CERTILOOP: AUDITING LEARNED AGENT ARTIFACTS UNDER BENEFIT–RISK CONTRACTS

Anonymous authors
Paper under double-blind review

ABSTRACT

An average benchmark gain does not establish that a learned agent artifact should be reused under explicit risk constraints. We introduce CERTILOOP, an audit protocol that compares each complete frozen artifact against its comparator, fixes risk strata from initial task information before execution, nests benefit–risk contracts (benefit-only, whole-scope, and stratum-guarded) on shared intervals so that each refutation is attributable to a specific requirement, and reports every decision as positive, negative, or unresolved. We audit 30 artifacts from ACE, GEPA, and ReasoningBank on fixed AppWorld and τ-Retail rosters, using 8,192 independent paired executions per artifact as reference evidence. Development selection agrees with every resolved benefit label, yet the reference refutes five of the 25 selected artifacts (20%) on risk grounds, three of them only once the stratum conditions are added. All eight judge configurations we evaluate recommend at least one reference-negative artifact, including six that retain every reference-positive artifact, and task-level prediction accuracy does not order the reliability of the resulting recommendation sets.

1 INTRODUCTION

Agent improvement increasingly takes the form of reusable context rather than new model weights. ACE evolves contextual playbooks, GEPA optimizes prompts through reflection on execution feedback, and ReasoningBank distills strategies from successes and failures into a memory store (Zhang et al., 2026; Agrawal et al., 2026; Ouyang et al., 2026). These methods produce artifacts that persist beyond the executions that created them. Once reused, a playbook, prompt, or memory becomes part of the context through which an agent interprets later requests and chooses actions. This raises a question distinct from how to learn an artifact: what evidence justifies reusing it?

Reuse can preserve mistakes as well as successful strategies. Studies of agent memory identify experience-following behavior, error propagation, and misleading replay of apparently successful executions (Xiong et al., 2026). Research on agent misevolution further documents safety degradation during memory accumulation and other forms of adaptation (Shao et al., 2026). These observations motivate evaluating an artifact's effects when reused, rather than relying on the apparent quality of the experiences from which it was learned. An average benchmark gain is relevant evidence of benefit, but it does not establish that the artifact meets the requirements of a constrained reuse decision.

Those requirements can depend on the task. Consider a retail memory that improves routine inquiries but increases unauthorized changes to refund amounts. A whole-scope risk average can dilute this increase when entitlement-sensitive tasks are uncommon. Prompt Risk Control already formalizes the importance of evaluating prompt choices through bounds on risk measures rather than average performance alone (Zollo et al., 2024). We ask whether a complete learned artifact improves utility relative to its comparator while limiting protected-risk increases, both overall and within prespecified task populations.

Automated judges provide useful assessments, but their predictive quality and the decisions made from those predictions are different objects of evaluation. Agentic evaluation, calibration against human judgments, and dependence-aware aggregation improve how outputs are assessed (Zhuge et al., 2025; Polo et al., 2025; Balasubramanian et al., 2026). JudgeBench nevertheless finds that evaluating response correctness remains challenging for strong judges (Tan et al., 2025). Even an

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.

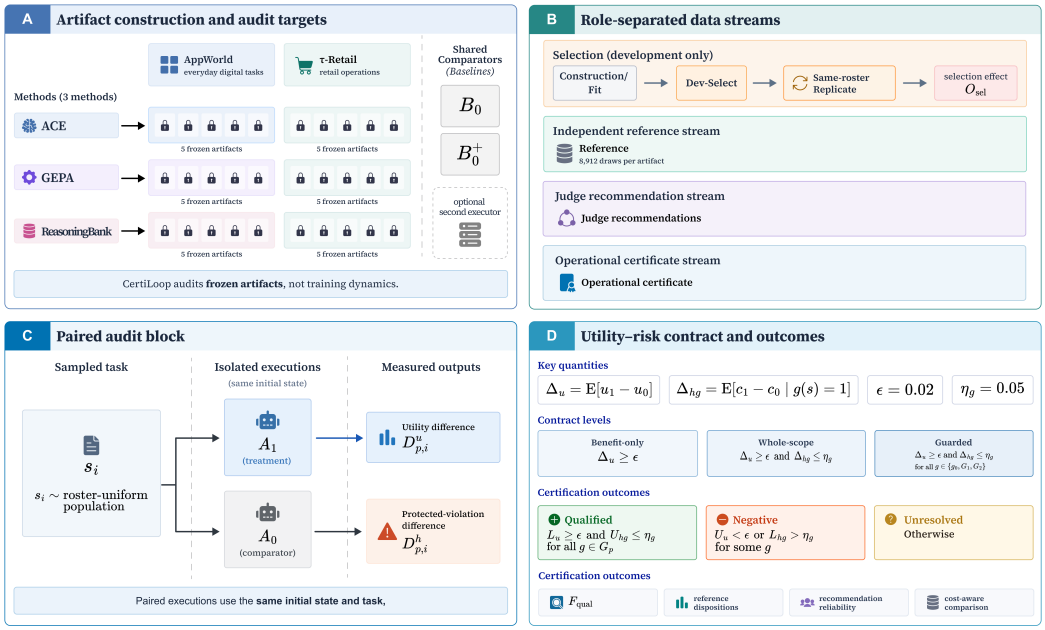


Figure 1: **CertiLoop overview.** (A) Five frozen artifacts per producer and environment are audited exactly as they would be reused. (B) Evidence streams are disjoint: development selection and replication never supply reference evidence, and independent reference labels score development selection and judge recommendations. Sequential certification backends are specified in the appendix. (C) Each draw samples a roster task and executes treatment and comparator from isolated copies of the same initial state, yielding utility and protected-violation differences. (D) Nested contracts map jointly valid intervals to positive, negative, or unresolved decisions.

accurate predictor can make a consequential error near a contract threshold: a small error in one conditional risk estimate may change an artifact's recommendation without substantially changing average prediction error.

This does not make statistical guarantees incompatible with model-assisted evaluation. Selective evaluation can provide guarantees on agreement with human judgments (Jung et al., 2025), and prediction-powered e-values incorporate predictions into valid statistical inference (Csillag et al., 2025). The question is which claim the evidence supports. We introduce CERTILOOP, an audit protocol that turns artifact reuse into an explicit benefit–risk claim and separates proposing a reuse decision from establishing support for it (Figure 1). Its contribution is the audit formulation and decision analysis, not a new concentration inequality or another artifact-generation algorithm.

We audit 30 artifacts from ACE, GEPA, and ReasoningBank on fixed AppWorld and text-based τ-Retail rosters (Trivedi et al., 2024; Barres et al., 2025). The independent reference refutes five of the 25 development-selected artifacts on risk grounds; three of these refutations require the added stratum conditions. All eight tested judge configurations recommend at least one reference-negative artifact. These results show why evidence of improvement and accurate task-level prediction do not, by themselves, establish support for reuse under the registered contract.

The protocol defines each audit by a complete frozen artifact, its comparator, and its task population, with separate roles for development, certification, and reference evidence (§3.1). It combines pre-execution risk strata with nested contracts on shared intervals, making additional refutations attributable to the requirements that were added (Lemma 1). A three-state decision rule distinguishes positive support, refutation, and unresolved evidence; reference-supported brackets then evaluate recommendation sets for both nonqualification and missed qualifying opportunity (§3.3–3.5). The empirical audit examines these decisions and accounts separately for the cost of obtaining evidence

(§4). Its decision-level results concern fixed-endpoint reference audits; sequential certification backends are specified separately and are not validated against reference labels in this study.

2 Related Work

Reusable artifacts and evolving risk. ACE, GEPA, and ReasoningBank learn contextual playbooks, prompt configurations, and reasoning memories, respectively, from execution feedback (Zhang et al., 2026; Agrawal et al., 2026; Ouyang et al., 2026). Shao et al. (2026) identify misevolution across model, memory, tool, and workflow updates, including safety degradation during memory accumulation. This motivates evaluating utility and safety together after adaptation. Whereas these studies develop adaptation mechanisms or characterize risks during evolution, CertiLoop examines a fixed product of that process: whether a complete artifact improves utility without exceeding prespecified risk tolerances relative to its comparator.

Automated evaluation and judge reliability. Agent-as-a-Judge evaluates task-solving processes (Zhuge et al., 2025); Bridge models and corrects human–LLM judgment discrepancies (Polo et al., 2025); and Ising aggregation accounts for dependence between judges (Balasubramanian et al., 2026). The former supplies process-level feedback; the latter two refine scoring and aggregation to improve agreement with reference judgments. These approaches address the quality of predicted assessments. CertiLoop evaluates decisions made with those predictions, comparing recommendation sets with independent reference labels and jointly reporting nonqualification and missed qualifying opportunities. This distinguishes task-level predictive performance from artifact-level contract support.

Safety evaluation and statistical evidence. AgentHarm benchmarks malicious multi-step tool use (Andriushchenko et al., 2025); our protected-risk conditions concern violations within the registered task population. Prediction-powered e-values incorporate predictions into anytime-valid statistical inference (Csillag et al., 2025). CertiLoop places such inference within an audit protocol with separated evidence roles and pre-execution strata. The protected task populations are fixed before outcomes are observed. Nested contracts attribute additional refutations to added requirements, while three-state reporting retains unresolved evidence. The principal experiments evaluate fixed-endpoint reference audits; sequential certification is a complementary implementation.

3 Methodology

CertiLoop audits a frozen agent artifact against a benefit–risk contract on a specified task population. The protocol connects the object of reuse to separate evidence roles, interpretable decisions, and the cost of obtaining those decisions.

3.1 Frozen Artifacts and Evidence Roles

An audit unit is $p = (C, A_1, A_0, \nu)$: task scope, treatment artifact, comparator, and configuration identity. We retain the entire ACE playbook, GEPA prompt configuration, or ReasoningBank memory store with its retrieval rule, against the base context, seed prompt, or agent without learned memory, respectively (Zhang et al., 2026; Agrawal et al., 2026; Ouyang et al., 2026). The audited object is therefore the object that would be reused, rather than a component chosen after seeing its outcomes. Construction and predictor fitting finish before auditing, and artifact updates remain suspended.

Each paired draw samples $J_i \sim \mathrm{Unif}(\mathcal{T}_p)$ with replacement from a locked roster and runs both configurations from isolated copies of the same initial state. Resource limits and task-conditional execution laws are fixed. Averaging fresh executions distinguishes an intervention's expected effect from stochastic output variation, a central concern in language-model auditing (Rauba et al., 2025). Construction, development selection, same-roster replication, certification, and reference evaluation have separate evidence roles. Reference executions remain hidden from operational rules. Task identities may recur, but completed records are not recycled as fresh evidence across roles or later draws; Appendix A.1 specifies permitted within-block shared controls.

3.2 Admissible Evidence and Replay Invariants

Evidence-role separation is part of the inferential contract rather than a bookkeeping convention. Let $\mathcal{F}_t^{\mathrm{cert}}$ denote the information visible to an operational certification backend after t committed blocks. The next unit choice, acquisition probability, betting rule, and stopping decision may depend on $\mathcal{F}_t^{\mathrm{cert}}$ and on current pre-outcome covariates, but not on the current trusted outcome or on any reference-only record. Reference evidence is maintained under a separate access path and is consulted only after the operational decision rule has been frozen. This prevents the evaluator-only reference from becoming an adaptive tuning set and makes the distinction between recommendation, certification, and reference assessment operationally testable.

A record updates an inference stream only when its frozen identities match the registered audit unit: artifact and comparator versions, executor and resource configuration, roster, predicate map, outcome rubric, and evidence role. A mismatch is retained in the audit log as a protocol or infrastructure event, but it cannot tighten a confidence interval. Within a valid paired block, the two arms may share the sampled task identity and validated initial snapshot, while mutable state, memory, caches, and post-execution outputs remain isolated. Repeated roster task identities are permitted because the estimand is the fresh-execution mean on the locked roster; a completed trajectory is never reused as a new draw.

These rules also make missing or inadmissible evidence conservative by construction. A component without justified support retains its full-range interval, so absence of evidence cannot by itself create a positive certificate. Conversely, every reported decision can be replayed from the append-only visible prefix and the registered randomness, without access to future blocks or hidden reference outcomes (Appendix A.6).

3.3 Initial-Information Strata and Nested Contracts

Let $u(Y) \in [0,1]$ denote task utility and $c(Y) \in \{0,1\}$ a protected violation. For initial task information s_i and a frozen predicate g, define

$$\begin{aligned} D^u_{p,i} &= u(Y_{p,i}(1)) - u(Y_{p,i}(0)), \qquad \Delta^u_p = \frac{1}{|\mathcal{T}_p|} \sum_{j \in \mathcal{T}_p} \mathbb{E}[D^u_{p,i} \mid J_i = j], \\ D^h_{p,i} &= c(Y_{p,i}(1)) - c(Y_{p,i}(0)), \quad \Delta^h_{p,g} = \mathbb{E}[D^h_{p,i} \mid g(s_i) = 1]. \end{aligned} \tag{1}$$

The whole-scope predicate is $g_0 \equiv 1$. Each stratum has positive population mass and uses its own conditional draws. Membership is fixed from the initial instruction, metadata, and state before either arm executes. The same mapping serves all artifacts in an environment, and strata may overlap. AppWorld strata identify temporal constraints and permitted persistent state changes; Retail strata identify authorized order-state changes and monetary or entitlement decisions (Table 4). Membership defines a task population, not an outcome: an authorized state change can have $c(Y) = 0$.

For a predicate family $\mathcal{G}_p$ containing g_0, qualification is

$$\mathrm{Qual}(p) \iff \Delta^u_p \geq \epsilon \quad \text{and} \quad \Delta^h_{p,g} \leq \eta_g \quad \forall g \in \mathcal{G}_p. \tag{2}$$

Benefit-only keeps the utility condition, *whole-scope* adds g_0 risk, and *guarded* adds both environment-specific strata. The experiments use $\epsilon = 0.02$ and $\eta_g = 0.05$: practical effect thresholds, distinct from error probabilities. Risk tolerances bound increases relative to the comparator, not absolute violation rates. Explicit tolerances also underlie risk-controlled prompt selection and model routing (Zollo et al., 2024; Overman & Bayati, 2025); here they constrain the paired effects of a complete artifact.

A whole-scope average can dilute a localized increase. For $\pi_g = \Pr(g(s_i) = 1) \in (0,1)$ and complement $\neg g$,

$$\Delta^h_{p,g_0} = \pi_g \, \Delta^h_{p,g} + (1 - \pi_g) \, \Delta^h_{p,\neg g}. \tag{3}$$

If risk is unchanged outside g, then $\pi_g = 0.2$ and $\Delta^h_{p,g} = 0.25$ give $\Delta^h_{p,g_0} = 0.05$. These population values satisfy the whole-scope tolerance while exceeding the stratum tolerance; a positive audit decision additionally requires supporting intervals. At a fixed total of n draws, $\mathbb{E}[n_g] = \pi_g n$. For comparable variances, equal error allocations, and sufficiently large counts, substituting $n_g \approx$

$\pi_g n$ enlarges the empirical-Bernstein variance and range terms by approximately $\pi_g^{-1/2}$ and π_g^{-1}, respectively (Maurer & Pontil, 2009). Actual resolution also depends on the observed variance and distance to each threshold (Appendix A.5).

3.4 Evidence Backends and Three-State Decisions

An evidence backend supplies jointly valid intervals $[L_p^u, U_p^u]$ and $[L_{p,g}^h, U_{p,g}^h]$ on the original difference scale. The common decision map is

$$\mathcal{D}_p(I) = \begin{cases} +, & L_p^u \geq \epsilon \text{ and } U_{p,g}^h \leq \eta_g \quad \forall g, \\ -, & U_p^u < \epsilon \text{ or } L_{p,g}^h > \eta_g \quad \text{for some } g, \\ ?, & \text{otherwise.} \end{cases} \tag{4}$$

Positive requires support for every condition; negative requires one interval-based refutation. Otherwise, the evidence remains unresolved. This logical asymmetry permits different resolution times without ordering them in general. Unsupported streams retain valid full-range intervals; invalid inputs are flagged separately.

Lemma 1 (Nested attribution)**.** *Apply Eq.* (4) *to the same nonempty component intervals and thresholds under the benefit-only, whole-scope, and guarded contracts, with positive sets P_b, P_w, P_g and negative sets N_b, N_w, N_g. Then $P_g \subseteq P_w \subseteq P_b$ and $N_b \subseteq N_w \subseteq N_g$. Every $p \in N_g \setminus N_b$ has $L_{p,g'}^h > \eta_{g'}$ for at least one risk predicate g'. For $p \in N_g \setminus N_w$, at least one such inequality holds in G1 or G2.*

Holding evidence fixed isolates the refutations contributed by added risk conditions. Thus N_b and $N_g \setminus N_b$ partition negatives by their refutation path; a benefit-refuted artifact may also violate risk constraints. Unresolved counts need not be monotone, because added conditions can remove positive support or refute an unresolved artifact. Appendix A.5 proves the lemma, and Section 4.2 applies it.

Proposition 1 (Validity of the audit decision)**.** *If all component intervals jointly cover their targets with probability at least $1-\delta$, then, with that probability, every positive decision satisfies the contract and every negative decision refutes it.*

On the coverage event, Eq. (4) implies the target inequalities in Eq. (2). Allocating component errors with $\sum_j \alpha_j \leq \delta$ establishes this event by a union bound, without cross-component independence (Appendix A.4). Every interval construction used by the decision, including optional armwise or zero-prefix bounds, receives an error share.

Any backend with joint coverage valid for the targets and sampling design can implement this rule. The main reference audit uses fixed-endpoint empirical-Bernstein intervals, each with its own sample count and variance (Maurer & Pontil, 2009). Sequential implementations use time-uniform coverage: DIRECT uses betting confidence sequences (Waudby-Smith & Ramdas, 2024), and PA uses prediction-powered e-values with known positive audit probabilities and nonnegative corrected betting factors (Csillag et al., 2025). Their constructions are given in Appendix A.2; the present experiments do not validate these backends against reference labels.

3.5 Recommendations and Reference-Supported Reuse

A recommendation rule applies a development selection criterion or contract thresholds to predicted artifact-level effects. Agentic evaluation, calibration, and dependence-aware aggregation provide different predictors (Zhuge et al., 2025; Polo et al., 2025; Balasubramanian et al., 2026). Judge benchmarks and studies of correlated errors motivate evaluating these predictions against independent targets rather than equating agreement with correctness (Tan et al., 2025; Kim et al., 2025). Reference labels $R_p = \mathcal{D}_p(I^{\text{ref}})$ use independent trusted executions and their own error allocation; operational statuses $S_{p,t} = \mathcal{D}_p(I_t^{\text{cert}})$ use only evidence available to the backend.

For fixed artifact weights $w_p \geq 0$ summing to one and a recommendation set S, define $P(S) = \sum_{p \in S} w_p \mathbf{1}\{R_p = +\}$, $N(S) = \sum_{p \in S} w_p \mathbf{1}\{R_p = -\}$, and $U(S) = \sum_{p \in S} w_p \mathbf{1}\{R_p = ?\}$, with $W(S) = P(S) + N(S) + U(S) > 0$. On the reference coverage event,

$$\frac{P(S)}{W(S)} \leq F_{\text{qual}}(S) \leq \frac{P(S) + U(S)}{W(S)}, \qquad \frac{N(S)}{W(S)} \leq F_{\text{bad}}(S) \leq \frac{N(S) + U(S)}{W(S)}. \tag{5}$$

Missed qualifying mass lies in $[P(S^c), P(S^c) + U(S^c)]$ on the full-panel scale. Reference-negative recommendations determine the lower nonqualification bound; unresolved artifacts outside the set enlarge the upper bound on missed qualifying mass. Reporting both retains confirmed failures and potential missed opportunities. Section 4.3 applies these summaries with equal artifact weights, separately from task-level prediction metrics. Same-roster replication measures development-selection optimism.

3.6 Audit Records and Evidence Cost

Each decision is linked to its frozen artifact and comparator, population, predicates, outcomes, evidence role, intervals, error allocation, and reason. Sequential records additionally retain acquisition probabilities and per-stream counts so that decisions can be reconstructed from the visible prefix. A changed artifact or comparator defines a new audit unit. Cost is itemized as $C = N_{\text{pair}} c_{\text{pair}} + N_{\text{judge}} c_{\text{judge}} + N_{\text{audit}} c_{\text{audit}} + C_{\text{setup}}$, where setup includes attributable construction, fitting, and calibration. Shared executions are charged once, and reference work is a separate study cost. Prediction assistance is cheaper when avoided audit expense exceeds additional non-audit expense (Appendix A.6). Section 4.4 distinguishes programmatic count–cost profiles from fixed-input human-adjudication accounting.

4 Experiments and Results

We apply the CertiLoop audit protocol to 30 frozen artifacts and ask whether development-side improvement and judge recommendations support reuse under a benefit–risk contract. Three results emerge. Development selection agrees with all 28 resolved benefit labels, yet five of the 25 selected artifacts receive risk-based reference refutations; adding the two strata contributes three additional negative decisions (§4.2). Six of eight judge configurations retain every reference-positive artifact while recommending fewer reference-negative ones, but every configuration recommends at least one (§4.3). In the reported programmatic profiles, neither PA backend is cheaper than DIRECT. Separate fixed-input accounting identifies when adjudication savings offset additional costs (§4.4).

4.1 Evaluation Setup

We evaluate five construction replicas of ACE, GEPA, and ReasoningBank in each of AppWorld and text-based τ-Retail, giving 30 complete artifacts (Trivedi et al., 2024; Barres et al., 2025; Zhang et al., 2026; Agrawal et al., 2026; Ouyang et al., 2026). ACE is evaluated as a full playbook against the base context, GEPA as a prompt configuration against its seed configuration, and ReasoningBank as a memory store and retrieval rule against the same agent without learned memory. Artifacts are built on development-side data, frozen before auditing, and weighted equally.

Each environment uses a locked roster: AppWorld Test-C or held-out τ-Retail tasks. A paired draw samples a task uniformly with replacement and executes treatment and comparator from isolated copies of its initial state, so the target is the roster-conditioned mean $|\mathcal{T}|^{-1} \sum_{j \in \mathcal{T}} \mathbb{E}[D_p \mid J = j]$. Utility $u(Y) \in [0, 1]$ measures task completion, and $c(Y) \in \{0, 1\}$ records a protected violation.

AppWorld G1 covers temporally constrained tasks and G2 authorized persistent state changes; Retail G1 covers authorized order-state changes and G2 monetary or entitlement decisions. Membership is determined from initial task information, is shared across artifacts and both arms, and may overlap (Table 4, Appendix B.3).

Benefit-only requires a utility gain of at least $\epsilon = 0.02$; whole-scope adds an overall risk-difference tolerance of 0.05; guarded applies the same tolerance to G1 and G2 as well. All three contracts use the same intervals and thresholds for shared conditions. Their comparison therefore isolates the decision effect of adding risk requirements (Lemma 1). The independent reference uses 8,192 fresh paired draws per artifact, unavailable to certification. Each artifact has four scalar targets (utility, whole-scope, G1, and G2 risk), giving 120 target intervals, each with error allocation $\alpha_{\text{ref}} = 0.05/318$. For the displayed paired-difference empirical-Bernstein construction (Maurer & Pontil, 2009), these shares sum to $120\alpha_{\text{ref}} \leq 0.05$; Appendix B.1 states the allocation scope. Reference-label conclusions use the joint coverage event and valid outcome measurement.

Dev-Mean selects artifacts with positive development-side utility estimates. Judge rules apply the benefit and risk thresholds to frozen predictions. Both are compared against the guarded reference, retaining positive, negative, and unresolved labels. The judge configurations include agentic evaluation, calibrated judgments, and dependence-aware aggregation (Zhuge et al., 2025; Polo et al., 2025; Balasubramanian et al., 2026). The executor is `Qwen/Qwen3.6-35B-A3B-FP8`; the campaign used about 3,360 allocated GPU-hours on four NVIDIA L40S GPUs (Table 3).

4.2 Risk Constraints Change the Reuse Decision

Table 1: **CertiLoop reference audit under nested contracts.** Each row applies the three-state decision rule to the full panel or its 25 development-selected artifacts, with the same intervals and thresholds for all shared conditions; only the registered requirements change. Brackets are $[R^+/N, (R^+ + R^?)/N]$ on the reference coverage event.

Requirements	Panel	N	$R^+/R^-/R^?$	Qualification (%)
Benefit only	All	30	24/4/2	$[80.0, 86.7]$
Benefit only	Selected	25	24/0/1	$[96.0, 100.0]$
Whole-scope risk	All	30	20/6/4	$[66.7, 80.0]$
Whole-scope risk	Selected	25	20/2/3	$[80.0, 92.0]$
Guarded risk	All	30	13/9/8	$[43.3, 70.0]$
Guarded risk	Selected	25	13/5/7	$[52.0, 80.0]$

Dev-Mean selects 25 artifacts. Their benefit-only reference counts are 24/0/1, and the five unselected artifacts have counts 0/4/1, so selection agrees with all 28 resolved benefit labels (Table 1).

Adding whole-scope risk reduces selected positives from 24 to 20 and yields two reference negatives; adding G1/G2 reduces positives to 13 and raises negatives to five. No selected artifact is benefit-negative under the shared utility interval, so each of these five refutations requires a risk lower bound above 0.05. By Lemma 1, G1/G2 adds exactly three refutations beyond whole-scope risk. Those artifacts were previously positive or unresolved. The selected qualification bracket falls from $[96.0, 100.0]\%$ under benefit-only to $[52.0, 80.0]\%$ under the guarded contract.

Across the full panel, the nine guarded negatives comprise four artifacts already refuted by the minimum-benefit condition and five additional artifacts whose negative decisions require risk evidence. Development selection excludes the former four and retains the latter five. This partitions refutation paths: a benefit-refuted artifact may also violate risk constraints. The risk audit therefore contributes five refutations beyond those obtained from the benefit condition.

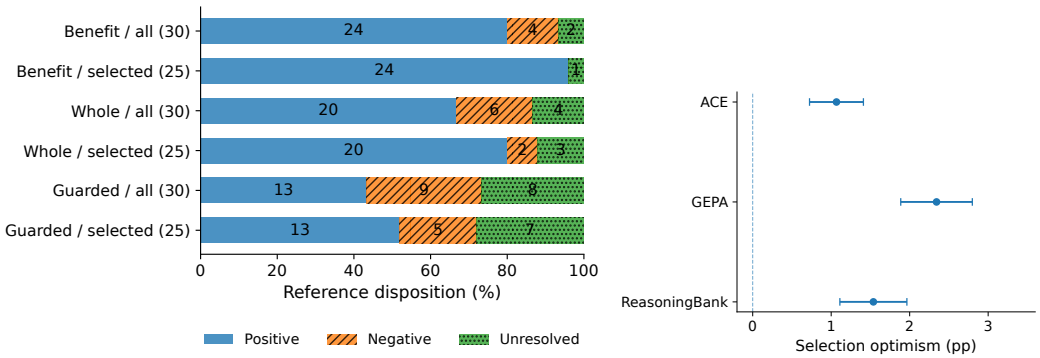
In this panel, selected unresolved counts rise from 1 to 3 to 7 as risk requirements are added. The guarded endpoint leaves seven selected artifacts unresolved at 8,192 draws. The increase from three to seven is a net change, not an identified artifact-level transition count (Appendix B.4).

Same-roster replication gives mean optimism of 1.07 pp for ACE, 2.34 pp for GEPA, and 1.54 pp for ReasoningBank, or 1.63 pp weighted by the nine, eight, and eight selected artifacts (Figure 2b, Appendix B.5). The 1.63 pp selected-panel average is comparable in magnitude to the 2 pp benefit threshold; GEPA's producer-level mean exceeds that threshold. The comparison quantifies the scale of selection optimism without assigning a correction to individual artifacts.

4.3 Judge Recommendations Retain Contract Failures

Single and Majority recommend 24 and 25 artifacts; each includes six reference negatives and omits one reference positive. The six other configurations recommend 16–22 artifacts and retain all 13 reference positives, with four (Source-Dedup, Dawid–Skene), two (Ising-CI, Bridge), or one (Ising-CD, Agent-as-a-Judge) reference negatives. Every tested judge recommends at least one reference-negative artifact.

Dev-Mean and Agent-as-a-Judge both retain all 13 reference-positive artifacts. Agent-as-a-Judge has a lower upper endpoint for the nonqualification fraction, 18.8% rather than 48.0%, but leaves more reference-unresolved artifacts outside its set. Its upper endpoint for missed qualifying mass



(a) Reference dispositions; segment labels are counts.

(b) Same-roster selection optimism.

Figure 2: **Development selection under nested contracts.** (a) All three states retain fixed panel denominators. (b) Producer-level selection–replication contrasts with descriptive replica intervals; Appendix B.5 states their interpretation.

Table 2: **Recommendation rules scored by the CertiLoop reference audit.** Bad fraction means nonqualification among recommendations; missed-good mass uses all 30 artifacts. Brackets retain unresolved labels and have the interpretation in Table 1.

Rule	Recommend	In-set $R^+/R^-/R^?$	Bad fraction (%)	Missed good (%)
Dev-Mean	25/30	13/5/7	[20.0, 48.0]	[0.0, 3.3]
Single	24/30	12/6/6	[25.0, 50.0]	[3.3, 10.0]
Majority	25/30	12/6/7	[24.0, 52.0]	[3.3, 6.7]
Source-Dedup	22/30	13/4/5	[18.2, 40.9]	[0.0, 10.0]
Dawid–Skene	21/30	13/4/4	[19.0, 38.1]	[0.0, 13.3]
Ising-CI	19/30	13/2/4	[10.5, 31.6]	[0.0, 13.3]
Ising-CD	17/30	13/1/3	[5.9, 23.5]	[0.0, 16.7]
Bridge	18/30	13/2/3	[11.1, 27.8]	[0.0, 16.7]
Agent-as-a-Judge	16/30	13/1/2	[6.3, 18.8]	[0.0, 20.0]

is therefore 20.0%, compared with 3.3% for Dev-Mean. The contrast concerns potentially missed qualification, not additional omissions of reference-positive artifacts.

Figure 3 plots each configuration's utility-difference MAE against its nonqualification bracket. Majority reduces MAE from 0.135 to 0.117 relative to Single, yet both recommend six reference negatives. Majority and Source-Dedup have almost the same MAE (0.117 and 0.116) but upper endpoints of 52.0% and 40.9%. Ising-CD has the lowest MAE (0.091), whereas Agent-as-a-Judge (0.095) has the lowest upper endpoint (18.8%). Other metrics also disagree: Agent-as-a-Judge has the lowest Brier score (0.052) and highest AUPRC (0.68), and Bridge the lowest ECE (0.047) (Appendix B.9). Task-level accuracy therefore informs, but does not determine, which recommendation set is reliable for reuse.

4.4 Evidence-Cost Trade-offs

At the reported reference checkpoints, no artifact is resolved at 2,048 or fewer paired draws. At 4,096 draws, the five resolved artifacts are all negative; positive labels first appear at the reported 8,192-draw checkpoint, which has 11 positive, nine negative, and ten unresolved artifacts. The primary endpoint instead has 13/9/8 under its larger per-interval allocation (Appendix B.7). The observed ordering is consistent with the decision rule's asymmetry: one decisive failure supports a negative, while a positive requires every condition. It is not a general ordering of resolution times, which also depend on effect sizes, variances, and stratum sample counts.

At the common 2,048-block endpoint, PA-Single and PA-Ising have normalized costs of 1.08 and 1.25, against 0.97 for DIRECT, and report 11 and 12 positive certificates, against 12. Within DI-

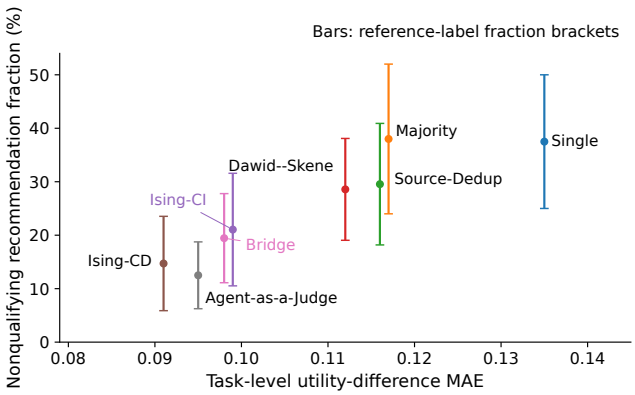


Figure 3: **Task-level prediction error and reference-supported recommendation quality.** Each judge configuration is placed at its utility-difference MAE; the vertical range is its nonqualification bracket from Table 2. Majority improves MAE over Single without reducing confirmed failures, configurations with nearly equal MAE have different brackets, and the lowest-MAE configuration (Ising-CD) does not have the lowest upper endpoint. Midpoints only position the marks.

RECT's cost, their last affordable checkpoints report 10 and 11 certificates (Table 7, Figure 4a). These profiles compare normalized accounting costs and reported certificate counts. Per-stream sample counts and certificate–reference intersections are not available for validating their sample complexity or decision errors (Appendix B.6).

PA lowers total cost only if the audit expense it avoids exceeds its added execution, prediction, and fitting expense (Eq. (30)). A break-even analysis on a separate three-artifact panel, using fixed rather than measured cost inputs, illustrates this condition: PA-Single and PA-Ising become cheaper than DIRECT when the adjudication-cost multiplier exceeds $77/137 \simeq 0.5620$ and $87/149 \simeq 0.5839$, respectively (Appendix B.10, Figure 4b). These thresholds depend on the assumed inputs; they are not measured labor savings.

4.5 Scope

The findings concern 30 equally weighted artifacts on two fixed rosters; more reference draws improve precision on these rosters, not coverage of new tasks. This study does not validate the sequential certification backends against reference labels, because their certified sets were not cross-tabulated with the reference. Reference-budget sensitivity and coupling analyses appear in Appendices B.7 and B.8.

5 Conclusion

An average benchmark gain is not evidence that a learned agent artifact should be reused. CertiLoop makes the required evidence explicit: it audits the complete frozen artifact against its comparator, fixes risk strata from initial task information, nests benefit and risk requirements on shared intervals so that each refutation can be attributed to a specific requirement, and keeps unresolved evidence separate from success and failure. Applied to 30 artifacts from ACE, GEPA, and ReasoningBank, the protocol shows that development selection reliably captures benefit but not risk: one in five selected artifacts is refuted by a risk condition, and three of these five refutations appear only under stratum conditions. Stronger judges narrow their recommendations without eliminating contract failures, and prediction assistance saves cost only when trusted adjudication is expensive. Our evidence concerns two fixed task rosters and a fixed-endpoint reference audit; validating sequential certification backends against reference labels, and extending the audit to new task distributions and additional executors, are the natural next steps.

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://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c493d23af93118975cdbc32cbe7323f5-Abstract-Conference.html.

Krishnakumar Balasubramanian, Aleksandr Podkopaev, and Shiva Prasad Kasiviswanathan. Dependence-aware label aggregation for LLM-as-a-judge via Ising models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://www.amazon.science/publications/dependence-aware-label-aggregation-for-llm-as-a-judge-via-ising-models.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. arXiv preprint arXiv:2506.07982, 2025. URL https://arxiv.org/abs/2506.07982.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Prediction-powered E-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11493–11514. PMLR, 2025. URL https://proceedings.mlr.press/v267/csillag25a.html.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or Escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

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*, 2025. URL https://openreview.net/forum?id=kzYq2hfyHB.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Mario Milanese and Antonio Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Automatica*, 27(6):997–1009, 1991. doi: 10.1016/0005-1098(91)90134-N.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10007887.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2364. URL https://papers.nips.cc/paper_files/paper/2025/hash/65a655c5a267f678fd3e897e4137ef53-Abstract-Conference.html.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging human and LLM judgments: Understanding and narrowing the gap. In *Advances*

in Neural Information Processing Systems, volume 38, 2025. doi: 10.52202/085713-0507. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/160f484617c9e3ce84332d8cbc8ba62b-Abstract-Conference.html.

Paulius Rauba, Qiyao Wei, and Mihaela van der Schaar. Auditing language models with distribution-based sensitivity analysis. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=ilNQ2m4GTy.

Shuai Shao, Qihan Ren, Dongrui Liu, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, and Jing Shao. Your agent may misevolve: Emergent risks in self-evolving LLM agents. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010580.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G0dksFayVq.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 623–645. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.27. URL https://aclanthology.org/2026.acl-long.27/.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Y. Zou, and Kunle Olukotun. Agentic Context Engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10008343.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-Judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80569–80611. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt Risk Control: A rigorous framework for responsible deployment of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5tGGWOijvq.

A Audit Protocol and Statistical Details

This appendix specifies the audit protocol and its statistical implementations. Sections A.1, A.5, and A.6 connect the main reference analysis to its sampling, decision, and reporting rules. Sequential inference, cross-unit sharing, and deployment implications are stated as extensions with explicit assumptions.

A.1 Artifact Identity, Sampling, and Evidence Roles

Whole-artifact intervention. The audit fixes $p = (C, A_1, A_0, \nu)$ before collecting its evaluation evidence. The complete ACE playbook and injection rule, GEPA prompt configuration and control flow, or ReasoningBank memory store and retrieval rule remain intact (Zhang et al., 2026; Agrawal et al., 2026; Ouyang et al., 2026). Online extraction, consolidation, and artifact updates are suspended within the audit. Configuration identity covers the executor, prompts, tool schema, environment version, comparator, resource caps, and outcome rubric. A standalone item contrast $\{k\}$ versus $\varnothing$ and a marginal contrast K versus $K \setminus \{k\}$ define different audit units; their registration states whether induced changes in retrieval or context length belong to the intervention.

Data roles and paired execution. Construction produces artifacts; development data determine predictors, selection rules, strata, and thresholds. The selection–replication contrast uses fresh executions on the same development roster after selection is fixed. Operational certification and evaluator-only reference have disjoint execution records. The protocol reserves a separate role for final-policy evaluation when that evaluation is conducted. Each arm starts from its own copy of the validated initial state, with matched resource limits and a registered ordering and randomness scheme. Agent timeouts remain scored outcomes; infrastructure failures follow the fixed retry and scoring rule and remain in the cost record. Missing judgments retain bounded uncertainty.

Roster-conditioned fresh executions. For the reported target, $J_i \sim \mathrm{Unif}(\mathcal{T}_p)$ is sampled independently with replacement and both configurations execute afresh conditional on J_i. For a normalized scalar $X_i = (1 + D_i)/2 \in [0, 1]$, sequential implementations require

$$\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = \theta. \tag{6}$$

The history includes previously revealed information used in sampling or betting. A prespecified predicate thins an independent task-draw stream to its conditional population. Different strata have different sample counts. Repeated draws of one roster task increase precision for the execution mean on that roster; a new-task-family target instead calls for sampling and analysis at the relevant task or cluster level.

Shared controls. A comparator already revealed in an earlier block is not a fresh random outcome conditional on that history. Shared controls are handled within an atomic task block: all applicable bets are fixed before the block is observed, permitted outcomes are revealed together, and the block is appended once. Reference and certification have separate caches, state, memory, and evidence access. Cross-stream independence is unnecessary for the final union bound, but each stream obeys its own sampling filtration.

Alternative fixed-outcome-bank target. For a fixed bank of N realized normalized outcomes, uniform sampling without replacement permits finite-population inference (Waudby-Smith & Ramdas, 2024). At candidate mean ϑ, the next remaining conditional mean is

$$\vartheta_i^{\mathrm{rem}} = \frac{N\vartheta - \sum_{j<i} X_j}{N - i + 1}. \tag{7}$$

Candidates inconsistent with the remaining range are excluded. For any observed subset O, a deterministic bounded completion interval is

$$\left[\frac{\sum_{i\in O} X_i}{N}, \frac{\sum_{i\in O} X_i + N - |O|}{N}\right]. \tag{8}$$

This estimates the realized bank. The main experiment instead targets fresh executions on a fixed task roster.

A.2 Sequential Implementations: DIRECT and PA

DIRECT. For a candidate mean $\vartheta \in [0,1]$, let a predictable betting strategy choose $\lambda_i(\vartheta)$ before the current draw and define

$$\psi_i(x;\vartheta) = 1 + \lambda_i(\vartheta)(x - \vartheta). \tag{9}$$

With nonnegative factors, their product is a test martingale at the true mean under Eq. (6). DIRECT applies variance-adaptive betting confidence sequences to trusted observations (Waudby-Smith & Ramdas, 2024). Fixed mixtures of opposite-sign betting products give two-sided confidence sets. After inversion, intervals are mapped back by $D = 2X - 1$. Running intersections preserve simultaneous coverage.

A simple simultaneous Hoeffding construction provides a numerical check:

$$I_n^{\mathrm{H}} = \left[\overline{X}_n - r_\alpha(n), \overline{X}_n + r_\alpha(n)\right] \cap [0,1], \qquad r_\alpha(n) = \sqrt{\frac{\log(\pi^2 n^2/(3\alpha))}{2n}}. \tag{10}$$

It allocates $6\alpha/(\pi^2 n^2)$ to each fixed time and sums these errors. This conservative check and the variance-adaptive implementation are distinct.

Prediction-assisted factors. Let $f_i(W_i) \in [0,1]$ predict the same normalized outcome without reading its current trusted label. Predictor fitting and the fallback for missing or tied verdicts are fixed beforehand. Generate

$$A_i \mid \mathcal{F}_{i-1}, W_i, X_i \sim \mathrm{Bernoulli}(\pi_i(W_i)), \qquad \pi_i(W_i) \geq \pi_* > 0. \tag{11}$$

Following Csillag et al. (2025), use

$$\psi_i^{\mathrm{PA}}(\vartheta) = \psi_i(f_i(W_i);\vartheta) + \frac{A_i}{\pi_i(W_i)}\left[\psi_i(X_i;\vartheta) - \psi_i(f_i(W_i);\vartheta)\right]. \tag{12}$$

For $0 < \chi < 1$ and $0 < \pi_* < 1$, one sufficient clipping rule is

$$-\frac{\chi\pi_*}{1-\pi_*\vartheta} \leq \lambda_i(\vartheta) \leq \frac{\chi\pi_*}{1-\pi_*(1-\vartheta)}. \tag{13}$$

The simpler choice $|\lambda_i| \leq \pi_*/2$ also suffices. For a positive bet, the corrected factor is bounded below by $[a - (1-\pi)b]/\pi$, where $a = 1 - \lambda\vartheta$ and $b = 1 + \lambda(1-\vartheta)$; the upper clipping bound makes this nonnegative. The negative-bet argument is symmetric. When $\pi_i = 1$, the correction reduces to the trusted factor.

Conditioning on the current record gives

$$\mathbb{E}[\psi_i^{\mathrm{PA}}(\vartheta) \mid \mathcal{F}_{i-1}, W_i, X_i] = \psi_i(X_i;\vartheta). \tag{14}$$

Averaging over a fresh record at $\vartheta = \theta$ then gives one. Thus nonnegative corrected products support time-uniform inference without requiring unbiased predictions or independent judge errors. Validity relies on the declared sampling law, acquisition probabilities, and clipping. The backend reserves funds for a possible audit before processing a new record.

Numerical and evidence conventions. Products are accumulated in log space; candidate inversion retains a conservative enclosure of the acceptance set. Contradictory intervals or unverified numerical results are flagged rather than converted into a positive decision. Predictor outputs and trusted labels occupy separate fields. A mixed-support decision records the inference route of each component; DIRECT and PA labels are not counted twice when they support the same artifact.

A.3 Risk Strata, Optional Bounds, and Coupling

For a positive-mass initial-information predicate g, membership-based thinning estimates $\mathbb{E}[D^h \mid g(s) = 1]$ using that stream's own observations. Overlapping predicates impose separate constraints. A stream without enough evidence keeps a full-range or justified completion interval. For an uncovered mass ρ and normalized covered-mean interval $[L_A, U_A]$, one valid completion is $[(1-\rho)L_A, (1-\rho)U_A + \rho]$; uncertainty in ρ is included in the optimization over valid input ranges.

Optional armwise and zero-prefix constructions. If ρ_1 and ρ_0 have jointly valid armwise intervals, a difference interval follows by subtracting opposite endpoints. On a representative binary treatment stream, an all-zero prefix of length n has upper risk bound

$$U^0_{1,n} = 1 - \alpha_0^{1/n}. \tag{15}$$

At candidate risk r, the process $\mathbf{1}\{c_1 = \cdots = c_n = 0\}(1-r)^{-n}$ is a nonnegative test martingale. Its threshold inversion gives Eq. (15). Since $\Delta^h = \rho_1 - \rho_0 \le \rho_1$, this can support a marginal-risk constraint. Nonzero-event data use a general risk construction. Armwise and zero-prefix routes are optional extensions, and any route affecting a decision receives an explicit error share. The number of target means alone does not count these additional constructions.

Coupling interpretation. Preserving the task-conditional arm laws preserves the marginal contrast $\rho_1 - \rho_0$, but the joint regression mass can vary. For aggregate binary arm rates r_1, r_0, unrestricted rematching gives

$$\max(0, r_1 - r_0) \le r_{10} \le \min(r_1, 1 - r_0), \qquad r_{10}^{\text{independent}} = r_1(1 - r_0). \tag{16}$$

Within-task restrictions can narrow this range. The experimental coupling analysis uses aggregate arm marginals and reports the corresponding algebraic sensitivity; its ranges are not confidence intervals for an observed pairing.

A.4 Simultaneous Validity and Frozen Versions

Proof of Proposition 1. Let $\mathcal{E}$ be the event that every component interval covers its target. On $\mathcal{E}$, a positive decision gives $\Delta^u_p \ge L^u_p \ge \epsilon$ and $\Delta^h_{p,g} \le U^h_{p,g} \le \eta_g$ for every risk condition. A negative decision gives either $\Delta^u_p \le U^u_p < \epsilon$ or $\Delta^h_{p,g} \ge L^h_{p,g} > \eta_g$ for some g. Hence both decisions have the stated contract interpretation. If component failure probabilities are bounded by α_j and $\sum_j \alpha_j \le \delta$, a union bound yields $\Pr(\mathcal{E}) \ge 1 - \delta$. No independence across components is needed.

Sequential extension. For confidence sequences, define $\mathcal{E}$ to cover every allocated stream at every time. The same argument controls the probability of any false positive over all declared units and stopping times by δ_{cert}. This includes selection among units after reading permissible audit evidence. Every construction influencing that decision, including subgroup and optional armwise bounds, is included in the same error accounting.

For successive frozen versions $e \ge 1$, one valid allocation is

$$\delta_{e,j} = \frac{\delta_{\text{cert}} w_{e,j}}{e(e+1)}, \qquad \sum_j w_{e,j} \le 1. \tag{17}$$

Version-level inference is valid conditional on prior history, and $\sum_{e \ge 1}[e(e+1)]^{-1} = 1$. A changed artifact, comparator, or population starts a new declared unit and uses prospective evidence under its allocation. Reference evaluation has a separate budget δ_{ref}.

Backend-specific families and a gate that selects among several backends have different error accounting. Three separately allocated 0.05 families give a union bound of 0.15 for a positive from any of them; a joint 0.05 rule uses a shared allocation fixed before observing evidence. These are theoretical family guarantees, distinct from within-panel recommendation fractions.

A.5 Reference Intervals and Recommendation Diagnostics

Fixed-endpoint reference. Reference evaluation keeps the intervention, target, predicates, and thresholds fixed while using fresh trusted executions unavailable to operational rules. For independent $D_i \in [-1, 1]$, sample count $n \ge 2$, and sample variance s^2 with denominator $n - 1$, the paired-difference empirical-Bernstein construction is (Maurer & Pontil, 2009)

$$\begin{aligned} r_{\text{EB}}(s, n, \alpha) &= \sqrt{\frac{2s^2 \log(4/\alpha)}{n}} + \frac{14 \log(4/\alpha)}{3(n-1)}, \\ I_{\text{EB}} &= [\overline{D} - r_{\text{EB}}, \overline{D} + r_{\text{EB}}] \cap [-1, 1]. \end{aligned} \tag{18}$$

For a fixed total-draw endpoint, membership thinning under the independent mixture design permits applying the bound conditional on a stratum's sample count. Each stratum uses its own n and s^2. Insufficient support retains a full-range interval. The untruncated full width in percentage points is $200r_{\mathrm{EB}}$.

The experiments use $8{,}192$ total paired draws per artifact and retain $\alpha_{\mathrm{ref}} = 0.05/318$ per interval. There are four targets per artifact and 120 paired-difference target intervals in the displayed panel. For this EB construction the allocated sum is $120\alpha_{\mathrm{ref}} \simeq 0.01887$. Additional armwise or zero-prefix constructions would require additional error accounting if they contributed to reference decisions. Appendix B.1 states the corresponding reference-budget scope.

The five reference checkpoints use $\alpha_{\mathrm{chk}} = \alpha_{\mathrm{ref}}/5$ at each target and each checkpoint. Their sum is again $120\alpha_{\mathrm{ref}}$. Primary and checkpoint results retain their own inputs and reference summaries; a fixed-endpoint interval is not reused as a confidence sequence. The SD-profile analysis evaluates Eq. (18) at the tabulated inputs, independently of the artifact-level status summaries.

Stratum precision. For fixed n independent uniform task draws, $\mathbb{E}[n_g] = \pi_g n$. Let s_g and s_0 be the stratum and whole-scope sample standard deviations. At equal error allocation, for $n_g, n \geq 2$ and $s_0 > 0$, the ratio of the variance terms in Eq. (18) is $(s_g/s_0)\sqrt{n/n_g}$; the ratio of the range terms is $(n-1)/(n_g-1)$. The approximations $\pi_g^{-1/2}$ and π_g^{-1} additionally use $s_g \approx s_0$, $n_g \approx \pi_g n$, and sufficiently large counts. They describe radius terms, not lower bounds on realized interval-width ratios or on the sample count needed to decide a contract. Resolution also depends on the empirical means relative to the thresholds.

Nested contract comparisons. With common scalar intervals and thresholds, write (P_b, N_b), (P_w, N_w), and (P_g, N_g) for the positive and negative sets under benefit-only, whole-scope, and guarded requirements. Then

$$P_g \subseteq P_w \subseteq P_b, \qquad N_b \subseteq N_w \subseteq N_g. \tag{19}$$

Adding requirements adds conjuncts to the positive rule and disjuncts to the negative rule, proving the set inclusions in Lemma 1. If $p \in N_g \setminus N_b$, then $U_p^u \geq \epsilon$, so a guarded negative requires $L_{p,g'}^h > \eta_{g'}$ for at least one risk predicate g'. If $p \in N_g \setminus N_w$, the absence of a whole-scope negative implies

$$U_p^u \geq \epsilon, \qquad L_{p,g_0}^h \leq \eta_{g_0}.$$

Neither shared condition therefore supplies a refutation. The guarded negative must instead have $L_{p,g'}^h > \eta_{g'}$ in G1 or G2. Crucially, $L_{p,g_0}^h \leq \eta_{g_0}$ does not imply $U_{p,g_0}^h \leq \eta_{g_0}$: the whole-scope condition is not refuted, but need not be established. Its contract status may be positive or unresolved. On the joint coverage event, an added lower bound above its tolerance refutes the corresponding population risk condition.

The partition $N_g = N_b \,\dot\cup\, (N_g \setminus N_b)$ classifies interval-based refutations; an artifact in N_b may also violate a risk condition. When G1 and G2 are added together, the set difference identifies at least one of them as a refutation, without requiring a unique failing stratum. Unresolved counts have no fixed direction of change: they can gain artifacts losing positive support and lose artifacts newly refuted. A change in their total is consequently a net change rather than an identified transition table. Appendix B.4 applies these identities to the selected panel.

Reference-supported brackets. On the reference coverage event, every $R_p = +$ is qualifying and every $R_p = -$ is nonqualifying. Thus the positive mass lower-bounds qualification and the positive plus unresolved mass upper-bounds it. Dividing by the chosen set's positive total weight proves Eq. (5). The complement gives the missed-good mass on the full-panel scale. This argument preserves all unresolved mass and uses simultaneous coverage over the fixed candidate panel, including for a development-selected subset. A zero-weight denominator is reported as not applicable.

For a certification backend m, the reference diagnostic is the intersection

$$N_{sr}^{(m)}(b) = \sum_{p \in \mathcal{P}} \mathbf{1}\{S_p^{(m)}(b) = s,\ R_p = r\}, \qquad s, r \in \{+, -, ?\}. \tag{20}$$

Positive-certificate count is its row sum $\sum_r N_{+r}^{(m)}(b)$; reference-negative and unresolved recommendations are different columns. The main experimental comparisons use recommendation–reference intersections.

Selection and prediction objects. For a development-selected set, selection optimism averages the difference between selection and fresh same-roster replication estimates, with the original fixed artifact weights. Task-level MAE evaluates the predicted utility difference; binary-risk scores evaluate their named binary target. A predictor of normalized paired risk is distinct from a classifier of an armwise violation. Judge aggregation, calibration, and agentic evaluation have separate identities and costs (Balasubramanian et al., 2026; Polo et al., 2025; Zhuge et al., 2025). Development-side fitting is separate from reference assessment. A PA-Ising backend identity does not determine which separate Ising diagnostic row it uses.

Family-error and power studies. A full-family validation trial includes its sampling, acquisition, stopping, and selection rules. False certification is evaluated under a strictly violated condition, such as $\Delta^u = \epsilon - \kappa$ or $\Delta_g^h = \eta_g + \kappa$, $\kappa > 0$. Power uses a specified qualifying alternative. Known-law trials and resampling from an empirical bank have different target distributions. These define possible validation studies; no numerical trial count or error rate is inferred from a panel row sum.

A.6 Audit Records, Replay, and Cost Accounting

An append-only audit record contains artifact/comparator and executor versions, target population, risk predicates, outcome rubric, task and arm identities, evidence role, acquisition probability, interval route, component allocation, per-stream count, decision, reason, and attributable costs. Shared comparator records have explicit identities and attribution. Reference evidence has its own access path. Replaying a visible prefix uses only that prefix and the registered randomness; unrevealed evaluator outcomes cannot change its decision.

Three counts are recorded separately: processed backend blocks, representative observations entering each scalar stream, and revealed trusted outcomes. The first is an acquisition/update counter. Its conversion to the other two depends on the declared block structure and permitted evidence. For a fixed M eligible records with one audit decision per record and probability 0.5, the design expectation is $M/2$ audits. Realized counts remain separate fields.

At cost b, certificate coverage for fixed weights is $\mathrm{Cov}_m(b) = \sum_p w_p \mathbf{1}\{S_p^{(m)}(b) = +\}$. Equal artifact weights describe panel coverage; deployment traffic has its own weights. Method-attributable cost is

$$C = N_{\mathrm{pair}} c_{\mathrm{pair}} + N_{\mathrm{judge}} c_{\mathrm{judge}} + N_{\mathrm{audit}} c_{\mathrm{audit}} + C_{\mathrm{setup}}, \tag{21}$$

with record-level sums for heterogeneous prices. Setup contains attributable construction, fitting, and calibration. Shared physical executions are charged once, failed attempts remain cost events, and reference-only work is a separate study cost.

For a common pool of N paired records and homogeneous execution and audit costs, PA is cheaper than verify-all exactly when

$$N c_{\mathrm{proxy}} + \Delta C_{\mathrm{setup}} < (N - n_A) c_{\mathrm{audit}}. \tag{22}$$

Here n_A is the audited-record count and $\Delta C_{\mathrm{setup}}$ is extra setup relative to that comparator. With different execution pools, the uncancelled method-specific counts in Eq. (21) are retained. The experimental repricing analysis holds its endpoint and other component inputs fixed while scaling adjudication cost. Its normalized units, annotation time, and campaign GPU allocation are separate quantities.

A.7 Supplementary Cross-Unit Sharing Model

The following model is a supplementary mathematical analysis, outside the main reference audit and sequential validity claim. For normalized outcome θ_p, frozen predictor mean μ_p, and source-configuration row a_p, consider the bounded affine restriction

$$\theta_p = a_p^\top \beta + \gamma \mu_p + \zeta_p, \quad |\zeta_p| \le \tau_p, \quad (\beta, \gamma) \in [-B_*, B_*]^d \times [0, \Gamma_*]. \tag{23}$$

This is a set-membership model (Milanese & Vicino, 1991). For input intervals $\Theta_p = [\ell_{\theta,p}, u_{\theta,p}]$ and $M_p = [\ell_{\mu,p}, u_{\mu,p}]$, the feasible coefficient set is

$$\mathcal{K} = \left\{(\beta,\gamma) : a_p^\top \beta + \gamma \ell_{\mu,p} - \tau_p \le u_{\theta,p}, \quad a_p^\top \beta + \gamma u_{\mu,p} + \tau_p \ge \ell_{\theta,p} \quad \forall p\right\}, \tag{24}$$

intersected with the coefficient box. If $\mathcal{K} \neq \varnothing$,

$$\begin{aligned} L_q^{\mathrm{LP}} &= \max\{\ell_{\theta,q}, \min_{\mathcal{K}}(a_q^\top \beta + \gamma \ell_{\mu,q} - \tau_q)\}, \\ U_q^{\mathrm{LP}} &= \min\{u_{\theta,q}, \max_{\mathcal{K}}(a_q^\top \beta + \gamma u_{\mu,q} + \tau_q)\}. \end{aligned} \tag{25}$$

For each fixed coefficient, feasibility is interval intersection at every cell; projecting the resulting joint set gives these extrema.

Proposition 2 (Fixed-witness dominance for identical information)**.** *If a known coefficient* $(\bar\beta, \bar\gamma)$ *lies in* $\mathcal{K}$*, then*

$$I_q^{\mathrm{FW}} = \Theta_q \cap [a_q^\top \bar\beta + \bar\gamma \ell_{\mu,q} - \tau_q, a_q^\top \bar\beta + \bar\gamma u_{\mu,q} + \tau_q] \subseteq [L_q^{\mathrm{LP}}, U_q^{\mathrm{LP}}].$$

The fixed slice is a subset of the feasible joint set, so its target extrema lie within the full extrema. With unchanged predictor, radii, and target intervals, more source audits test compatibility but cannot shrink this known-coefficient slice. Unknown shared coefficients impose a different information contract.

Calibration and precision requirements. Freeze a coefficient witness and positive scale functions on training data. For M exchangeable calibration cells and J exchangeable target cells, a score bounding the maximum normalized residual over outcomes can use order statistic $k = \lceil (M+1)(1-\delta_S/J) \rceil$, with an appended infinite score. A finite radius under this construction needs $M + 1 \ge J/\delta_S$. Calibration interval failure consumes a separate share. Two bounded trusted means per cell with Hoeffding half-width r_C require $n_C \ge \log(4M/\delta_C)/(2r_C^2)$, plus the corresponding prediction, execution, and fitting costs.

Let $z_p = (a_p^\top, \widehat{\mu}_p)^\top$, $b_p = \tau_p + h_p^T + \Gamma_* h_p^P$, and $R_* = \max(B_*, \Gamma_*/2)$. Differences v between feasible coefficients satisfy $|Z_S v| \le 2b$ and $\|v\|_\infty \le 2R_*$. For any λ,

$$|z_q^\top v| \le 2 \sum_{p \in S} b_p |\lambda_p| + 2R_* \|z_q - Z_S^\top \lambda\|_1. \tag{26}$$

Target residual and proxy uncertainty add at most $2(\tau_q + \Gamma_* h_q^P)$. With exact representation, $\|\lambda\|_1 \le \Lambda$, and uniform errors, the corresponding width screen is $2(1+\Lambda)(\bar\tau + \Gamma_* \bar h^P) + 2\Lambda \bar h^T$. At a marginal-risk null, its necessary residual feasibility condition is $4(1+\Lambda)(\bar\tau + \Gamma_* \bar h^P) < \eta$. These requirements concern this sufficient construction. Evaluating its empirical usefulness would compare shared and target-local fits under the same predictor, interval inputs, coefficient bounds, and calibration costs.

A.8 Conditional Deployment Implications

Within one system, suppose a frozen router defines mutually exclusive cells, G_p is a certificate-based gate, and π_{cmp} uses the registered comparator in each cell. Let ω_p be deployment mass. If audit and deployment cell populations and conditional execution laws agree, and all invoked certificates are valid, then

$$J(\pi_{\mathrm{CL}}) - J(\pi_{\mathrm{cmp}}) = \sum_p \omega_p G_p \Delta_p^u \ge \epsilon \sum_p \omega_p G_p. \tag{27}$$

For positive invocation mass, conditional marginal risk is bounded by the invocation-weighted mean of the whole-scope tolerances. Overlapping stratum claims remain conditional constraints rather than disjoint deployment cells. An additional within-cell gate changes the population and belongs to the router definition before auditing.

With unchanged conditional execution kernels and $d_{\mathrm{TV}}(Q'_p, Q_p) \le \upsilon_p$, the difference range $[-1, 1]$ gives $|\Delta_p^{m\prime} - \Delta_p^m| \le 2\upsilon_p$, hence

$$\begin{aligned} J(\pi_{\mathrm{CL}}) - J(\pi_{\mathrm{cmp}}) &\ge \sum_p \omega'_p G_p (\epsilon - 2\upsilon_p), \\ \mathbb{E}[D^h \mid \text{invoke}] &\le \frac{\sum_p \omega'_p G_p (\eta_p + 2\upsilon_p)}{\sum_p \omega'_p G_p}. \end{aligned} \tag{28}$$

The second denominator is positive. Subgroup guarantees require their own conditional shift bounds. These are consequences of valid certificates under stated populations; the reported fixed-panel reference study does not measure the realized return of a deployed policy.

B Experimental Details and Complete Aggregate Results

This appendix specifies the risk strata, reference decision rule, and aggregate results underlying Section 4. The analysis package reproduces all count decompositions, ratios, interval widths, coupling ranges, and break-even thresholds. Reference-label interpretation follows the measurement and coverage conditions stated in the main text.

B.1 Configuration and Units of Analysis

We audit complete ACE playbooks, GEPA prompt configurations, and ReasoningBank memory stores with their retrieval rules, each against its registered comparator. Five replicas per producer and environment form the 30-artifact panel. Development selection and its same-roster replication are a separate descriptive contrast. Reference labels evaluate the contract, whereas operational counts describe the certification backend outputs.

Table 3: **Evaluation configuration and campaign resources.** Model identifiers specify the evaluation configurations. Campaign allocation is wall-clock device time, separate from per-method evidence costs.

Field	Setting
Artifact panel	2 environments $\times$ 3 producers $\times$ 5 replicas; 30 artifacts
Rosters	AppWorld Test-C; text-based τ-Retail held-out roster
Stratum membership	Frozen per-task predicate membership shared across certification and reference
Executor	`Qwen/Qwen3.6-35B-A3B-FP8`
Base judge bank	`google/gemma-4-31B-it`, `mistralai/Mistral-Small-3.2-24B-Instruct-2506`, and `openai/gpt-oss-120b`; two templates per family
Agentic judge	`openai/gpt-oss-120b`
Bridge encoder	`Qwen/Qwen3-Embedding-8B`
Reference contract	Benefit $\epsilon = .02$; whole-scope and G1/G2 risk tolerances .05
Reference endpoint	8,192 draws; $\alpha_{\mathrm{ref}} = .05/318$ per interval; 120 intervals in the displayed panel
Sensitivity checkpoints	512, 1,024, 2,048, 4,096, and 8,192 draws; $\alpha_{\mathrm{chk}} = \alpha_{\mathrm{ref}}/5$
Operational comparison	DIRECT, PA-Single, PA-Ising; 128–2,048 processed blocks
PA acquisition setting	Design audit probability .5
Operational family level	.05 per declared backend family, separate from reference error allocation
GPU	NVIDIA L40S $\times 4$; reported memory 46,068 MiB per GPU
Campaign	Approximately five weeks; about 3,360 allocated GPU-hours on four devices
CPU	Intel Xeon Gold 6326 at 2.90 GHz
System memory	2.0 TiB

The audit target is the roster-conditioned fresh-execution mean $|\mathcal{T}|^{-1}\sum_{j\in\mathcal{T}}\mathbb{E}[D_p \mid J = j]$. Each draw generates a new paired execution on a uniformly sampled roster task. Certification and reference have separate execution streams on these rosters. Repeated task identities contribute execution precision on the same fixed target population.

Reference allocation for the displayed panel. Each primary artifact has four scalar targets: utility, whole-scope risk, G1 risk, and G2 risk. The 30-artifact panel therefore comprises $30 \times 4 = 120$ target intervals. We retain $\alpha_{\mathrm{ref}} = 0.05/318$ per interval, so

$$\sum_{p=1}^{30}\sum_{j=1}^{4}\alpha_{p,j} = 120\frac{0.05}{318} \le 0.05.$$

This is the numerical error accounting for the displayed primary panel. For the five-checkpoint analysis, each scalar interval at each checkpoint uses $\alpha_{\mathrm{chk}} = \alpha_{\mathrm{ref}}/5$, yielding the same summed

allocation $120\alpha_{\mathrm{ref}}$ across those checkpoints. Primary and checkpoint results retain their respective interval inputs and reported states. This calculation neither reallocates the original shares nor specifies the allocation of additional study targets.

Campaign size. Five weeks correspond to approximately 840 wall-clock hours, or 3,360 allocated GPU-hours on four L40S devices. The reference design comprises $30 \times 8{,}192 = 245{,}760$ artifact–draw pairs and 491,520 logical arm-record incidences before shared-comparator deduplication. Allocation time and logical record incidences are distinct from device utilization and unique executions.

B.2 Four Frozen Risk-Stratum Predicates

For environment e, write s for the initial task instruction, frozen task metadata, and validated initial environment state. The rules are fixed before certification and reference outcomes are examined. The membership function $g_{e,k}(s) \in \{0, 1\}$, $k \in \{1, 2\}$, is evaluated before either arm runs. The same environment-specific mapping is used across all artifact replicas and for both arms and evidence roles. Membership is not assigned from an agent's trajectory, an observed violation, a judge's verdict, or a reference interval. G1 and G2 identify task populations, not model families or two alternative measurements of whether an outcome is safe.

Table 4: **Human-readable predicates used for the guarded contract.** All four use initial information. The two predicates within an environment may overlap; they do not define a mutually exclusive partition.

Environment	Stratum	Membership definition
AppWorld	G1	AppWorld-G1 contains tasks for which correct execution depends materially on temporal information available in the initial task specification, including absolute dates or times, relative-time expressions, deadlines, event ordering, or timestamp-dependent object selection.
AppWorld	G2	AppWorld-G2 contains tasks whose registered solution scope permits or requires a persistent externally observable state change, such as creating, deleting, sending, purchasing, booking, or modifying an application object. Membership is determined from the frozen task specification before either arm is executed; read-only information-retrieval tasks are excluded.
τ-Retail	G1	Retail-G1 contains requests for which the authorized resolution entails a modification to an existing order or fulfillment state, including cancellation, return, exchange, replacement, or another registered order-state transition. Pure inquiries about an order or policy are excluded unless the initial request authorizes such a state transition.
τ-Retail	G2	Retail-G2 contains requests in which the authorized resolution changes, determines, or verifies a monetary quantity or policy entitlement, including refunds, payment instruments, credits, discounts, reimbursements, compensation, or fee-related eligibility. Membership is fixed from the initial request and task metadata before execution.

Inclusion and exclusion boundaries. AppWorld G1 includes relative dates, time intervals, deadlines, chronological ordering, calendar scheduling, and selection of the latest or earliest object when they change a correct action or its parameters. An incidental date that does not affect the solution, or a timestamp produced only during execution, does not establish membership. AppWorld G2 includes permitted changes to calendar events, messages, settings, contacts, subscriptions, application or database records, and transactions such as purchases, bookings, and transfers. Its defining condition is that the task scope *permits or requires* the persistent change, not that an agent happened to perform one.

Retail G1 includes authorized changes to an existing order, its items, or its fulfillment state, including cancellation, return, exchange, replacement, and registered shipment or delivery changes. Asking for order status or merely asking whether cancellation is possible does not authorize that transition. Retail G2 includes determining amounts or eligibility for refunds, payment instruments, prices, discounts, coupons, store credit, reimbursement, compensation, and fee waivers. A monetary-eligibility

inquiry can therefore belong to G2 even when it requires no order mutation. Pure product or shipping queries without monetary or entitlement consequences do not belong to G2. An agent's erroneous mutation cannot move an otherwise excluded task into any stratum after execution.

Overlap and the outcome distinction. A request to return an item and refund its original payment instrument belongs to both Retail strata. A time-constrained calendar modification can belong to both AppWorld strata. These examples illustrate membership. A task outside both predicates remains covered by the whole-scope condition. The predicates answer *which task population is being assessed*; the protected-outcome variable $c(Y)$ answers *whether the registered violation occurred*. A correct authorized mutation can have $c(Y) = 0$, and a temporal mistake can reduce utility without creating a protected risk event. No outcome-dependent relabeling is used to define G1 or G2.

B.3 From Membership to Risk Intervals and Decisions

For artifact p in environment e, let $\mathcal{I}_{p,k}$ select its paired draws whose initial task information satisfies $g_{e,k}(s_i) = 1$. For the whole scope, use every permitted draw. With $D^h_{p,i} = c(Y_{p,i}(1)) - c(Y_{p,i}(0))$, the conditional target and unweighted estimator are

$$\Delta^h_{p,k} = \mathbb{E}[D^h_{p,i} \mid g_{e,k}(s_i) = 1], \qquad \widehat{\Delta}^h_{p,k} = \frac{1}{n_{p,k}} \sum_{i \in \mathcal{I}_{p,k}} D^h_{p,i}, \quad n_{p,k} = |\mathcal{I}_{p,k}|.$$

The mean requires $n_{p,k} > 0$; the endpoint empirical-Bernstein calculation requires $n_{p,k} \geq 2$ and uses sample-variance denominator $n_{p,k} - 1$. Whole-scope utility retains the full registered task population, not the union of the two selected strata. Under the stated uniform-with-replacement design, pre-execution membership thinning selects the conditional population; a task in both strata can update both distinct risk streams. It is not counted as two new draws in the whole-scope estimator or two artifacts in a coverage total. Unknown or insufficient support prevents a positive guarded decision; a separate confirmed failing condition can still yield a negative decision. Otherwise the artifact remains unresolved. Every registered stratum remains part of the guarded contract.

For valid independent-reference intervals, the guarded reference rule is

$$R_p = \begin{cases} +, & L^u_p \geq 0.02 \text{ and } U^h_{p,k} \leq 0.05 \quad \text{for every } k \in \{\text{whole}, 1, 2\}, \\ -, & U^u_p < 0.02 \text{ or } L^h_{p,k} > 0.05 \quad \text{for some } k, \\ ?, & \text{otherwise.} \end{cases}$$

A confirmed failing condition suffices for a negative reference label; a positive label requires support for every condition. Benefit-only uses utility, and whole-scope uses utility and overall risk. All shared scalar intervals and thresholds remain unchanged in the three reference comparisons. Let P_b, P_w, P_g and N_b, N_w, N_g denote their positive and negative sets. Then

$$P_g \subseteq P_w \subseteq P_b, \qquad N_b \subseteq N_w \subseteq N_g.$$

An added risk condition can turn a positive label into a negative or unresolved label, or an unresolved label into a negative one. The intervals remain fixed, so these transitions isolate the decision effect of the additional condition.

The endpoint reference bound uses the sample size and variance of the relevant scalar stream and its allocated α; Appendix B.7 gives its range-correct width. The two SD profiles in that appendix provide an analytical precision comparison; reference decisions use their artifact-specific scalar intervals. Overlap does not require independence between risk streams for the final union bound, but each stream must satisfy its own sampling and measurement assumptions.

Recommendation rules. Dev-Mean retains artifacts with positive development-side utility estimates. Single/Judge-Gate applies the benefit and every relevant risk threshold to its frozen predicted artifact-effect tuple. It does not turn a single episode prediction into an artifact certificate and does not use the current trusted reference outcome. The other judge rows apply the same rule to their own frozen predictions. Table 2 reports how each recommendation set intersects the reference labels. Supplementary operational profiles appear in Table 8.

B.4 Reference Counts and Bracket Calculations

For a panel of size N with reference counts (P, Q, U), define

$$I_{\text{qual}} = \left[\frac{P}{N}, \frac{P+U}{N}\right].$$

For a recommended set of size $m > 0$ containing (p, q, u) reference labels,

$$I_{\text{bad,fraction}} = \left[\frac{q}{m}, \frac{q+u}{m}\right], \qquad I_{\text{bad,mass}} = \left[\frac{q}{N}, \frac{q+u}{N}\right],$$

$$I_{\text{missed,mass}} = \left[\frac{P-p}{N}, \frac{P-p+U-u}{N}\right].$$

The numerical tables multiply these quantities by 100. A recommendation fraction is conditional on the recommended set, while a missed-good mass uses the full panel denominator; the two denominators are intentionally different. They are brackets that retain unresolved reference mass, not binomial confidence intervals; on the simultaneous reference coverage event, and given valid outcome measurement, they bound the true fractions.

Table 5 includes the unselected complement for each requirement. Complement counts are full minus selected counts and introduce no new artifacts. The 25 selected artifacts represent 83.3% of the full panel. Under the guarded contract, Dev-Mean has bad mass $[5/30, 12/30] = [16.7\%, 40.0\%]$ and missed-good mass $[0/30, 1/30] = [0.0\%, 3.3\%]$.

Table 5: **Complete disposition summaries and count complements.** Unselected rows are derived by subtraction. Resolved mass is $(R^+ + R^-)/N$; it includes both positive and negative reference labels.

Requirements	Panel	N	$R^+/R^-/R^?$	Qualification (%)	Resolved (%)
Benefit only	All	30	24/4/2	[80.0, 86.7]	93.3
Benefit only	Selected	25	24/0/1	[96.0, 100.0]	96.0
Benefit only	Unselected	5	0/4/1	[0.0, 20.0]	80.0
Whole-scope risk	All	30	20/6/4	[66.7, 80.0]	86.7
Whole-scope risk	Selected	25	20/2/3	[80.0, 92.0]	88.0
Whole-scope risk	Unselected	5	0/4/1	[0.0, 20.0]	80.0
Guarded risk	All	30	13/9/8	[43.3, 70.0]	73.3
Guarded risk	Selected	25	13/5/7	[52.0, 80.0]	72.0
Guarded risk	Unselected	5	0/4/1	[0.0, 20.0]	80.0

What the nested counts identify. For the development-selected set A, the counts give $|A \cap N_b| = 0$, $|A \cap N_w| = 2$, and $|A \cap N_g| = 5$. Thus every guarded negative in A requires a risk refutation, and

$$|A \cap (N_g \setminus N_w)| = 5 - 2 = 3.$$

Those three new negatives require G1 or G2 evidence, because all weaker conditions retain their intervals. Their prior whole-scope states can be positive or unresolved.

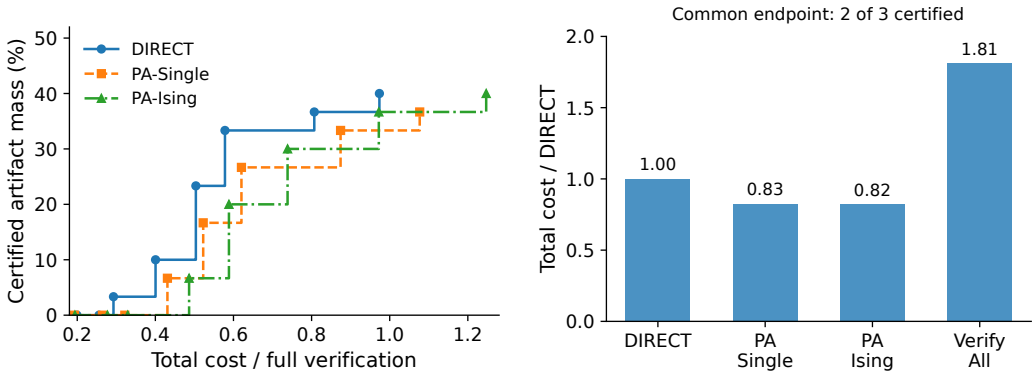
The unresolved count rises from 3 to 7, a net increase of four. If x previously positive artifacts become negative, then $3 - x$ previously unresolved artifacts become negative and $7 - x$ previously positive artifacts become unresolved, for $0 \le x \le 3$. All these transitions yield the same 20/2/3 and 13/5/7 margins. The net change therefore permits several artifact-level transition tables.

B.5 Same-Roster Selection Optimism

Selection optimism is the selected-artifact average of the difference between the development-selection estimate and its same-roster replication estimate. For producer-level means $\widehat{O}_k$ and selected counts n_k,

$$\widehat{O}_{\text{sel}} = \frac{\sum_k n_k \widehat{O}_k}{\sum_k n_k}$$

$$= \frac{9(1.0673554551) + 8(2.3423971952) + 8(1.5376625677)}{25} = 1.6258670879 \text{ pp.}$$



(a) Programmatic cost–coverage profiles.

(b) Human-adjudication accounting at $\lambda = 1$.

Figure 4: **Coverage and cost under distinct verification channels.** (a) Reported positive-certificate counts against normalized programmatic accounting cost. (b) Fixed cost inputs at a common two-of-three endpoint in a separate accounting panel. The right panel is the $\lambda = 1$ snapshot; the break-even analysis varies its adjudication component. The adjudication component is an accounting input.

This weights artifacts equally rather than producers. Table 6 gives all producer summaries. The intervals are reported replica t intervals and are used descriptively, since the result files do not specify their nominal confidence level. The 1.63 pp comparison with $\epsilon = 2$ pp concerns the selected-panel average.

Table 6: **Producer-level same-roster selection contrasts.** Means, intervals, and half-widths are in percentage points.

Producer	Selected	Mean	Reported interval	Half-width
ACE	9	1.067	$[0.724, 1.411]$	0.343
GEPA	8	2.342	$[1.886, 2.798]$	0.456
ReasoningBank	8	1.538	$[1.111, 1.965]$	0.427

B.6 Supplementary Backend Count and Cost Profiles

Table 7 compares the reported terminal counts and the last recorded checkpoints within a common accounting-cost limit. Table 8 includes every checkpoint. These aggregate profiles provide backend progress counters, positive-certificate counts, and normalized costs. Their operational predicate set, per-stream sample counts, and certificate–reference intersections are unavailable in the result files; consequently, these profiles assess count–cost behavior rather than certification validity or label efficiency.

Table 8 contains every processed-block checkpoint from Figure 4a. Each entry reports normalized cost and positive-certificate count. Coverage is that count divided by 30. The zero-coverage origins in Figure 4a are setup costs (0.1993, 0.1930, and 0.1940), not processed-block observations.

At the endpoint, dividing the three costs by DIRECT's 0.9736 gives 1.0000, 1.1060, and 1.2808. For the common-cost snapshot in Table 7, each method contributes its last checkpoint with cost at most 0.9736, giving block counts 2,048, 1,536, and 1,536. No outcomes are interpolated; this is a prefix comparison of the recorded runs, not a separately executed budget-stopped experiment.

B.7 Analytical Reference Widths and Checkpoint Resolution

For independent differences in $[-1, 1]$ and sample variance s^2 with denominator $n - 1$, the range-correct empirical Bernstein full width in pp is

$$W(s, n, \alpha) = 200 \left[\sqrt{\frac{2s^2 \log(4/\alpha)}{n}} + \frac{14 \log(4/\alpha)}{3(n-1)} \right]. \tag{29}$$

Table 7: **Programmatic accounting costs and reported positive certificates.** Costs use the full-verification accounting normalizer. Panel B takes each method's last recorded checkpoint with cost at most 0.9736, without interpolation. Blocks denote the recorded backend counters.

Method	Blocks	Normalized cost	Certificates	Coverage (%)
A. Common processed-block endpoint				
DIRECT	2,048	0.9736	12/30	40.0
PA-Single	2,048	1.0768	11/30	36.7
PA-Ising	2,048	1.2470	12/30	40.0
B. Common cost limit				
DIRECT	2,048	0.9736	12/30	40.0
PA-Single	1,536	0.8742	10/30	33.3
PA-Ising	1,536	0.9721	11/30	36.7

Table 8: **All programmatic checkpoints.** Costs use the full-verification accounting normalizer. Counts are reported positive certificates, not a reference-verified confusion matrix.

	DIRECT		PA-Single		PA-Ising	
Blocks	Cost	Count/30	Cost	Count/30	Cost	Count/30
128	0.2557	0/30	0.2643	0/30	0.2774	0/30
256	0.2928	1/30	0.3218	0/30	0.3293	0/30
512	0.4005	3/30	0.4309	2/30	0.4864	2/30
768	0.5037	7/30	0.5225	5/30	0.5884	6/30
1,024	0.5782	10/30	0.6204	8/30	0.7385	9/30
1,536	0.8069	11/30	0.8742	10/30	0.9721	11/30
2,048	0.9736	12/30	1.0768	11/30	1.2470	12/30

The factor 200 converts a half-width on the original difference scale to a full width in percentage points (Maurer & Pontil, 2009). We evaluate it with $\alpha_{\mathrm{chk}} = 0.05/(318 \cdot 5) \approx 3.14 \times 10^{-5}$ at the two budget-specific SD profiles labeled p25 and p75. Table 9 lists their exact numerical inputs. The names p25 and p75 identify the two input profiles used in this calculation. The p25 profile falls from 35.11 pp at 512 draws to 4.75 pp at 8,192, and the p75 profile from 38.34 pp to 5.75 pp (Figure 5a). At 2,048 draws their full widths are 12–14 pp. Resolution depends jointly on width and distance to each applicable threshold; the corresponding three-state counts are tabulated separately.

Table 9: **All precision inputs, derived widths, and checkpoint counts.** W is calculated by the displayed formula, in pp. Three-state counts come from the separate checkpoint resolution summary; they are not regenerated from two SD profiles.

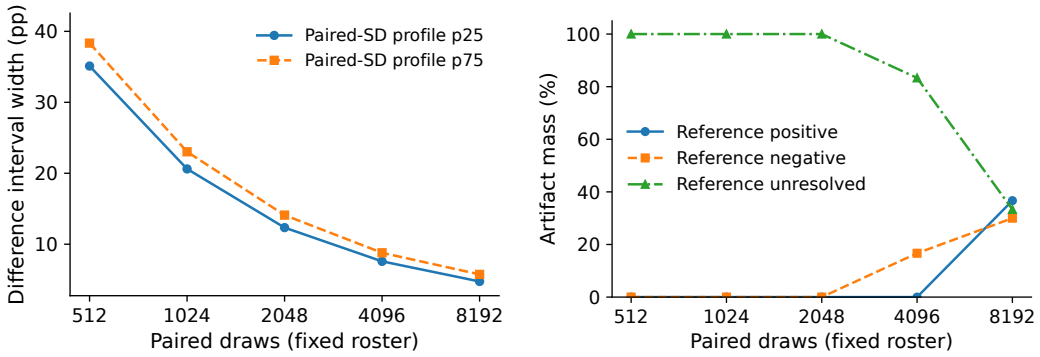
Draws	s_{25}	s_{75}	$W(s_{25})$	$W(s_{75})$	$R^+/R^-/R^?$	Resolved (%)
512	0.318343	0.393736	35.11	38.34	0/0/30	0.0
1,024	0.326365	0.406066	20.61	23.03	0/0/30	0.0
2,048	0.325804	0.407963	12.34	14.10	0/0/30	0.0
4,096	0.323809	0.403531	7.58	8.79	0/5/25	16.7
8,192	0.318313	0.411238	4.75	5.75	11/9/10	66.7

No artifact is resolved at 512, 1,024, or 2,048 draws; five are resolved (all negative) at 4,096, and 20 at 8,192 (Figure 5b). The 8,192-draw checkpoint resolves $20/30 = 66.7\%$ of the panel, whereas the primary endpoint, with its larger per-stream allocation, resolves $22/30 = 73.3\%$. More draws improve precision on a fixed roster; they do not increase the number of distinct tasks.

B.8 Algebraic Coupling Calculations

For treatment and comparator marginal rates r_1, r_0, the admissible joint regression mass $r_{10} = \Pr(c_1 = 1, c_0 = 0)$ satisfies

$$\max(0, r_1 - r_0) \le r_{10} \le \min(r_1, 1 - r_0).$$



(a) Analytical reference-width profiles.

(b) Checkpoint resolution.

Figure 5: **Reference-budget sensitivity.** (a) Empirical-Bernstein full widths at the two budget-specific SD profiles labeled p25 and p75. (b) Guarded reference states at each checkpoint under α_{chk}. The horizontal axis counts fresh paired draws on the fixed rosters, not distinct tasks.

The marginal difference is $r_1 - r_0$ under every coupling. Independent random matching gives expected regression mass $r_1(1 - r_0)$. Table 10 reports all inputs and outputs; each displayed range is a deterministic function of its two marginal rates.

Table 10: **Arm marginals and Fréchet coupling sensitivity.** Arm rates, matching ranges, and random-matching expectations are percentages; marginal differences are percentage points.

Case	r_1	r_0	Difference	Matching range	Random matching
N0	14.84375	13.28125	1.56250	$[1.56250, 14.84375]$	12.8723
N1	17.96875	14.06250	3.90625	$[3.90625, 17.96875]$	15.4419

N0 denotes the identical-configuration control and N1 the length-matched placebo. Configuration identity does not force two finite-sample arm means to coincide. The N0 inputs give a 1.56 pp marginal difference, an admissible matching range of $[1.56\%, 14.84\%]$, and independent-matching expectation 12.87%; N1 gives 3.91 pp, $[3.91\%, 17.97\%]$, and 15.44%. These are algebraic outputs from fixed marginal inputs, not a significance test of either intervention or measured paired-regression frequencies. Although the input rates are multiples of $1/128$, this does not identify the original sample size. The aggregate inputs specify neither that denominator nor the control environment, so no per-task causal regression rate or population confidence interval is inferred.

B.9 Complete Judge Diagnostics and Recommendation Counts

Table 11 reports all eight configurations' task-level prediction summaries. The risk-positive prevalence is reported as 9.4%. The tie-or-missing field is a combined percentage. MAE refers to the reported utility-difference prediction target; Brier, AUPRC, and ECE refer to the reported risk-prediction evaluation. All values come from our evaluation with the judge bank in Table 3, not from the cited papers. They measure task-level prediction quality and are not errors of an artifact certificate.

For each rule with m recommendations, b_L of which are reference-negative and b_U reference-negative or unresolved, the in-set counts are $(p, q, u) = (m - b_U, b_L, b_U - b_L)$. Subtracting from the guarded totals $(13, 9, 8)$ gives the counts outside the recommendation set. Table 12 displays both partitions, all recommended fractions, and bad mass; Table 2 reports the corresponding bad-fraction and missed-good brackets. The Single row is the Judge-Gate rule. The PA-Ising operational backend is evaluated separately and is not identified with the Ising-CD row.

Ising-CD's MAE is 32.6% below Single's in this evaluation. Its 17 recommendations retain all 13 reference-qualified artifacts but also one negative and three unresolved labels. Agent-as-a-Judge recommends 16, comprising 13 positive, one negative, and two unresolved labels. Its smaller rec-

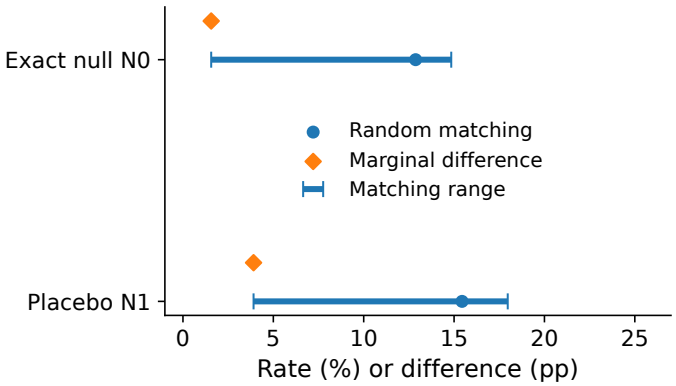


Figure 6: **Algebraic coupling sensitivity.** Horizontal ranges are Fréchet matching ranges; circles indicate independent random-matching expectations and diamonds show marginal differences. The ranges are not confidence intervals, and rematching supplies no additional independent observations.

Table 11: **Task-level judge metrics.** Lower MAE, Brier, and ECE and higher AUPRC are preferable. The reported risk-positive prevalence is 9.4%; tie/missing is a combined field.

Predictor	MAE ↓	Brier ↓	AUPRC ↑	ECE ↓	Tie/missing (%)
Single	0.135	0.079	0.43	0.088	4.1
Majority	0.117	0.076	0.50	0.083	4.8
Source-Dedup	0.116	0.067	0.53	0.072	4.5
Dawid–Skene	0.112	0.065	0.56	0.063	4.7
Ising-CI	0.099	0.062	0.60	0.053	4.4
Ising-CD	0.091	0.053	0.64	0.048	4.2
Bridge	0.098	0.055	0.60	0.047	3.9
Agent-as-a-Judge	0.095	0.052	0.68	0.051	5.2

Table 12: **Reference-label composition of judge recommendation sets.** The two three-state columns partition the same 30-artifact panel. Bad mass uses denominator 30 and has the conditional interpretation stated in Appendix B.4.

Predictor	Recommend (%)	In-set $+/-/?$	Outside $+/-/?$	Bad mass (%)
Single	80.0	12/6/6	1/3/2	[20.0, 40.0]
Majority	83.3	12/6/7	1/3/1	[20.0, 43.3]
Source-Dedup	73.3	13/4/5	0/5/3	[13.3, 30.0]
Dawid–Skene	70.0	13/4/4	0/5/4	[13.3, 26.7]
Ising-CI	63.3	13/2/4	0/7/4	[6.7, 20.0]
Ising-CD	56.7	13/1/3	0/8/5	[3.3, 13.3]
Bridge	60.0	13/2/3	0/7/5	[6.7, 16.7]
Agent-as-a-Judge	53.3	13/1/2	0/8/6	[3.3, 10.0]

ommendation count has a lower upper nonqualification endpoint and a higher upper missed-good endpoint, reflecting the different numbers of unresolved labels inside and outside the set. Both intervals retain the reference-unresolved labels when comparing the two recommendation rules. Figure 3 in Section 4.3 plots these brackets against utility-difference MAE.

B.10 Human-Adjudication Cost Accounting

This analysis asks when the fixed accounting inputs favor PA if trusted outcomes require human adjudication. It conditions every method on the reported two-certificate endpoint of a separate three-artifact panel. Table 13 provides the component inputs, in normalized units rather than measured

hours or currency. We analyze this schedule as a break-even calculation, not a completed human-labeling cost measurement.

Construction and execution costs are fixed inputs to this scenario. The comparisons below condition on the complete component schedule rather than estimating those quantities from the common endpoint.

Cost balance. Write $C_m = C_m^{\text{other}} + C_m^{\text{audit}}$ for a method's costs, with C_m^{other} containing construction, execution, proxy, and fitting. Then

$$\begin{aligned} C_{\text{PA}} - C_{\text{DIRECT}} = (C_{\text{PA}}^{\text{other}} - C_{\text{DIRECT}}^{\text{other}}) \\ - (C_{\text{DIRECT}}^{\text{audit}} - C_{\text{PA}}^{\text{audit}}). \end{aligned} \tag{30}$$

PA is cheaper precisely when avoided audit expense exceeds the change in other expense. The condition is evaluated componentwise in the accounting scenario below.

Table 13: **Human-adjudication accounting components.** Each total is the sum of its five component inputs. All four rows are conditioned on the same reported two-of-three accounting endpoint.

Method	Construct	Execute	Proxy	Adjudicate	Fit	Total
DIRECT	19	49	0	277	0	345
PA-Single	21	103	13	140	8	285
PA-Ising	17	102	30	128	6	283
Verify-All-Legal	17	104	0	504	0	625

Table 14: **Derived relative costs.** A negative saving denotes a higher cost than DIRECT under the same inputs.

Method	Total	Cost / DIRECT	Saving versus DIRECT (%)
DIRECT	345	1.0000	0.0000
PA-Single	285	0.8261	17.3913
PA-Ising	283	0.8203	17.9710
Verify-All-Legal	625	1.8116	-81.1594

The PA savings are $1 - 285/345 = 17.3913\%$ and $1 - 283/345 = 17.9710\%$. Verify-All-Legal costs $625/345 = 1.8116$ times DIRECT. To expose the dependence on adjudication cost, multiply only that component by $\lambda \geq 0$ while holding the other accounting inputs and the common endpoint fixed:

$$\begin{aligned} C_{\text{DIRECT}}(\lambda) = 68 + 277\lambda, \quad C_{\text{Single}}(\lambda) = 145 + 140\lambda, \\ C_{\text{Ising}}(\lambda) = 155 + 128\lambda, \quad C_{\text{Verify}}(\lambda) = 121 + 504\lambda. \end{aligned}$$

PA-Single costs less than DIRECT when $\lambda > 77/137 \simeq 0.5620$; PA-Ising does so when $\lambda > 87/149 \simeq 0.5839$. PA-Ising costs less than PA-Single when $\lambda > 5/6$. At the original inputs, $\lambda = 1$, their difference is two units, or 0.70% of PA-Single's total. These thresholds hold with the other accounting inputs and endpoint fixed.

B.11 Analysis Package and Reproducibility

The analysis package contains unchanged aggregate CSV inputs, derived tables, figure-generation code, and numeric consistency checks. The reference tables and Figure 2 share one disposition summary; the judge tables and Figure 3 share one recommendation summary. The cost, width, and coupling analyses list their complete numerical inputs. The four predicate definitions specify the strata used for reference interpretation.

The available release does not include task-ID rosters and stratum counts, the Retail simulator identifier, or the complete run-level allocation manifest. The 120-target calculation in Appendix B.1 establishes the numerical reference-budget bound for the displayed panel; it does not reconstruct that manifest. The separate operational-profile boundary is stated in Appendix B.6. These limitations apply to reproducing the execution protocol, whereas the aggregate analyses and their figures are reproducible from the released inputs.