) \neq d(1)$. A pair counts as ever flipped if this inequality holds at any recorded cap. The baseline margin is $|s_{\mathrm{raw}}(1)|$. All calculations use the unadjusted mean.

The Gini coefficient measures the concentration of judgment counts across roots:

$$G(\mathbf{m}) = \frac{\sum_{k=1}^{K} \sum_{j=1}^{K} |m_k - m_j|}{2K \sum_{k=1}^{K} m_k}. \tag{13}$$

Equal quotas give $G = 0$. Pre-intervention root conflict is

$$C_{\mathrm{root}} = 1 - \frac{\left| \sum_{k=1}^{K} z_{k1} \right|}{\sum_{k=1}^{K} |z_{k1}| + \epsilon}, \qquad \epsilon = 10^{-12}. \tag{14}$$

For nonzero baseline judgments, values near zero indicate aligned support, whereas values near one indicate cancellation of positive and negative support. The all-zero convention also gives one; every observed pair has a nonzero baseline margin. Conflict is fixed before any extra interpretations and is constant across a pair's layouts; Gini can vary with the layout. Both variables are used only to describe the layouts.

Let $\Delta_i^{\mathrm{audit}}$ be the held-out audit gain for pair i, with positive values favoring the candidate. Audit-direction mismatch compares $d_i(L)$ with $\mathbf{1}\{\Delta_i^{\mathrm{audit}} > 0\}$ and excludes pairs with zero audit gain. All 240 gains are nonzero, so all pairs enter every reported mismatch rate. Mismatch measures agreement with the observed audit direction.

B.6.3 Figure construction and interpretation

In Figure 2(a), candidate pairs are sorted in ascending order of held-out audit gain. The five columns show the recorded caps; color encodes $s_{\mathrm{raw}} \in [-1, 1]$, and each diamond marks a decision differing from that pair's $L = 1$ decision. Sorting leaves the computed rates unchanged.

Panel (b) bins all 9,600 layouts by G and C_{root} using a hexagonal grid with resolution $(18, 12)$. Each displayed cell contains at least 12 layouts and shows the arithmetic mean of their binary flip indicators. White cells fall below the occupancy threshold. The color scale runs from 0 to 60%. The regional frequencies in Table 6 use all records satisfying their thresholds, including records in undisplayed cells.

Define the within-root mean as $\bar{z}_k(L) = m_k(L)^{-1} \sum_{\ell=1}^{m_k(L)} z_{k\ell}$ and its weight as $p_k(L) = m_k(L) / \sum_j m_j(L)$. The score change decomposes as

$$s_{\mathrm{raw}}(L) - s_{\mathrm{raw}}(1) = \underbrace{\sum_{k=1}^{K} \left(p_k(L) - \frac{1}{K} \right) z_{k1}}_{\text{reweighting baseline judgments}} + \underbrace{\sum_{k=1}^{K} p_k(L) \big(\bar{z}_k(L) - z_{k1} \big)}_{\text{changes in within-root judgments}}. \tag{15}$$

Both terms can contribute to decision sensitivity under fixed observations. The preliminary study measures their combined effect on unadjusted acceptance; Section 5.4 reports the two components for a separate set of counterfactual layouts.

B.7 Reporting conventions and literature alignment

We compare commit gates, judge aggregators, and complete optimization systems separately. NA denotes an undefined ratio, including a zero denominator.

Evaluation settings. The task-performance tables use the metrics in Appendix B.1: accuracy for mathematics and ARC, pass@1 for code, and prompt-level loose accuracy for IFBench. Their evaluation denominators are 175 for LiveCodeBench and 300 for IFBench. The official IFBench test set contains 300 prompts (Pyatkin et al., 2025); GEPA uses 294 test prompts (Agrawal et al., 2026). LiveCodeBench uses the 175-problem January–April 2025 slice of `release_v6`. These settings apply to Tables 2, 8, and 1.

The direct-feedback experiment uses five closed-loop runs with $K = 16$ and $T = 20$, yielding $2KT = 640$ acceptance task completions per run. Its calls exclude proposals and audits. The replay experiment has 100 frozen candidate pairs per benchmark, 16 singleton roots per pair, and three valid judgments per root. Its binary adapter uses $b = \mathbf{1}[z > 0]$ and averages $2b - 1$, mapping every completed nonpositive comparison to -1; there are no missing judgments in this experiment. The optimizer and ANN comparisons pair each native selector (Own) with PAE under fresh-block adaptation.

Matched optimizer pairs. Table 1 changes the outer commit rule within each proposer, using the same initialization, budget ceilings, and task-block schedule. Trajectories may diverge after different commits. All rows use Qwen3-8B; demonstration-enabled MIPROv2 forms a separate block, with demonstration tokens included in cost. Shared GEPA, TextGrad, and MIPROv2-No-Demos entries reuse the corresponding aggregates in Table 2.

Estimates and supporting counts. Final scores, gains, commit counts, and call counts are means across five runs. In Table 7, FCR, HCR, and recall pool counts across runs; Table 3 pools counts over its fixed replay bank. Pooled rates use summed numerators and denominators, including runs with zero commits. The task and aggregation tables report descriptive aggregates. Table 10 supplies paired 95% intervals for the primary GEPA contrast specified in Appendix B.3; Table 11 supplies the direct-replay numerator/denominator counts. The other comparisons are reported descriptively. All reported commit metrics use $\delta = 0$. Scores and rates use percent units, whereas gains use percentage points. Unless explicitly noted, percentage-point differences are computed from count-based rates before rounding once to two decimal places. Pale blue identifies PAE. Bold and underlined scores mark the best and second-best distinct values according to each metric's direction. Table 1 ranks selectors within each optimizer pair; Table 9 ranks systems within each backbone. Tables 2, 7, and 8 rank methods within each benchmark and backbone; Table 3 additionally separates continuous and binary evidence. Ties share marks, NA is excluded, and commit counts are not ranked.

For the direct-score comparison, the common acceptance budget counts the two task completions for each evaluated incumbent–candidate pair, including completed acceptance evaluations that do not result in a commit. Proposal, calibration, judge acquisition, and audit costs occupy separate ledgers. For system comparisons the final checkpoint is evaluated without audit-based checkpoint selection, and each development domain has its own optimization run. Transfer benchmarks use the final checkpoint optimized on their corresponding development domain. In particular, the GSM8K checkpoint transfers to SVAMP, the MATH checkpoint to AIME 2025, and the MBPP checkpoint to HumanEval+ and LiveCodeBench. Costs are reported separately for each development domain. Per-domain ledgers separate task/proposer/judge calls, input and output tokens, GPU hours, and audit costs. The primary cost comparison is specified in Table 10; CPU aggregation time is confined to the overlap diagnostic.

Differences from published evaluation settings. PACE's Tables 1, 2, and 5 use small audit pools and reused development tasks, including a planted-improvement regime and sampled decoding (Shawn, 2026). Table 7 uses full audit sets and fresh acceptance blocks. Our optimizer comparison follows the structure of GEPA's Tables 1–2. Its Qwen3-8B runs use temperature 0.6; IFBench uses two response stages and AIME averages repeated sampled responses (Agrawal et al., 2026). Its GPT-4.1-Mini results for TextGrad and MIPROv2-No-Demos also use a different backbone. CARE's Tables 1–2 and Ising's Table 1 evaluate quality/preference datasets with different judge panels and target labels (Zhao et al., 2026; Balasubramanian et al., 2026) and report classification accuracy or scoring MAE. ANN's Tables 1–2 and Appendix C use GPT-family models, a 33-problem

HumanEval test subset, and a 39-problem MATH subset (Ma et al., 2026). Our ANN comparison evaluates the full held-out sets in Appendix B.1.

Equivalence and interpretation. In Table 7, complete direct singleton feedback makes PAE exactly the mean paired gain; greedy acceptance and root averaging therefore share PAE's row. Likewise, a system wrapper that already uses that exact mean with the same zero threshold is equivalent to PAE on the same inputs. Differences from an optimizer's own selection procedure measure the effect of replacing that selector with root-mean acceptance. Table 3 fixes candidate banks independently of gate decisions. Its complete, balanced singleton cases also supply an equality check for mean and PAE. Each comparison uses a fixed judge panel and retains all cases, with missingness, channel coverage, fit costs, and multiplicity interventions recorded explicitly. Controlled overlapping-bundle diagnostics remain separate from these task results.

B.8 Decision-shift protocol and diagnostics

Score decomposition. The mechanism study uses a separate set of 9,600 layouts, with 40 layouts for each of 240 incumbent–candidate pairs. For $K = 16$ fixed roots, let $z_{k\ell} \in [-1, 1]$ denote the candidate-minus-incumbent judgment of interpretation ℓ from root k. A layout retains m_k interpretations from that root, assigning weight $p_k = m_k / \sum_j m_j$ and within-root mean $\bar{z}_k = m_k^{-1} \sum_{\ell=1}^{m_k} z_{k\ell}$. Relative to the single-interpretation baseline $s_0 = K^{-1} \sum_k z_{k1}$, the artifact mean $s = \sum_k p_k \bar{z}_k$ satisfies

$$\begin{aligned} s - s_0 &= \Delta_{\text{weight}} + \Delta_{\text{content}}, \\ \Delta_{\text{weight}} = \sum_{k=1}^{K} \left(p_k - \frac{1}{K} \right) z_{k1}, &\qquad \Delta_{\text{content}} = \sum_{k=1}^{K} p_k (\bar{z}_k - z_{k1}). \end{aligned} \tag{16}$$

This is Equation (15) with the two contributions named explicitly. The decomposition first changes the weights of baseline judgments, then changes the judgments at the resulting weights.

Normalization and decision geometry. Normalize both components by the nonzero baseline score:

$$x_{\text{weight}} = \frac{\Delta_{\text{weight}}}{s_0}, \qquad x_{\text{content}} = \frac{\Delta_{\text{content}}}{s_0}. \tag{17}$$

Dividing by s_0 is equivalent to multiplying by $\text{sign}(s_0)/|s_0|$; negative values oppose the baseline decision. Consequently,

$$\begin{aligned} s &= s_0 \big(1 + x_{\text{weight}} + x_{\text{content}}\big), \\ \text{sign}(s) \neq \text{sign}(s_0) &\iff x_{\text{weight}} + x_{\text{content}} < -1. \end{aligned} \tag{18}$$

All baseline and layout scores in this study are nonzero. The line $x_{\text{weight}} + x_{\text{content}} = 0$ denotes zero net shift. Within the flip region, $|x_{\text{weight}}| > |x_{\text{content}}|$ is equivalent to $x_{\text{weight}} < x_{\text{content}}$, so the diagonal separates the two dominance regimes. No flipped layout has equal component magnitudes.

Density visualization. The hexagonal grid counts all layouts on a logarithmic color scale. Contours use Gaussian kernel density estimation with bandwidth factor 0.24. Circle and cross markers overlay the two audit groups at their recorded coordinates. The inset reports dominance among flipped layouts.

B.9 Source-exposure protocol and diagnostics

Provenance layouts and exposure. The study contains 1,600 layouts, each with 16 represented roots and 10–28 distinct, nonempty ancestry groups. Layouts vary in singleton coverage, ancestry-set size, and overlap among roots. Group g has source vector $q_g = \mathbf{1}_{A_g} / \sqrt{|A_g|}$, as in Equation (2). We compare equal-group weights $w^{\text{grp}} = \mathbf{1}/M$ with PAE's minimum-norm projection in Equation (4). The resulting root exposure and normalized share are

$$u(w) = Q^\top w, \qquad \pi(w) = \frac{u(w)}{\mathbf{1}^\top u(w)}. \tag{19}$$

The normalized shares express redistribution relative to total exposure.

Overlap and ranked transfer. We measure overlap by the mean cosine similarity between distinct source vectors:

$$\Omega(Q) = \frac{2}{M(M-1)} \sum_{g<h} q_g^\top q_h. \tag{20}$$

For each layout, roots are sorted by $\pi_k(w^{\mathrm{grp}})$ in ascending order. The same ordering is applied to PAE, so rank r refers to the same root in both methods. The empirical deciles of Ω partition the layouts into ten bins $\mathcal{D}_d$, each with 160 layouts. The mean transfer at rank r in decile d is

$$\mathcal{R}_{d,r} = \frac{K}{|\mathcal{D}_d|} \sum_{Q \in \mathcal{D}_d} \left[\pi_{(r)}(w^\star) - \pi_{(r)}(w^{\mathrm{grp}}) \right]. \tag{21}$$

A transfer of $+1$ is a gain of $1/16$ of the normalized total, and $\sum_r \mathcal{R}_{d,r} = 0$ in every bin. Figure 6 uses a symmetric color scale centered at zero. Dashed contours interpolate nonzero transfer levels; the solid contour separates exposure gain from loss.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 PRELIMINARY-STUDY RESULTS

Table 5 reports the full multiplicity path. Each percentage uses the same denominator of 240 pairs. The number of pairs that ever differ from baseline reaches 36, whereas 31 differ at $L = 16$; five pairs therefore return to their baseline decision by the final cap.

Table 5: Decision changes and audit-direction mismatches across the multiplicity path. All counts are out of 240 pairs; flips are relative to $L = 1$.

Cap L	Flips, n (%)	Ever flipped through L, n	Audit mismatches, n (%)
1	0 (0.0)	0	51 (21.3)
2	11 (4.6)	11	50 (20.8)
4	23 (9.6)	26	50 (20.8)
8	29 (12.1)	32	56 (23.3)
16	31 (12.9)	36	54 (22.5)

Relative to $L = 1$, the layouts at $L = 2, 4, 8, 16$ remove 6, 12, 12, and 14 baseline audit disagreements, respectively, while introducing 5, 11, 17, and 17 new disagreements. Consequently, the lower mismatch count at $L = 2$ reflects a net reduction of one pair.

Table 6 expands the two regions highlighted in the main text into four concentration–conflict combinations. Regional rates pool layout records: each numerator counts flipped layouts and each denominator counts all layouts meeting that row's thresholds. The pair counts identify how many candidate pairs contribute at least one layout; pairs can contribute to multiple concentration regions. These regions form a subset of the 9,600 layouts, excluding intermediate values of G or C_{root}.

Table 6: Descriptive flip frequencies in selected layout regions. Rates pool repeated layouts within candidate pairs.

Concentration	Root conflict	Flips / layouts	Rate (%)	Pairs
$G < 0.15$	$C_{\mathrm{root}} < 0.35$	0 / 287	0.0	53
$G < 0.15$	$C_{\mathrm{root}} \geq 0.80$	79 / 386	20.5	69
$G \geq 0.45$	$C_{\mathrm{root}} < 0.35$	2 / 933	0.2	53
$G \geq 0.45$	$C_{\mathrm{root}} \geq 0.80$	448 / 1,137	39.4	69

Instability also occurs at low concentration: the high-conflict, low-concentration region has a 20.5% flip rate. These frequencies describe associations across 9,600 layouts clustered within candidate pairs.

C.2 SHARED-EVIDENCE RISK AND RECALL

Table 3 holds candidate pairs and judge records fixed. Complete, balanced singleton evidence makes continuous PAE identical to the mean and binary PAE identical to majority vote. CARE-SVD attains the lowest continuous-score FCR and HCR_0, while CARE-Tensor attains the lowest binary rates; both have lower beneficial-update recall than PAE. Thus source accounting does not replace judge-dependence estimation, which can affect risk even with balanced source representation.

C.3 DIRECT-FEEDBACK ACCEPTANCE CONTROL

Table 7 complements the task-performance comparison in Table 1 by reporting commit risk, beneficial-update recall, and final audit gain together. Complete singleton feedback makes PAE identical to the zero-threshold root-mean gate. Fixed-n accepts only 0.4–0.8 updates per run and has no observed false or harmful commits; PACE makes no commits, leaving its risk ratios undefined. PAE accepts 6.2–7.2 updates per run and achieves larger audit gains, but its FCR is 29.03–30.56% and HCR_0 is 19.35–22.22%. These results show a risk–progress trade-off. Recall is computed over the candidates encountered by each gate along its own closed-loop trajectory.

Table 7: **Acceptance quality with direct feedback.** Qwen3-8B: $K = 16$, $T = 20$, five runs, with 640 acceptance task completions per gate per run. Rates are pooled percentages; Δ is audit gain (percentage points), Cmt is commits/run. NA denotes no commits. **Bold**/underline mark best/second-best values within each dataset, including ties; Cmt is unranked.

Dataset	Gate	FCR↓	HCR_0 ↓	$Recall_0$ ↑	Δ ↑	Cmt
GSM8K	Fixed-n paired test (Shawn, 2026)	**0.00**	**0.00**	5.26	+0.23	0.4
	PACE (Shawn, 2026)	NA	NA	0.00	0.00	0.0
	PAE / root mean	29.03	19.35	**56.41**	**+2.68**	6.2
MATH-500	Fixed-n paired test (Shawn, 2026)	**0.00**	**0.00**	8.70	+0.96	0.8
	PACE (Shawn, 2026)	NA	NA	0.00	0.00	0.0
	PAE / root mean	30.56	22.22	**56.82**	**+3.64**	7.2
MBPP+	Fixed-n paired test (Shawn, 2026)	**0.00**	**0.00**	6.98	+0.69	0.6
	PACE (Shawn, 2026)	NA	NA	0.00	0.00	0.0
	PAE / root mean	30.30	21.21	**54.76**	**+3.33**	6.6

C.4 Additional backbones

GEPA + PAE has the highest mean score on all nine benchmarks for both additional backbones (Table 8), exceeding GEPA by 0.34–1.27 percentage points. Figure 4 visualizes the GEPA/PAE pairs on the four primary benchmarks; Table 8 includes all methods and benchmarks. Across the eight primary backbone–benchmark combinations in the figure, gains range from 0.80 to 1.27 points; IFBench gains are largest at +1.20 and +1.27 points for Qwen2.5-3B-Instruct and Llama-3.1-8B-Instruct, respectively. Proposal models, splits, decoding, and budgets match Table 2.

Table 8: **Task performance across model families.** Five-run mean scores (%, ↑); metrics, transfer, and test sizes follow Table 2. †: fresh-block adaptation. Comparators: TextGrad (Yuksekgonul et al., 2025), MIPROv2 (Opsahl-Ong et al., 2024), and GEPA (Agrawal et al., 2026). Bold/underline mark best/second-best scores per row.

Benchmark	Initial	TextGrad†	MIPROv2† No-Demos	GEPA†	**GEPA + PAE (ours)**
Qwen2.5-3B-Instruct					
GSM8K	80.59	82.02	82.38	83.49	**84.32**
SVAMP	86.60	87.94	88.18	89.26	**89.96**
MATH-500	56.40	58.44	58.80	60.32	**61.36**
AIME 2025	10.00	10.67	10.67	12.67	**13.33**
HumanEval+	60.37	62.32	61.95	63.90	**65.12**
MBPP+	61.11	62.96	63.33	64.60	**65.50**
LiveCodeBench	11.43	11.89	12.23	13.03	**13.83**
ARC-Challenge	75.17	76.74	77.01	77.75	**78.52**
IFBench	25.67	27.13	27.73	28.87	**30.07**
Llama-3.1-8B-Instruct					
GSM8K	78.01	79.55	79.77	81.03	**81.83**
SVAMP	84.60	86.02	86.40	87.64	**88.26**
MATH-500	44.80	46.56	46.92	48.44	**49.52**
AIME 2025	6.67	6.67	7.33	8.00	**8.67**
HumanEval+	57.93	59.39	59.63	61.46	**62.68**
MBPP+	59.26	60.79	61.11	62.65	**63.60**
LiveCodeBench	8.57	9.26	9.14	10.29	**10.63**
ARC-Challenge	79.01	80.24	80.55	81.55	**82.22**
IFBench	29.67	30.87	30.60	32.13	**33.40**

C.5 MULTI-AGENT COMPARISON

Table 9 retains ANN's network construction and textual updates (Ma et al., 2026) while changing its outer selector. PAE increases mean scores by 0.67–1.96 percentage points across the 12 backbone–benchmark combinations. Each backbone uses matched initial architectures and budgets, with MATH and MBPP as development tasks.

Table 9: **PAE as an outer gate for ANN.** Five-run mean scores (%, ↑) on full audit sets. All network invocations count toward the budget. †: fresh-block adaptation. Bold/underline rank selectors within each backbone.

System	MATH-500	AIME 2025	HumanEval+	MBPP+
Qwen2.5-3B-Instruct				
ANN† (Ma et al., 2026)	<u>59.60</u>	<u>12.67</u>	<u>64.63</u>	<u>64.81</u>
ANN + PAE (ours)	**61.56**	**13.33**	**65.98**	**66.14**
Qwen3-8B				
ANN† (Ma et al., 2026)	<u>71.16</u>	<u>24.67</u>	<u>74.39</u>	<u>73.12</u>
ANN + PAE (ours)	**72.64**	**26.00**	**75.73**	**74.29**
Llama-3.1-8B-Instruct				
ANN† (Ma et al., 2026)	<u>47.48</u>	<u>8.00</u>	<u>62.44</u>	<u>62.86</u>
ANN + PAE (ours)	**49.20**	**8.67**	**64.15**	**64.18**

C.6 SUPPLEMENTARY CONTROLS AND DIAGNOSTICS

The tables in this section report paired task gains and costs, common-bank gate decisions, evidence interventions, and numerical diagnostics. NA marks an undefined or inapplicable quantity. Pale blue identifies PAE. Bold/underline rank the best/second-best distinct task scores within matched pairs, replay rates within datasets, and structural exposure within layouts; ties share marks. Costs and decision-flip rates are not ranked.

C.6.1 PAIRED TASK GAINS AND ACTUAL COST

The primary contrast uses Qwen3-8B and the five paired GEPA runs. Table 10(a) gives paired intervals for the point estimates in Table 1. Cost totals in panel (b) include proposal, reflection, calibration, and acceptance, with audits separated. The task-model and generated-token ceilings apply to each run.

Table 10: **Primary contrast and resource accounting.** Qwen3-8B, five paired runs. Scores and gains are in percent and percentage points. Intervals use the paired procedure in Appendix B.3. Cost entries are per-run means; tokens are in millions (M), GPU time in hours. Task/prop./judge lists the three call counts separately. Audit GPU time is excluded from optimization GPU time.

(a) GEPA + PAE versus GEPA; point estimates reused from Table 1

Benchmark	GEPA (Agrawal et al., 2026)	**GEPA + PAE**	Δ ↑	95% CI
GSM8K	<u>89.31</u>	**90.14**	+0.83	[+0.31, +1.36]
MATH-500	<u>70.52</u>	**71.48**	+0.96	[+0.29, +1.61]
MBPP+	<u>71.43</u>	**72.43**	+1.01	[+0.39, +1.66]
IFBench	<u>38.40</u>	**39.73**	+1.33	[+0.55, +2.09]

(b) Actual optimization and audit costs

Benchmark	Selector	Calls task/prop./judge	Input (M)	Output (M)	Optim. GPU h	Audit GPU h
GSM8K	Own†	696/20/0	1.84	0.61	3.18	0.42
	PAE (ours)	700/20/0	1.87	0.62	3.24	0.43
MATH-500	Own†	692/20/0	2.21	0.78	3.86	0.24
	PAE (ours)	700/20/0	2.26	0.80	3.94	0.25
MBPP+	Own†	688/20/0	2.36	0.71	4.08	0.31
	PAE (ours)	696/20/0	2.42	0.73	4.16	0.32
IFBench	Own†	684/20/0	1.56	0.46	2.74	0.18
	PAE (ours)	692/20/0	1.61	0.48	2.82	0.18

Paired gains and cost. PAE improves the four primary mean scores by 0.83–1.33 percentage points; each benchmark-wise paired 95% interval excludes zero. The gains accompany 4–8 additional task calls and 0.06–0.08 additional optimization GPU hours per run. Proposal counts are unchanged at 20, and neither selector uses auxiliary judges in this contrast.

C.6.2 GATE DECISIONS ON A COMMON CANDIDATE BANK

Table 11 compares direct-score gates on frozen pairs generated independently of the tested gates, with the same ordered root blocks. All gates face a common candidate distribution; Table 7 evaluates their separate closed-loop trajectories. Pair IDs, generation sessions, and audit-task families determine the pairing. The unchanged-incumbent action supplies the no-update reference; greedy acceptance and PAE share the root-mean row.

Table 11: **Direct-score replay at fixed operating points.** Qwen3-8B; 100 candidate pairs per benchmark, $K = 16$. A/D gives accepted/deferred counts out of 100. Rate cells give a percentage with its numerator/denominator: FCR F/A, HCR_0 H/A, Recall_0 TP/P. P is shared across gates; zero denominators are NA. $\alpha = 0.05$ and PACE $\lambda = 0.5$. Calls/decision includes both task executions and records any early stopping.

Dataset	Gate	A/D	FCR↓	HCR_0 ↓	Recall_0 ↑	Calls/ decision
GSM8K	Fixed-n (Shawn, 2026)	4/0	**0.00** (0/4)	**0.00** (0/4)	<u>9.52</u> (4/42)	32.0
	PACE (Shawn, 2026)	0/0	NA	NA	0.00 (0/42)	29.7
	PAE / root mean	32/0	<u>28.13</u> (9/32)	<u>18.75</u> (6/32)	**54.76** (23/42)	32.0
MBPP+	Fixed-n (Shawn, 2026)	3/0	**0.00** (0/3)	**0.00** (0/3)	<u>7.32</u> (3/41)	32.0
	PACE (Shawn, 2026)	0/0	NA	NA	0.00 (0/41)	30.2
	PAE / root mean	34/0	<u>29.41</u> (10/34)	<u>20.59</u> (7/34)	**58.54** (24/41)	32.0

Commit risk and recall. PAE accepts 32/34 candidates and recovers 23/42 and 24/41 beneficial updates on GSM8K/MBPP+, respectively. Fixed-n accepts only 4/3 candidates, all beneficial; PACE accepts none and therefore has undefined conditional risk. PAE's higher recall accompanies nonzero false and harmful commits. These operating points show a risk–recall trade-off. The mean PACE cost is 29.7/30.2 task calls per decision versus 32 for the other gates.

C.6.3 MULTIPLICITY AND NEW-EVIDENCE CONTROLS

All operators in Table 12 use the same candidate pairs, scores, and intervention layouts. Cache and interpretation controls compare $L = 1$ with $L = 16$ at $K = 16$. Uniform layouts assign 16 records per root; concentrated layouts assign 16 to four roots chosen before reading scores and one to each remaining root. Cache aliases retain the original invocation ID. Fresh interpretations use separate invocations under one fixed acquisition manifest, even when their outputs coincide. New-root controls retain one interpretation per root and use nested task blocks. A flip is a change between accepting the candidate and retaining the incumbent; both Keep and Defer count as retention.

Table 12: **Decision changes under evidence interventions.** Flip rates (%) relative to each paired pre-intervention layout; each rate retains its count over the 100-pair bank in the result records. Raw: artifact mean; Dedup.: invocation-deduplicated mean; Group: equal ancestry-group mean; Root: singleton root mean. Defer retains the incumbent and is logged separately. Exact invariance is required only for cache aliases. Fresh interpretations and fresh roots can change PAE. Flip rates are descriptive.

Dataset	Intervention	Layout	Raw	Dedup.	Group	Root	**PAE**
GSM8K	Cache aliases	Uniform	0	0	0	0	0
	Cache aliases	Concentrated	9	0	0	0	0
	New interpretations	Uniform	7	7	7	7	7
	New interpretations	Concentrated	18	18	9	9	9
	Fresh roots	$K : 8 \to 16$	13	13	13	13	13
	Fresh roots	$K : 16 \to 32$	7	7	7	7	7
MBPP+	Cache aliases	Uniform	0	0	0	0	0
	Cache aliases	Concentrated	11	0	0	0	0
	New interpretations	Uniform	8	8	8	8	8
	New interpretations	Concentrated	20	20	10	10	10
	Fresh roots	$K : 8 \to 16$	14	14	14	14	14
	Fresh roots	$K : 16 \to 32$	9	9	9	9	9

Effects of each intervention. Concentrated cache copies flip 9%/11% of raw-mean decisions on GSM8K/MBPP+, while all four remaining operators are unchanged. With concentrated new interpretations, raw and deduplicated averaging flip 18%/20%, compared with 9%/10% for group/root averaging and PAE. Uniform new interpretations and fresh roots change all five operators equally in these singleton layouts. PAE is invariant to cache copies and coincides with root averaging under these complete singleton layouts. Cache copies incur no new model calls; fresh interpretations and tasks enter acquisition costs.

C.6.4 SENSITIVITY TO THE NUMBER OF TASK ROOTS

Table 13 holds the proposer, task backbone, initialization, and split policy fixed. Each K uses five paired seeds. Within a K setting, selectors share the fresh-block schedule and budget ceilings; later candidates may diverge. U counts unique acceptance tasks and equals KT only when all scheduled rounds finish. All task-model calls, including proposer-side task evaluations, enter the call column.

Table 13: **Task-root sensitivity.** GSM8K, Qwen3-8B, GEPA proposer (Agrawal et al., 2026). T is completed rounds and U unique acceptance roots; costs are per-run means. All runs complete 20 rounds. Scores are five-run means (%); Δ is PAE minus Own with a paired 95% interval in percentage points. The $K = 16$ score/gain entries reuse Table 1; the $K = 8$ and $K = 32$ rows extend the comparison. Own rows have no paired-difference entry.

K	Selector	T/U	Score↑	Δ [95% CI]	Task calls	Output tokens (M)
8	Own†	20/160	<u>88.95</u>	NA	372	0.33
	PAE (ours)	20/160	**89.46**	+0.51 [+0.11, +0.94]	376	0.34
16	Own†	20/320	<u>89.31</u>	NA	696	0.61
	PAE (ours)	20/320	**90.14**	+0.83 [+0.31, +1.36]	700	0.62
32	Own†	20/640	<u>89.58</u>	NA	1328	1.16
	PAE (ours)	20/640	**90.37**	+0.79 [+0.27, +1.32]	1336	1.18

Rounding. At $K = 8$, Own/PAE totals are 5,866/5,900 correct out of 6,595. The displayed gain uses rounded-score subtraction, $89.46 - 88.95 = 0.51$.

Budget interpretation. The schedule satisfies $T_{\max} = \min\{20, \lfloor |D_{\text{dev}}|/K \rfloor\}$ before execution; the common system budget can stop a run earlier. Cross-K changes are accompanied by their actual root counts and costs. Each acceptance block contains previously unused tasks. The displayed PAE gains are 0.51/0.83/0.79 points at $K = 8/16/32$, respectively. Raising K from 16 to 32 increases PAE's score by 0.23 points while increasing task calls from 700 to 1,336, so the absolute gain is accompanied by substantially more evidence and computation. The $K = 16$ scores, interval, calls, and output tokens agree with Table 10.

C.6.5 CONTROLLED OVERLAP AND NUMERICAL VALIDITY

Table 14 uses constructed source layouts and fixed score vectors. Complete coverage contains 16 singleton groups and four additional pairs; the disjoint layout partitions 16 roots into groups of sizes 1, 2, 4, and 9. The incomplete chain uses $\{1\}, \{1, 2\}, \{2, 3\}$; the cycle uses the four ancestry sets in Appendix A.2. Multiplicities and solver starts are fixed independently of comparison scores. Panel (a) measures structural exposure; panels (b–c) assess numerical and provenance sensitivity. All layouts use Algorithm 1 and the guard in Equation (7).

Table 14: **Constructed overlap diagnostics.** (a) Source exposure $c(w) = \|Q^\top w\|_2^2$ (lower is better) on the declared layouts. Root averaging is NA when full singleton coverage is absent. (b) PAE numerical checks: maximum cross-start weight/support spread, gap/residual, failure/Defer counts, and median CPU ms. (c) PAE sensitivity to corrupted ancestry relative to intact inputs; flip rates use accept versus retain.

(a) Structural exposure with identical record inputs

Operator	Complete	Disjoint	Incomplete	Cycle
Raw artifact mean	0.0846	0.2964	0.6552	0.5671
Invocation-deduplicated mean	0.0812	<u>0.2718</u>	0.6257	<u>0.5284</u>
Equal ancestry-group mean	<u>0.0783</u>	**0.2500**	<u>0.6016</u>	**0.5000**
Singleton root mean	**0.0625**	NA	NA	NA
PAE (ours)	**0.0625**	**0.2500**	**0.5000**	**0.5000**

(b) PAE solver and entropy checks

Layout	Starts n	Max spread w / s	Max gap / residual	Failure n/N	Defer n/N	CPU ms
Incomplete chain	16	2.4×10^{-11} 6.1×10^{-12}	3.2×10^{-10} 2.8×10^{-12}	0/64	0/64	1.8
Nonunique cycle	16	7.6×10^{-10} 1.9×10^{-10}	5.4×10^{-9} 8.1×10^{-11}	0/64	1/64	3.4
Near-zero supports	16	3.1×10^{-10} 8.4×10^{-11}	4.8×10^{-9} 6.5×10^{-11}	0/64	58/64	2.9

(c) PAE ancestry-corruption checks on the same cases

Input graph	Cases	Max $\lvert s - s_{\text{intact}}\rvert$	Flips (%)	Defer n/N
Intact	128	0.000	0.00 (0/128)	6/128
Shuffled parent edges	128	0.183	8.59 (11/128)	7/128
Omitted parent edges	128	0.247	12.50 (16/128)	9/128

Numerical reporting. Spreads compare successful solves of identical inputs across starts; gap/residual maxima include all checked values. Support changes require a returned support in both conditions, with this count recorded separately. Flip and Defer rates include all attempts, with Defer retaining the incumbent. The cycle includes $(1, -1, 1, -1)$ with exact-zero support; near-zero cases exercise the 10^{-6} guard.

Observed structural and numerical behavior. PAE matches root/group averaging on complete, disjoint, and cycle layouts, while reducing incomplete-chain exposure from 0.6016 to 0.5000. Maximum gaps/residuals stay below $10^{-8}/10^{-10}$, with no solver failures in 64 cases per family. The guard defers 58/64 near-zero cases. Shuffled/omitted ancestry flip 11/128 and 16/128 decisions: incorrect provenance can affect decisions even when its graph is structurally valid.

C.7 DECISION-SHIFT MECHANISM ANALYSIS

Component dominance and audit direction. Table 15 reports all flipped layouts. An audit-correcting flip changes a baseline mismatch into agreement with the held-out audit direction; an audit-harming flip changes agreement into a mismatch. Frequencies pool the 40 layouts per pair. The weighting-dominated group has median coordinates $(-1.55, -0.17)$, whereas the content-dominated group has medians $(-0.22, -1.32)$, matching the two branches in Figure 5. Reweighting dominates both audit groups at similar rates.

Table 15: **Decision reversals by component dominance and audit direction.** Percentages use the flip count in the same row as the denominator. Overall, 4,905 of 9,600 layouts flip (51.1%).

Flip group	Total	Weighting-dominated	Content-dominated
All flips	4,905	3,500 (71.4%)	1,405 (28.6%)
Audit-correcting	1,338	944 (70.6%)	394 (29.4%)
Audit-harming	3,567	2,556 (71.7%)	1,011 (28.3%)

C.8 SOURCE-EXPOSURE REDISTRIBUTION

Table 16: **Redistribution across overlap deciles.** Each row contains 160 layouts. Mean transfers use the least- and most-exposed roots under equal-group averaging. Interval endpoints are rounded to three decimals.

Decile	Overlap interval	Rank 1 transfer	Rank 16 transfer
1	$[0.000, 0.049)$	$+0.02$	-0.47
2	$[0.049, 0.078)$	$+0.10$	-0.75
3	$[0.078, 0.100)$	$+0.12$	-0.93
4	$[0.100, 0.129)$	$+0.17$	-1.05
5	$[0.129, 0.160)$	$+0.21$	-1.31
6	$[0.160, 0.198)$	$+0.26$	-1.59
7	$[0.198, 0.234)$	$+0.29$	-1.69
8	$[0.234, 0.281)$	$+0.37$	-1.89
9	$[0.281, 0.338)$	$+0.41$	-2.03
10	$[0.338, 0.476]$	$+0.51$	-2.17

Objective and concentration. For $c(w) = \|Q^\top w\|_2^2$, the reduction $100[1 - c(w^\star)/c(w^{\mathrm{grp}})]$ has median 40.2% and interquartile range 27.1–52.4%. Exposure Gini, $G(\pi) = \sum_{i,j} |\pi_i - \pi_j|/(2K)$, decreases in 1,531 of 1,600 layouts (95.7%). In all 320 layouts in the top two overlap deciles, the bottom exposure quartile (ranks 1–4) gains total share and the top quartile (ranks 13–16) loses it. Across all ten deciles, mean transfer increases monotonically at rank 1 and decreases monotonically at rank 16 (Table 16).

D PROOFS

D.1 PROJECTION, LINEAR FIBERS, AND ENTROPY

Lemma 1 (Unique support)**.** *For nonempty source groups, the minimum-norm vector $u^\star$ in Equation* (4) *is unique. Its complete set of minimizing weights is $\{w \in \mathcal{W}_M : Q^\top w = u^\star\}$. Equation* (5) *defines a unique weight vector, independent of the comparison values.*

Proof. The convex hull of finitely many source vectors is nonempty and compact. The squared Euclidean norm is strictly convex in its source-space argument, hence has a unique minimizer $u^\star$ on that hull. Every primary minimizer must map to this vector; every simplex weight mapping to it attains the same minimum. The resulting fiber is compact and convex. Entropy is continuous and strictly concave on the simplex, giving a unique maximizer on the fiber. Both feasible sets depend only on provenance. Permuting records, groups, and root coordinates preserves the corresponding aggregate. □

For $c^\star = \|u^\star\|_2^2$, the projection condition is

$$q_g^\top u^\star \geq c^\star \quad (g = 1, \ldots, M), \qquad w_g^\star > 0 \implies q_g^\top u^\star = c^\star. \tag{22}$$

Indeed, $\langle u^\star, u - u^\star \rangle \geq 0$ on the convex hull is equivalent to checking its vertices. The weighted average of $q_g^\top u^\star$ equals $c^\star$, giving equality on positive weights. Conversely these vertex inequalities certify the projection. The vertex conditions apply even when C is singular.

D.2 STRUCTURAL PROPERTIES

D.2.1 SOURCE PERTURBATIONS AND NORMALIZATION

Proposition 1 (Minimum sensitivity)**.** *For each $\rho \geq 0$, Equation* (4) *minimizes $\sup_{\|e\|_2 \leq \rho} |w^\top Q e|$ over $w \in \mathcal{W}_M$.*

Proof. Cauchy–Schwarz gives $|w^\top Q e| \leq \rho \|Q^\top w\|_2$. The vector $Q^\top w$ is nonzero because the source vectors are nonnegative, nonzero, and convexly combined. For $\rho > 0$, equality holds at $e = \rho Q^\top w / \|Q^\top w\|_2$; for $\rho = 0$, both sides vanish. Minimizing the norm is equivalent to minimizing its square. □

The perturbation class assigns the same self-sensitivity to every source group and couples groups through shared roots. It permits perturbed scores outside $[-1, 1]^M$; restricting the scores to that interval preserves the upper bound but can remove equality. A statistical interpretation follows under an additional model. Under $\bar{z} = \mu \mathbf{1} + Q\xi$, with fixed Q, $\mathbb{E}\xi = 0$, and $\mathrm{Cov}(\xi) = \nu^2 I$, the aggregate is unbiased for μ and has variance $\nu^2 w^\top C w$. PAE weights depend only on ancestry; disjoint groups may still share systematic errors.

Unit normalization deliberately removes source-set size from the diagonal. Without normalization, minimizing $\| \sum_g w_g \mathbf{1}_{A_g} \|_2^2$ would assign less weight to larger disjoint groups. Normalizing each incidence vector to sum to one would assign more weight to those groups. The unit-norm representation makes disjoint groups exchangeable before their scores are read. Interpretations within a group may have dependent errors; their number determines the group mean, while ancestry determines its weight.

D.2.2 COPIES, ABSTENTIONS, AND SOURCE COVERAGE

Proposition 2 (Record invariance)**.** *Adding valid aliases and valid terminal abstentions leaves Q, $\bar{z}$, $w^\star$, and s_{PAE} unchanged. Adding a new comparable judgment $z \in [-1, 1]$ to an existing group with n unique judgments preserves its total weight and changes support by $w_g^\star (z - \bar{z}_g)/(n+1)$, whose magnitude is at most $2w_g^\star/(n+1)$.*

Proof. Aliases are removed by invocation identity and abstentions are excluded before grouping, leaving the comparable records unchanged. A new judgment with an existing ancestry leaves Q

unchanged and replaces that group's mean by $(n\bar{z}_g + z)/(n+1)$. The weights depend only on Q. Substitution gives the stated change; bounded scores give the inequality. □

A comparable equality is different. For groups $\{1\}$ and $\{1,2\}$ with scores $+1, -1$, the optimal weights are $(1/2, 1/2)$ and support is zero. An abstaining record on $\{2\}$ is excluded and changes nothing. A *completed comparison* with equal scores on $\{2\}$ creates a genuine singleton group; the weights become $(1/2, 0, 1/2)$ and support becomes $1/2$. PAE intentionally permits this latter change because the retained evidence has changed.

Proposition 3 (Coverage bounds and singleton reduction)**.** *Write $K_\cup = |\bigcup_g A_g|$, $c^\star = \|u^\star\|_2^2$, and $N_{\mathrm{prov}} = 1/c^\star$. Then $1 \le N_{\mathrm{prov}} \le \min(M, K_\cup)$. If every represented root has a singleton group, the unique primary optimum assigns weight $1/K_\cup$ to each singleton and zero to all multi-root groups.*

Proof. Since $C_{gg} = 1$ and $0 \le C_{gh} \le 1$, every simplex weight satisfies $1/M \le \sum_g w_g^2 \le w^\top C w \le 1$. Moreover, for $u = Q^\top w$, nonnegativity gives $\|u\|_1 = \sum_g w_g \sqrt{|A_g|} \ge 1$. At most $K_\cup$ coordinates are nonzero, hence $\|u\|_2^2 \ge \|u\|_1^2 / K_\cup \ge 1/K_\cup$. Minimization and reciprocation prove the bounds.

Uniform singleton weights attain $c^\star = 1/K_\cup$. Any positive weight on a multi-root group would make $\|u\|_1 > 1$ and violate this optimum. Among singleton weights, equality in Cauchy–Schwarz requires uniform weights. There is one group per ancestry set, so the primary optimum is unique. The entropy stage therefore returns the same unique optimum. □

Proposition 4 (Disjoint addition)**.** *Adding one group disjoint from every existing group increases N_{prov} by exactly one. Its weight is $c^\star/(1+c^\star)$, and the old maximum-entropy weights are scaled by $1/(1+c^\star)$.*

Proof. Let a be the new group's weight and v the normalized old weights. The new objective is $(1-a)^2 v^\top C v + a^2$. Its minimum occurs at an old primary minimizer and $a = c^\star/(1+c^\star)$, with value $c^\star/(1+c^\star)$. The new entropy is $h(a) + (1-a)H(v)$, where h is binary entropy, so its tie rule retains the old entropy maximizer. Reciprocating the new objective gives $N_{\mathrm{prov}} + 1$. □

The diagnostic measures the geometry of represented groups. Adding a new ancestry group on existing roots can increase it without acquiring an observation: appending $\{1\}$ to the sole group $\{1,2\}$ changes N_{prov} from 1 to $2/(1+1/\sqrt{2})$. A single summary has $M = 1$ and $N_{\mathrm{prov}} = 1$ even when it references many roots.

D.3 Evidence information limit

We prove Theorems 1 and 2 and Corollary 1, including the exact finite-sample benchmark, parity-dependent bounds whose ratio tends to one, a second-order budget expansion for the randomized minimax rule, and a finite-sample counterexample for PAE. The proofs build on comparison of experiments (Blackwell, 1953), optimized exponential testing bounds (Chernoff, 1952), and bounded-variable convexity (Hoeffding, 1963).

D.3.1 Experiments, histories, and admissible transcripts

Statistical experiment. Fix the incumbent, candidate, and evaluator before sampling the block. Let $X_1, \dots, X_K$ be independent, identically distributed complete root observations, and let $D_k = d(X_k) = R(S', x_k) - R(S, x_k) \in [-1, 1]$, with $\Delta = \mathbb{E}D_k$. We assume complete valid direct scores almost surely on the unconditionally sampled block. All fixed checkpoints and prior history are conditioning metadata. A transcript A can contain any finite number of derived judgments, summaries, invocation identifiers, and stopping decisions. Its conditional law $\mathsf{T}(\mathrm{d}a \mid X_{1:K})$ is the same under every hypothesis: the transcript uses only the roots and fresh internal randomness. This permits adaptive processing and variable transcript length. Any new task execution enters the root experiment.

Decision risk. For $0 < \gamma < 1$, define $\mathcal{P}_\gamma = \{P \text{ on } [-1,1] : |\mathbb{E}_P D| \geq \gamma\}$. A randomized decision ϕ takes values in $\{-,+\}$, where $+$ accepts and $-$ retains the incumbent. Its worst-case error on direct scores is

$$\mathcal{R}_K(\phi;\gamma) = \sup_{P\in\mathcal{P}_\gamma} \Pr_{P^{\otimes K},\phi}\{\phi \neq \operatorname{sign}(\mathbb{E}_P D)\}. \tag{23}$$

The probability includes the decision's internal randomness. A deferral counts as retention and therefore incurs an error under the positive-gain alternative. Let $p_\gamma = (1+\gamma)/2$, $B \sim \mathrm{Bin}(K, p_\gamma)$, and define

$$\beta_K(\gamma) = \Pr\{B < K/2\} + \tfrac{1}{2}\Pr\{B = K/2\}, \qquad J(\gamma) = -\tfrac{1}{2}\log(1-\gamma^2). \tag{24}$$

The equality event is empty when K is odd; all logarithms are natural.

All measurable spaces are standard Borel, so the conditional distributions used below exist. Under hypothesis $h \in \{-,+\}$, the complete roots have law $P_h^{\otimes K}$. For every measurable transcript set E, the common kernel $\mathsf{T}(E \mid x_{1:K})$ has no remaining dependence on h. This assumes that the roots contain all observations available to the transcript generator, while allowing dependence among derived artifacts.

For example, a finite adaptive pipeline can draw its next artifact using $T_j(\mathrm{d}a_j \mid x_{1:K}, a_{<j})$ and choose whether to stop using the same observed inputs. Adding an absorbing stop symbol puts all such pipelines in the common-kernel model. Variable transcript length is included in the output. Cached copies, deterministic summaries, randomized explanations, and repeated interpretations of fixed roots are all allowed. New external observations and parameter-dependent side information must be included in the root experiment.

For conditional use after a history $\mathcal{F}$, the statements apply separately to any realization of $\mathcal{F}$ for which the candidate is fixed, the fresh roots remain conditionally i.i.d., and the conditional transcript kernel is common to the compared hypotheses. The candidate must therefore be fixed before observing the fresh evaluation block.

D.3.2 Root-based decisions reproduce transcript decisions

Lemma 2 (Exact lifting and information contraction)**.** *Let $X = X_{1:K}$ and let $A \mid X$ have a common kernel T. Every randomized transcript decision has a randomized decision on X with identical error probabilities under both hypotheses. For any prior on the binary hypothesis H, $I(H;A) \leq I(H;X)$.*

Proof. Write $\varphi(a) \in [0,1]$ for the probability of accepting given transcript a. Average this decision over the transcript conditional on the roots:

$$\psi(x) = \int \varphi(a)\,\mathsf{T}(\mathrm{d}a \mid x), \qquad 0 \leq \psi(x) \leq 1. \tag{25}$$

The same function ψ is valid under both hypotheses because T is common. Tonelli's theorem and the conditional sampling construction give the two exact equalities

$$\begin{aligned}
\Pr_+\{\text{retain from } A\} &= \int\!\!\int [1-\varphi(a)]\,\mathsf{T}(\mathrm{d}a \mid x)\,P_+^{\otimes K}(\mathrm{d}x) \\
&= \int [1-\psi(x)]\,P_+^{\otimes K}(\mathrm{d}x) = \Pr_+\{\text{retain from } X\}, \\
\Pr_-\{\text{accept from } A\} &= \int\!\!\int \varphi(a)\,\mathsf{T}(\mathrm{d}a \mid x)\,P_-^{\otimes K}(\mathrm{d}x) \\
&= \int \psi(x)\,P_-^{\otimes K}(\mathrm{d}x) = \Pr_-\{\text{accept from } X\}.
\end{aligned} \tag{26}$$

A root observer with the same computational freedoms can thus reproduce the decision, including its randomization and adaptive transcript length.

An entropy argument proves the information statement while allowing dependence among artifacts. Let $\pi(X) = \Pr\{H = + \mid X\}$ and let $h_2(u) = -u\log u - (1-u)\log(1-u)$, with continuous

endpoint values. Conditional independence of H and A given X implies

$$\begin{aligned}\Pr\{H = + \mid A\} &= \mathbb{E}\big[\mathbb{E}[\mathbf{1}_{\{H=+\}} \mid X, A] \mid A\big] \\ &= \mathbb{E}[\pi(X) \mid A], \qquad h_2''(u) = -\frac{1}{u} - \frac{1}{1-u} < 0 \quad (0 < u < 1).\end{aligned} \tag{27}$$

Conditional Jensen, followed by the tower property, now yields

$$\begin{aligned}H(H \mid A) &= \mathbb{E}\, h_2(\mathbb{E}[\pi(X) \mid A]) \\ &\geq \mathbb{E}[\mathbb{E}[h_2(\pi(X)) \mid A]] = \mathbb{E}h_2(\pi(X)) = H(H \mid X), \\ I(H; A) &= H(H) - H(H \mid A) \leq H(H) - H(H \mid X) = I(H; X).\end{aligned} \tag{28}$$

All entropies involving binary H are finite, including at degenerate priors. □

D.3.3 An exact least-favorable paired-score experiment

Lemma 3 (Exact endpoint testing risk)**.** *Let $D \in \{-1, +1\}$ satisfy $\Pr_h\{D = +1\} = (1 + h\gamma)/2$, identifying h with its sign. Every randomized decision based on K such roots has maximum error at least $\beta_K(\gamma)$. Majority with a fair decision at a tie has both errors exactly equal to $\beta_K(\gamma)$. The same lower bound holds after every common transcript channel.*

Proof. These are genuine paired-score experiments: take two task types indexed by $d \in \{-1, +1\}$ and fixed system scores

$$R(S, x_d) = \frac{1-d}{2}, \qquad R(S', x_d) = \frac{1+d}{2}, \qquad R(S', x_d) - R(S, x_d) = d. \tag{29}$$

Only the frequencies of the two task types change with the hypothesis. Take all remaining root contents to be deterministic functions of d, so D determines the complete root. The two population gains are exactly $\pm\gamma$.

Put $p = (1 + \gamma)/2$, $q = (1 - \gamma)/2$, and let $b(\mathbf{d})$ count the positive coordinates of $\mathbf{d} \in \{-1, +1\}^K$. The product masses and their likelihood ratio are

$$f_+(\mathbf{d}) = p^{b(\mathbf{d})} q^{K - b(\mathbf{d})}, \quad f_-(\mathbf{d}) = q^{b(\mathbf{d})} p^{K - b(\mathbf{d})}, \quad \frac{f_+(\mathbf{d})}{f_-(\mathbf{d})} = \left(\frac{p}{q}\right)^{2b(\mathbf{d}) - K}. \tag{30}$$

For any acceptance probability $\psi(\mathbf{d}) \in [0, 1]$, write its two errors as $e_+(\psi)$ and $e_-(\psi)$. Pointwise minimization gives the full testing lower-bound chain

$$\begin{aligned}\max\{e_+(\psi), e_-(\psi)\} &\geq \frac{e_+(\psi) + e_-(\psi)}{2} \\ &= \frac{1}{2} \sum_{\mathbf{d}} [f_+(\mathbf{d})(1 - \psi(\mathbf{d})) + f_-(\mathbf{d})\psi(\mathbf{d})] \\ &\geq \frac{1}{2} \sum_{\mathbf{d}} \min\{f_+(\mathbf{d}), f_-(\mathbf{d})\} \\ &= \Pr_+\{b < K/2\} + \frac{1}{2} \Pr_+\{b = K/2\} = \beta_K(\gamma).\end{aligned} \tag{31}$$

The penultimate equality pairs each $\mathbf{d}$ with $-\mathbf{d}$. By Equation (30), the minimizing action is acceptance when $b > K/2$ and retention when $b < K/2$. At a tie the two masses coincide, so a fair action attains the same pointwise minimum. Reflection shows that the two errors of this fair-tie rule coincide. Consequently both inequalities in Equation (31) are equalities for that rule, proving the exact minimax value for the endpoint pair. Lemma 2 transfers the lower bound to any transcript-based decision, independently of transcript length. □

D.3.4 Exact minimax risk for arbitrary bounded direct scores

Lemma 4 (A randomized endpoint reduction)**.** *Over all i.i.d. laws on $[-1, 1]$ whose mean has magnitude at least γ, the minimax error of randomized direct-score decisions is exactly $\beta_K(\gamma)$.*

Proof. The lower bound follows from Lemma 3, since its two laws belong to $\mathcal{P}_\gamma$. To prove attainability for the entire class, introduce independent $U_1, \dots, U_K \sim \mathrm{Unif}[0,1]$, independent of the scores, and define an auxiliary rounded sign by

$$Y_k = 2\mathbf{1}_{\{U_k \le (1+D_k)/2\}} - 1, \qquad \Pr\{Y_k = +1\} = \mathbb{E}\frac{1+D_k}{2} = \frac{1+\Delta}{2}. \tag{32}$$

The pairs (D_k, U_k) are independent and identically distributed. Hence the Y_k are i.i.d. signs. Use their majority and a separate fair coin at a tie. Define the decreasing function $g_K(b) = \mathbf{1}_{\{b<K/2\}} + \frac{1}{2}\mathbf{1}_{\{b=K/2\}}$. For $p_1 \le p_2$, couple binomial variables with common independent uniforms V_k:

$$\begin{aligned} B(p_1) = \sum_{k=1}^{K} \mathbf{1}_{\{V_k \le p_1\}} \le \sum_{k=1}^{K} \mathbf{1}_{\{V_k \le p_2\}} = B(p_2), \\ g_K(B(p_2)) \le g_K(B(p_1)), \qquad \mathbb{E}g_K(B(p_2)) \le \mathbb{E}g_K(B(p_1)). \end{aligned} \tag{33}$$

If $\Delta \ge \gamma$, the error of the rounded rule is bounded by

$$\begin{aligned} \Pr_P\{\text{retain}\} &= \mathbb{E}g_K\left(B\left(\frac{1+\Delta}{2}\right)\right) \\ &\le \mathbb{E}g_K\left(B\left(\frac{1+\gamma}{2}\right)\right) = \beta_K(\gamma). \end{aligned} \tag{34}$$

If $\Delta \le -\gamma$, reflect every Y_k. The fair-tie rule is symmetric, so the same calculation bounds the probability of acceptance. Combining both cases with the least-favorable pair gives

$$\beta_K(\gamma) \le \inf_{\phi} \sup_{P \in \mathcal{P}_\gamma} \Pr_P\{\phi \ne \operatorname{sign}\Delta\} \le \sup_{P \in \mathcal{P}_\gamma} \Pr_P\{\phi_{\mathrm{round}} \ne \operatorname{sign}\Delta\} \le \beta_K(\gamma). \tag{35}$$

This proves equality. The rounding and fair-tie coins define the randomized benchmark. PAE uses the original scores and retains the incumbent at a tie. □

D.3.5 The optimized PAE upper bound with exact tie semantics

Lemma 5 (Uniform exponential bound for the root mean)**.** *For independent, identically distributed $D_k \in [-1,1]$ with $|\mathbb{E}D_k| \ge \gamma$, the rule that accepts exactly when $\sum_k D_k > 0$ has error at most $(1-\gamma^2)^{K/2} = e^{-KJ(\gamma)}$.*

Proof. First let $\Delta = \mathbb{E}D_k \ge \gamma$. Each $d \in [-1,1]$ is the convex combination $d = \frac{1+d}{2}(+1) + \frac{1-d}{2}(-1)$. Convexity of the exponential, followed by taking expectations, gives for every $\lambda \ge 0$

$$\begin{aligned} e^{-\lambda d} &\le \frac{1+d}{2}e^{-\lambda} + \frac{1-d}{2}e^{\lambda}, \\ \mathbb{E}e^{-\lambda D_k} &\le \frac{1+\Delta}{2}e^{-\lambda} + \frac{1-\Delta}{2}e^{\lambda} \\ &= \cosh\lambda - \Delta\sinh\lambda \le \cosh\lambda - \gamma\sinh\lambda. \end{aligned} \tag{36}$$

Write $S_K = \sum_{k=1}^{K} D_k$. The positive-gain error includes an exact zero, because PAE retains at a tie. Markov's inequality and independence yield

$$\begin{aligned} \Pr\{\phi_{\mathrm{PAE}} = -\} = \Pr\{S_K \le 0\} &= \Pr\{e^{-\lambda S_K} \ge 1\} \\ &\le \mathbb{E}e^{-\lambda S_K} = \prod_{k=1}^{K} \mathbb{E}e^{-\lambda D_k} \\ &\le (\cosh\lambda - \gamma\sinh\lambda)^K \qquad (\lambda > 0). \end{aligned} \tag{37}$$

We optimize the constant exactly. Let $a = \frac{1}{2}\log[(1+\gamma)/(1-\gamma)] > 0$. Direct substitution and the hyperbolic subtraction identity give

$$\begin{aligned} \cosh a = (1-\gamma^2)^{-1/2}, \qquad \sinh a = \gamma(1-\gamma^2)^{-1/2}, \\ \cosh\lambda - \gamma\sinh\lambda = \sqrt{1-\gamma^2}\cosh(\lambda - a) \ge \sqrt{1-\gamma^2}, \\ \inf_{\lambda \ge 0}(\cosh\lambda - \gamma\sinh\lambda)^K = (1-\gamma^2)^{K/2} = e^{-KJ(\gamma)}. \end{aligned} \tag{38}$$

Equality in the last optimization holds at $\lambda = a$, which is permitted in Equation (37).

For $\Delta \le -\gamma$, apply the same argument to $-D_k$. Here the error is the strict acceptance event, so

$$\begin{aligned}\Pr\{\phi_{\mathrm{PAE}} = +\} &= \Pr\{S_K > 0\} = \Pr\{-S_K < 0\} \\ &\le \Pr\{-S_K \le 0\} \le e^{-KJ(\gamma)}.\end{aligned} \tag{39}$$

This covers both errors under deterministic retention at ties. Finally, the elementary integral bounds

$$\frac{\gamma^2}{2} = \frac{1}{2}\int_0^{\gamma^2} 1\,\mathrm{d}t \le J(\gamma) = \frac{1}{2}\int_0^{\gamma^2} \frac{\mathrm{d}t}{1-t} \le \frac{\gamma^2}{2(1-\gamma^2)} \tag{40}$$

show that $e^{-KJ(\gamma)} \le e^{-K\gamma^2/2}$. The optimized exponent is strictly stronger for every $\gamma \in (0,1)$. □

D.3.6 Finite-sample envelopes and sharp parity-dependent prefactors

We bound the exact benchmark in Equation (24) with elementary, parity-dependent expressions to quantify the exponent and sample-budget results.

Lemma 6 (Affinity bounds and a fair-coin representation)**.** *For $K \ge 1$ and $0 < \gamma < 1$,*

$$\frac{1-\sqrt{1-e^{-2KJ(\gamma)}}}{2} \le \beta_K(\gamma) \le \frac{1}{2}e^{-KJ(\gamma)}. \tag{41}$$

If $Z_1, \dots, Z_K$ are independent fair signs and $a = \operatorname{arctanh}\gamma$, then

$$\beta_K(\gamma) = \frac{1}{2}e^{-KJ(\gamma)}\mathbb{E}\exp\left(-a\left|\sum_{k=1}^{K} Z_k\right|\right). \tag{42}$$

Proof. Use the product masses f_+, f_- in Equation (30). Their affinity $\mathcal{A}_K$ factors coordinatewise:

$$\begin{aligned}\mathcal{A}_K &= \sum_{\mathbf{d}} \sqrt{f_+(\mathbf{d}) f_-(\mathbf{d})} = \prod_{k=1}^{K} (\sqrt{pq} + \sqrt{qp}) \\ &= (2\sqrt{pq})^K = (1-\gamma^2)^{K/2} = e^{-KJ(\gamma)}.\end{aligned} \tag{43}$$

Let $\mathrm{TV}_K = \frac{1}{2}\sum_{\mathbf{d}} |f_+(\mathbf{d}) - f_-(\mathbf{d})|$. Factoring the difference into square roots and applying Cauchy–Schwarz yields

$$\begin{aligned}\mathrm{TV}_K &= \frac{1}{2}\sum_{\mathbf{d}} \left|\sqrt{f_+(\mathbf{d})} - \sqrt{f_-(\mathbf{d})}\right| \left(\sqrt{f_+(\mathbf{d})} + \sqrt{f_-(\mathbf{d})}\right) \\ &\le \frac{1}{2}\left[\sum_{\mathbf{d}} (\sqrt{f_+} - \sqrt{f_-})^2\right]^{1/2} \left[\sum_{\mathbf{d}} (\sqrt{f_+} + \sqrt{f_-})^2\right]^{1/2} \\ &= \frac{1}{2}\sqrt{(2-2\mathcal{A}_K)(2+2\mathcal{A}_K)} = \sqrt{1-\mathcal{A}_K^2}.\end{aligned} \tag{44}$$

The pointwise identity $\min(u,v) = (u+v-|u-v|)/2$, Lemma 3, and $\min(u,v) \le \sqrt{uv}$ therefore imply

$$\begin{aligned}\beta_K(\gamma) &= \frac{1}{2}\sum_{\mathbf{d}} \min(f_+, f_-) = \frac{1-\mathrm{TV}_K}{2} \ge \frac{1-\sqrt{1-\mathcal{A}_K^2}}{2}, \\ \beta_K(\gamma) &\le \frac{1}{2}\sum_{\mathbf{d}} \sqrt{f_+ f_-} = \frac{\mathcal{A}_K}{2}.\end{aligned} \tag{45}$$

This proves Equation (41). To obtain the representation, write $s(\mathbf{d}) = \sum_k d_k$. Then

$$\begin{aligned}f_+(\mathbf{d}) = 2^{-K}e^{-KJ(\gamma)}e^{as(\mathbf{d})}, \qquad f_-(\mathbf{d}) = 2^{-K}e^{-KJ(\gamma)}e^{-as(\mathbf{d})}, \\ \min\{f_+(\mathbf{d}), f_-(\mathbf{d})\} = 2^{-K}e^{-KJ(\gamma)}e^{-a|s(\mathbf{d})|}.\end{aligned} \tag{46}$$

Summing the last expression and multiplying by $1/2$ is precisely expectation under K fair signs, proving Equation (42). □

Lemma 7 (Central binomial coefficients with explicit constants). *For $m \geq 1$, let $a_m = 4^{-m}\binom{2m}{m}$. Then*

$$\frac{1}{\sqrt{\pi(m+1/2)}} \leq a_m \leq \frac{1}{\sqrt{\pi m}}, \qquad \lim_{m\to\infty} \sqrt{\pi m}\, a_m = 1. \tag{47}$$

Proof. Set $I_n = \int_0^{\pi/2} \sin^n t \, \mathrm{d}t$. Integration by parts, with vanishing boundary term for $n \geq 2$, gives

$$\begin{aligned} I_n &= (n-1)\int_0^{\pi/2} \sin^{n-2} t \cos^2 t \, \mathrm{d}t \\ &= (n-1)(I_{n-2} - I_n), \qquad I_n = \frac{n-1}{n} I_{n-2}, \qquad I_0 = \frac{\pi}{2}, \quad I_1 = 1. \end{aligned} \tag{48}$$

Iterating the recurrence separately on even and odd indices yields

$$I_{2m} = \frac{\pi}{2} a_m, \qquad I_{2m-1} = \frac{1}{2m a_m}, \qquad I_{2m+1} = \frac{1}{(2m+1)a_m}. \tag{49}$$

Because $0 \leq \sin t \leq 1$, the integral sequence decreases. Substitute Equation (49) into that monotonicity to obtain the chain

$$\begin{aligned} &\frac{1}{(2m+1)a_m} = I_{2m+1} \leq I_{2m} = \frac{\pi}{2} a_m \leq I_{2m-1} = \frac{1}{2m a_m}, \\ &\frac{1}{\pi(m+1/2)} \leq a_m^2 \leq \frac{1}{\pi m}, \\ &\sqrt{\frac{m}{m+1/2}} \leq \sqrt{\pi m}\, a_m \leq 1. \end{aligned} \tag{50}$$

Taking square roots proves the finite bounds, and squeezing proves the limit. □

Proposition 5 (Finite-sample envelopes with matching prefactors). *Let $p = (1+\gamma)/2$, $q = (1-\gamma)/2$, and $J = J(\gamma)$. For even $K = 2m \geq 2$, define $H_K = a_m e^{-KJ}/(2\gamma)$ and $\eta_K = (1-\gamma^2)/[\gamma^2(K+2)]$. For odd $K = 2m+1 \geq 1$, define $H_K = a_{m+1}\sqrt{1-\gamma^2}\, e^{-KJ}/(2\gamma)$ and $\eta_K = (1-\gamma^2)/[\gamma^2(K+3)]$. Then*

$$\begin{aligned} &\max\{L_K, (1-\eta_K)H_K\} \leq \beta_K(\gamma) \leq \min\{H_K, \tfrac{1}{2}e^{-KJ}\}, \\ &L_K = \begin{cases} \frac{1}{2} a_m e^{-KJ}, & K = 2m, \\ a_{m+1}\sqrt{q/p}\, e^{-KJ}, & K = 2m+1. \end{cases} \end{aligned} \tag{51}$$

In particular, at fixed γ, the following equivalent forms hold along the respective parity subsequences:

$$\beta_K(\gamma) \sim \begin{cases} \dfrac{e^{-KJ(\gamma)}}{\gamma\sqrt{2\pi K}}, & K \text{ even}, \\ \dfrac{\sqrt{1-\gamma^2}\, e^{-KJ(\gamma)}}{\gamma\sqrt{2\pi K}}, & K \text{ odd}. \end{cases} \tag{52}$$

Proof. Put $r = q/p \in (0,1)$ and $t_b = \binom{K}{b} p^b q^{K-b}$. We retain the exact combinatorial ratios before bounding the tails.

Even sample size. For $K = 2m$, the central mass is $t_m = a_m e^{-KJ}$. Define $R_j^{\mathrm{e}} = \binom{2m}{m-j}/\binom{2m}{m}$ for $0 \leq j \leq m$ and set it to zero for $j > m$. The exact tail expression is

$$\beta_K(\gamma) = t_m \left[\frac{1}{2} + \sum_{j=1}^{\infty} R_j^{\mathrm{e}} r^j\right], \qquad R_j^{\mathrm{e}} = \prod_{s=0}^{j-1} \frac{m-s}{m+s+1} \quad (1 \leq j \leq m). \tag{53}$$

For numbers $u_s \in [0,1]$, the inequality $\prod_s (1-u_s) \geq 1 - \sum_s u_s$ follows by induction: multiplication of an existing lower bound by $1-u \geq 0$ produces the additional nonnegative term $u\sum_s u_s$. Apply

it with $u_s = (2s+1)/(m+s+1)$. For $j \le m$,

$$\begin{aligned} 1 \ge R_j^{\mathrm{e}} &= \prod_{s=0}^{j-1} \left(1 - \frac{2s+1}{m+s+1}\right) \\ &\ge 1 - \sum_{s=0}^{j-1} \frac{2s+1}{m+s+1} \ge 1 - \frac{1}{m+1}\sum_{s=0}^{j-1}(2s+1) = 1 - \frac{j^2}{m+1}. \end{aligned} \tag{54}$$

For $j > m$, the rightmost expression is nonpositive, so the same lower bound holds for the zero extension. The geometric-series derivatives give

$$\sum_{j=1}^{\infty} r^j = \frac{r}{1-r}, \qquad \sum_{j=1}^{\infty} j^2 r^j = \frac{r(1+r)}{(1-r)^3}, \qquad \sum_{j=0}^{\infty} j(j+1) r^j = \frac{2r}{(1-r)^3}. \tag{55}$$

These identities follow by applying $r\,\mathrm{d}/\mathrm{d}r$ to $\sum_{j\ge 0} r^j = (1-r)^{-1}$ once and twice and adding the appropriate derivatives; termwise differentiation is valid on every compact subinterval of $(0,1)$. Substitute Equation (54) into Equation (53) to obtain

$$\begin{aligned} \frac{\beta_K}{t_m} &\le \frac{1}{2} + \sum_{j\ge 1} r^j = \frac{1+r}{2(1-r)} = \frac{1}{2\gamma}, \\ \frac{\beta_K}{t_m} &\ge \frac{1}{2} + \sum_{j\ge 1} \left(1 - \frac{j^2}{m+1}\right) r^j \\ &= \frac{1}{2\gamma} - \frac{r(1+r)}{(m+1)(1-r)^3} \\ &= \frac{1}{2\gamma}\left[1 - \frac{1-\gamma^2}{\gamma^2(K+2)}\right]. \end{aligned} \tag{56}$$

Keeping just half the central mass also gives $\beta_K \ge t_m/2 = L_K$. This remains useful when the lower bracket in Equation (56) is negative.

Odd sample size. For $K = 2m+1$, there is no tie. The largest mass in the lower tail is

$$t_m = \binom{2m+1}{m} p^m q^{m+1} = a_{m+1}\sqrt{q/p}\, e^{-KJ}, \qquad 2^{-(2m+1)}\binom{2m+1}{m} = a_{m+1}. \tag{57}$$

Set $R_j^{\mathrm{o}} = \binom{2m+1}{m-j}/\binom{2m+1}{m}$ for $0 \le j \le m$ and zero otherwise. The same product argument now gives, for $j \le m$,

$$\begin{aligned} 1 \ge R_j^{\mathrm{o}} &= \prod_{s=0}^{j-1} \left(1 - \frac{2s+2}{m+s+2}\right) \\ &\ge 1 - \sum_{s=0}^{j-1} \frac{2s+2}{m+s+2} \ge 1 - \frac{j(j+1)}{m+2}. \end{aligned} \tag{58}$$

For $j > m$, the rightmost term is again nonpositive, including $m = 0$, so the bound extends to all $j \ge 0$. Consequently

$$\begin{aligned} \frac{\beta_K}{t_m} &= \sum_{j\ge 0} R_j^{\mathrm{o}} r^j \le \sum_{j\ge 0} r^j = \frac{1}{1-r}, \\ \frac{\beta_K}{t_m} &\ge \sum_{j\ge 0} \left(1 - \frac{j(j+1)}{m+2}\right) r^j \\ &= \frac{1}{1-r} - \frac{2r}{(m+2)(1-r)^3} \\ &= \frac{1}{1-r}\left[1 - \frac{1-\gamma^2}{\gamma^2(K+3)}\right]. \end{aligned} \tag{59}$$

Since $\sqrt{q/p}/(1-r) = \sqrt{1-\gamma^2}/(2\gamma)$, the right upper envelope is exactly H_K. The boundary mass alone gives $\beta_K \ge t_m = L_K$. Combining both parity calculations with Lemma 6 proves Equation (51).

Finally, for fixed γ, $\eta_K \to 0$. Thus $(1-\eta_K) \le \beta_K/H_K \le 1$ for all sufficiently large K. Lemma 7 then supplies

$$\begin{aligned} K=2m: &\quad \frac{\beta_K}{e^{-KJ}/(\gamma\sqrt{2\pi K})} = \frac{\beta_K}{H_K}\, a_m\sqrt{\pi m} \longrightarrow 1, \\ K=2m+1: &\quad \frac{\beta_K}{\sqrt{1-\gamma^2}\, e^{-KJ}/(\gamma\sqrt{2\pi K})} = \frac{\beta_K}{H_K}\, a_{m+1}\sqrt{\frac{\pi K}{2}} \longrightarrow 1. \end{aligned} \tag{60}$$

The last limit uses $K/[2(m+1)] \to 1$. These finite bounds establish both sharp parity prefactors. □

D.3.7 Exact finite-sample refinement and the optimal exponent

The following refinement separates the exact randomized benchmark from the actual PAE rule.

Theorem 2 (Root information limit and optimal error exponent)**.** *For every $K \ge 1$ and $0 < \gamma < 1$, the following statements hold.*

(i) ***Information bound for derived evidence.*** *For any binary hypothesis H and any admissible transcript, $I(H;A) \le I(H;X_{1:K})$. Every transcript-based decision has a randomized root-based decision with identical errors. Moreover, there are two valid paired-score experiments with gains $+\gamma$ and $-\gamma$ under which every transcript-based decision has maximum error at least $\beta_K(\gamma)$, regardless of its number of derived artifacts.*

(ii) ***Exact direct-score benchmark.*** *Over all randomized rules observing $D_{1:K}$,*

$$\inf_{\phi} \mathcal{R}_K(\phi;\gamma) = \beta_K(\gamma). \tag{61}$$

(iii) ***PAE attains the optimal exponent.*** *The primary rule $\phi_{\mathrm{PAE}} = +$ exactly when $\sum_{k=1}^{K} D_k > 0$ satisfies*

$$\beta_K(\gamma) \le \mathcal{R}_K(\phi_{\mathrm{PAE}};\gamma) \le \exp[-KJ(\gamma)] \le \exp(-K\gamma^2/2), \tag{62}$$

and, for every fixed $\gamma \in (0,1)$,

$$\lim_{K\to\infty} -\frac{1}{K}\log\beta_K(\gamma) = \lim_{K\to\infty} -\frac{1}{K}\log\mathcal{R}_K(\phi_{\mathrm{PAE}};\gamma) = J(\gamma). \tag{63}$$

Complete proof of Theorem 2. Part (i) follows from Lemmas 2 and 3. In the least-favorable pair, the binary task type determines the complete root, including all additional contents. The transcript lower bound therefore applies to the full observations in the theorem. Part (ii) follows from the two inequalities in Lemma 4.

For part (iii), PAE is one of the permissible direct-score rules. Lemmas 4 and 5 therefore imply every inequality in Equation (62). To verify the claimed limit directly, define

$$c_\gamma = \min\left\{\frac{1}{\sqrt{2\pi}}, \sqrt{\frac{2}{\pi}}\sqrt{\frac{1-\gamma}{1+\gamma}}\right\} > 0. \tag{64}$$

The central lower terms L_K in Proposition 5 and the lower Wallis bound imply, for all $K \ge 1$,

$$\beta_K(\gamma) \ge \frac{c_\gamma}{\sqrt{K+2}} e^{-KJ(\gamma)}. \tag{65}$$

Indeed, the parity-specific estimates are

$$\begin{aligned} K \text{ even}: &\quad L_K \ge \frac{e^{-KJ}}{\sqrt{2\pi(K+1)}}, \\ K \text{ odd}: &\quad L_K \ge \sqrt{\frac{2}{\pi}}\sqrt{\frac{1-\gamma}{1+\gamma}}\,\frac{e^{-KJ}}{\sqrt{K+2}}. \end{aligned} \tag{66}$$

Both dominate Equation (65). Taking logarithms, the full squeeze for PAE is

$$\begin{aligned} J(\gamma) &\le -\frac{1}{K}\log\mathcal{R}_K(\phi_{\mathrm{PAE}};\gamma) \le -\frac{1}{K}\log\beta_K(\gamma) \\ &\le J(\gamma) + \frac{\log(K+2)}{2K} - \frac{\log c_\gamma}{K} \longrightarrow J(\gamma). \end{aligned} \tag{67}$$

This proves both limits and completes the theorem. Proposition 5 gives the stronger multiplicative asymptotics for β_K itself. □

Complete proof of Theorem 1. Lemma 2 gives the claimed exact decision lifting. Choose c_γ as in Equation (64). Combining Equations (65) and (62) proves Equation (8). On the admissible endpoint pair, Lemma 3 lower-bounds every artifact-based rule's maximum error by $\beta_K(\gamma)$, which is at least the same polynomial-prefactor bound. Finally, Equation (67) proves that PAE attains exponent $J(\gamma)$, while the endpoint lower bound prevents any artifact-based rule from attaining a larger uniform exponent. This proves all claims in the main-text theorem. □

D.3.8 ROOT-BUDGET INVERSION AND SECOND-ORDER REFINEMENT

Corollary 1 (Necessary and sufficient root budgets)**.** *Let $0 < \delta < 1/2$. Any admissible procedure that controls both errors by δ uniformly over the experiments above must satisfy the exact condition $\beta_K(\gamma) \le \delta$, hence necessarily*

$$K \ge \frac{\log\big(1/[4\delta(1-\delta)]\big)}{2J(\gamma)}. \tag{68}$$

Direct-score PAE has error at most δ whenever

$$K \ge \left\lceil \frac{\log(1/\delta)}{J(\gamma)} \right\rceil. \tag{69}$$

For fixed γ, both the optimal direct-score budget and the smallest budget sufficient for PAE's actual worst-case risk are asymptotic to $J(\gamma)^{-1}\log(1/\delta)$ as $\delta \downarrow 0$. In the iterated limit $\delta \downarrow 0$ followed by $\gamma \downarrow 0$, the leading constant is $2/\gamma^2$.

Complete proof of Corollary 1. If a procedure has both errors at most δ on every admissible experiment, it does so on the endpoint pair. Lemma 3 forces $\beta_K(\gamma) \le \delta$. By Lemma 6, and using $1 - 2\delta > 0$, this yields the sequence of implications

$$\begin{aligned} \delta \ge \beta_K(\gamma) &\ge \frac{1-\sqrt{1-e^{-2KJ}}}{2}, \\ 1-2\delta &\le \sqrt{1-e^{-2KJ}}, \\ (1-2\delta)^2 &\le 1-e^{-2KJ}, \\ e^{-2KJ} &\le 4\delta(1-\delta), \\ K &\ge \frac{\log(1/[4\delta(1-\delta)])}{2J}. \end{aligned} \tag{70}$$

All quantities being squared are nonnegative, so that step preserves the implication. Conversely, the stated sufficient integer budget ensures

$$\mathcal{R}_K(\phi_{\mathrm{PAE}};\gamma) \le e^{-KJ} \le \exp\left[-J\left\lceil \frac{\log(1/\delta)}{J}\right\rceil\right] \le \delta. \tag{71}$$

This proves the nonasymptotic assertions.

For the asymptotic claim, define the two minimal integer budgets

$$K_\star(\gamma,\delta) = \min\{K \ge 1 : \beta_K(\gamma) \le \delta\}, \qquad K_{\mathrm{PAE}}(\gamma,\delta) = \min\{K \ge 1 : \mathcal{R}_K(\phi_{\mathrm{PAE}};\gamma) \le \delta\}. \tag{72}$$

Both exist by the exponential upper bounds. Let $L = \log(1/\delta)$ and fix $\gamma \in (0,1)$. The minimax lower bound and PAE sufficiency give

$$K_\star \le K_{\mathrm{PAE}} \le \lceil L/J \rceil. \tag{73}$$

Apply Equation (65) at $K_\star$, and then the upper budget in Equation (73), to obtain

$$\begin{aligned} e^{-L} = \delta \ge \beta_{K_\star}(\gamma) &\ge \frac{c_\gamma}{\sqrt{K_\star+2}} e^{-K_\star J}, \\ K_\star J &\ge L + \log c_\gamma - \tfrac{1}{2}\log(K_\star+2) \\ &\ge L + \log c_\gamma - \tfrac{1}{2}\log(\lceil L/J\rceil + 2), \\ \frac{1}{J}\left[1 + \frac{\log c_\gamma}{L} - \frac{\log(\lceil L/J\rceil+2)}{2L}\right] &\le \frac{K_\star}{L} \le \frac{K_{\mathrm{PAE}}}{L} \le \frac{\lceil L/J\rceil}{L}. \end{aligned} \tag{74}$$

Both extremes converge to $1/J$. This squeeze proves both first-order limits directly. Finally, Equation (40) implies

$$\frac{1}{2} \le \frac{J(\gamma)}{\gamma^2} \le \frac{1}{2(1-\gamma^2)}, \qquad \lim_{\gamma\downarrow 0}\lim_{\delta\downarrow 0} \frac{\gamma^2 K_\star(\gamma,\delta)}{\log(1/\delta)} = \lim_{\gamma\downarrow 0}\lim_{\delta\downarrow 0} \frac{\gamma^2 K_{\mathrm{PAE}}(\gamma,\delta)}{\log(1/\delta)} = 2. \tag{75}$$

The limit is iterated in the displayed order. □

Proposition 6 (Second-order budget for the minimax benchmark)**.** *For fixed $\gamma \in (0,1)$ and $\delta \downarrow 0$, the optimal randomized direct-score budget satisfies*

$$K_\star(\gamma,\delta) = \frac{\log(1/\delta) - \frac{1}{2}\log\log(1/\delta) + O_\gamma(1)}{J(\gamma)}. \tag{76}$$

This refinement applies to the exact randomized minimax benchmark.

Proof. Proposition 5 implies the existence of constants $0 < c_1(\gamma) \le c_2(\gamma) < \infty$ and $K_0(\gamma)$ such that for every $K \ge K_0$,

$$c_1 K^{-1/2} e^{-KJ} \le \beta_K(\gamma) \le c_2 K^{-1/2} e^{-KJ}. \tag{77}$$

Both parities are covered by taking the smaller and larger of their positive limiting constants, with fixed slack. Put $n = K_\star(\gamma,\delta)$ and $L = \log(1/\delta)$. Since each fixed-K error is positive, $n \to \infty$. For sufficiently small δ, both n and $n-1$ are at least K_0. Minimality gives $\beta_n \le \delta < \beta_{n-1}$, whether or not successive benchmark values tie. Hence

$$\begin{aligned} c_1 n^{-1/2} e^{-nJ} &\le \delta < c_2 (n-1)^{-1/2} e^{-(n-1)J}, \\ L + \log c_1 &\le nJ + \tfrac{1}{2}\log n, \\ (n-1)J + \tfrac{1}{2}\log(n-1) &< L + \log c_2, \\ L + \log c_1 \le nJ + \tfrac{1}{2}\log n &< L + \log c_2 + J + \tfrac{1}{2}\log\frac{n}{n-1}. \end{aligned} \tag{78}$$

The first-order result already proved gives $n/L \to 1/J$, so $\log n = \log L - \log J + o(1)$. The two bounds in Equation (78) therefore imply $nJ = L - \frac{1}{2}\log L + O_\gamma(1)$, as claimed. Bounded parity and integer-rounding effects are contained in the final constant. □

D.3.9 Finite-sample separation, margins, and interpretation

Proposition 7 (Finite-sample separation from minimax risk)**.** *For $K = 1$ and any $\gamma \in (0,1)$, there are valid bounded paired-score laws with mean γ for which the error of deterministic PAE exceeds the minimax benchmark $\beta_1(\gamma) = (1-\gamma)/2$.*

Proof. Choose any $\varepsilon \in (0,1)$ and assign mass $p_\varepsilon = (\gamma+\varepsilon)/(1+\varepsilon)$ to $D = 1$, with the remaining mass at $D = -\varepsilon$. These are valid differences of scores in $[0,1]$, for example through $R = (1-D)/2$ and $R' = (1+D)/2$. Their mean and PAE error satisfy

$$\begin{aligned} \mathbb{E}D &= p_\varepsilon - \varepsilon(1-p_\varepsilon) = (1+\varepsilon)p_\varepsilon - \varepsilon = \gamma, \\ \Pr\{\phi_{\mathrm{PAE}} = -\} &= 1 - p_\varepsilon = \frac{1-\gamma}{1+\varepsilon} > \frac{1-\gamma}{2} = \beta_1(\gamma). \end{aligned} \tag{79}$$

As $\varepsilon \downarrow 0$, this error tends to $1-\gamma$. This separates PAE's finite-sample risk from the minimax benchmark while preserving the optimal exponent in Theorem 2. □

Even on the endpoint pair, deterministic Keep-at-zero is distinct from the fair-tie comparator. If e_+^{PAE} and e_-^{PAE} denote its two errors there, reflection gives the exact relation

$$\begin{aligned} e_+^{\mathrm{PAE}} = \beta_K + \tfrac{1}{2}\Pr_+\{B = K/2\}, \qquad e_-^{\mathrm{PAE}} = \beta_K - \tfrac{1}{2}\Pr_+\{B = K/2\}, \\ \beta_K \le \max(e_+^{\mathrm{PAE}}, e_-^{\mathrm{PAE}}) \le 2\beta_K. \end{aligned} \tag{80}$$

The tie mass vanishes for odd K. For even K, the last inequality follows because β_K already includes half that mass.

The margin restriction is essential. For fixed K, as $\gamma \downarrow 0$, the two endpoint laws coincide and $\beta_K(\gamma) \to 1/2$ by the finite binomial formula and fair-coin symmetry. Therefore no fixed root budget uniformly distinguishes all arbitrarily small positive gains from negative ones with error below $1/2$. Controlling only erroneous acceptance would admit the trivial always-retain procedure; the two-sided decision risk deliberately also counts missed improvements.

The bound characterizes worst-case difficulty for a fixed root experiment. Derived transcripts can still improve restricted evaluators by preserving information from those roots.