ROBUST AI-AGENT REPUTATION VIA DELAYED-VALIDATION ONLINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

An audit can miss a deployment failure; a user report can reveal it, but can also be manipulated. We study this tension as online decision-making for nonstationary entities under strategic multi-source evidence and delayed, correctable outcomes. AI-agent reputation is one instance: the learner estimates an instance or version's current task reliability before outcomes mature. We propose FARR-SC, which separates bounded use of a report from credit earned by its source. Current reports inform the estimate; matured validation and exact correction update persistent source credit, specialist losses, and calibration. For its fixed-anchor capped-evidence problem, we prove matching strategic-mass bounds with $\Theta(\sum_t \varepsilon_t)$ dependence, alongside implementation-aligned fixed-specialist and realized-path calibration guarantees. With one frozen head, causal replays favor FARR-SC over a descriptive envelope of predeclared non-FARR-SC baselines on four of six traces in each of two blindspot regimes under adaptive attack; these envelope comparisons are descriptive, and conservative baselines win all six clean cells. On two public trust networks, the default has lower decision loss than every non-FARR-SC baseline, although its no-source ablation is lower still. Source scores learned without fraud-label supervision achieve fraud-source AUC $0.91/0.85$. Together, these results identify where validated feedback aids recovery and where conservative operation remains preferable.

1 INTRODUCTION

A deployed agent can change while its latest audit remains unchanged. A new version, task, or context may expose a failure that an earlier check missed. User reports offer timely evidence, but their sources may be unreliable or coordinated, and the outcome needed to judge a report arrives later. The platform must nevertheless decide whether to list, review, or delist the agent, with a continuous deposit multiplier alongside that choice. Reputation here means the current task reliability of an instance or version, not a ranking of model families.

The resulting learning problem is broader than the platform: estimate a changing entity's quality from strategic multi-source reports, act before validation, and revise learned credit when an outcome is corrected. Ignoring reports closes the feedback attack surface but leaves audit blindspots unresolved. Using report agreement to establish source credibility lets coordinated sources reinforce one another. Dynamic truth discovery studies evolving truth and source reliability (Li et al., 2015; 2017; Yao et al., 2018); our focus is the additional constraint imposed when source credit must be earned from later validation.

A report may inform the current estimate without earning its source additional future credit. FARR-SC (Frequency-Aware Robust Reputation with delayed-validation Selector and Calibration) implements this distinction: clipping and caps bound current report use, while matured outcomes and exact correction update source reliability, validation-earned capacity, specialist credit, and calibration. In this sense, *influence over reputation must be purchased with validated work*. Shrinkage, robust aggregation, delayed Hedge, and calibration provide the component tools; the mechanism specifies which observations may change which states, and when.

This separation has a quantitative consequence. Under explicit source-creation and validation costs, warmup caps turn attacker spending into a per-round bound on retained strategic mass (Appendix A.4). Matching evidence-layer upper and lower bounds characterize its effect on estimation.

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.

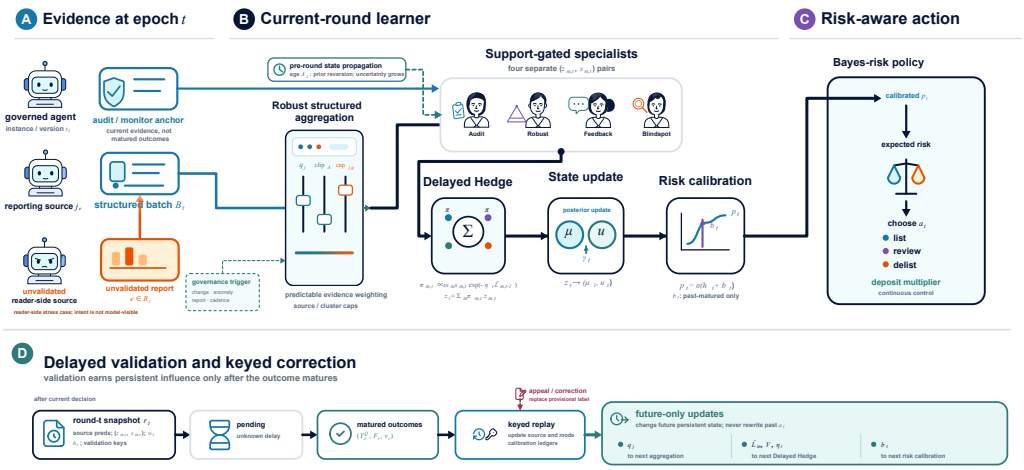


Figure 1: FARR-SC casts dynamic reputation as delayed-validation online platform decision learning. Bounded structured evidence is aged, source-weighted, capped, and combined by delayed specialist Hedge. A separate policy head maps reputation into cost-sensitive actions. Only matured validation and exact appeal replacement update source, specialist-loss, and calibration state.

Separate guarantees cover the implemented selector and calibrator. The theoretical contribution is therefore a bound on strategic influence together with an analysis of adaptation, with each result tied to its own observable quantities and comparator.

Our expanded grid ruled out universal superiority: conservative bandits win clean operation, while feedback recovers some audit blindspots. We evaluate that trade-off with one frozen head, using controlled mechanism tests, external agent traces, a recruited-human pilot on the deployed website workflow, and public trust networks. The experiments test three claims: whether feedback helps, which components carry that benefit or limit manipulation, and how far the findings transfer.

2 Problem Formulation

Round t is an event-triggered decision epoch. The learner observes a structured evidence batch B_t for entity i_t; in the deployment application, the entity is an agent instance/version. Trigger scheduling and audit allocation are exogenous: an epoch is neither every service call nor a fixed-time tick. Each item $e \in B_t$ has a source j_e, source cluster g_e, event type τ_e, bounded score $x_e \in [0,1]$, confidence $a_e \in [0,1]$, and timestamp. Only bounded structured ratings, audits, monitors, or task outcomes update reputation; free-form comments may support confidence or later adjudication.

Let $\mathcal{F}_t^0$ denote clean pre-action history. The quantity to be estimated is current expected quality,

$$\theta_t := \mathbb{E}[Y_t^Q \mid \mathcal{F}_t^0, i_t, \text{current context}] \in [0,1].$$

The learner maintains an entity state (μ_i, u_i): μ_i estimates this conditional mean and u_i records uncertainty. Source reliabilities q_j, matured specialist losses $\widehat{L}_m$, and calibration intercept b are separate learned states. The implementation uses

$$\mathcal{M} = \{\text{audit}, \text{robust}, \text{feedback}, \text{blindspot}\}.$$

After observing B_t, the learned discrete action is $a_t \in \{\text{list}, \text{review}, \text{delist}\}$, with a continuous deposit multiplier applied alongside it; restrict, slash, and compensate are downstream trace actions rather than separately optimized learner choices. A delayed validation yields $(Y_t^Q, F_t, v_t) \in [0,1] \times \{0,1\} \times [0,1]$: Y_t^Q supervises reputation and specialists, whereas F_t supervises failure calibration and platform loss. Thus θ_t is a conditional mean, Y_t^Q is an observed quality target, and F_t is a binary failure target; their relationship is specified by each trace protocol. Corrections replace either target and replay its keyed contributions.

Platform decision loss is the deployment objective and the empirical endpoint. The online analysis uses the two comparators actually optimized by the implementation. For a fixed specialist m,

$$\mathrm{Reg}_{\mathrm{sel},T}(m) = \sum_{t \in \mathcal{V}_T} v_t s_{m,t} [\ell(z_t, Y_t^Q) - \ell(z_{m,t}, Y_t^Q)], \qquad \ell(z,y) = (z-y)^2,$$

and calibration is compared with a fixed intercept on the realized base-logit sequence. These sequential comparators match the two implemented updates. Platform-risk consequences enter through an explicit estimation-to-decision transfer.

3 FARR-SC

Algorithms 1 and 2 in Appendix B give the full online update and delayed-validation replay loops. The current batch changes the estimate through bounded aggregation; later validation changes the credit assigned to its sources and specialists. Appendix E specifies the action costs.

Frequency-aware state propagation. Agents receive evidence at uneven rates. Rather than applying a fixed wall-clock decay, FARR-SC uses expected-update age

$$A_i(s,t) = \int_s^t \lambda_i(u)\, du,$$

where λ_i estimates the agent's update/event rate. Reputation reverts toward a prior and uncertainty grows as a function of A_i, so an inactive high-frequency agent and an inactive low-frequency agent are not treated identically.

Source reliability and caps. Source reliability is learned only from delayed validation or corrected labels. For source j,

$$q_j(T) = \frac{K_{\mathrm{type}} q_{\mathrm{type}} + S_j(T)}{K_{\mathrm{type}} + N_j(T)},$$

where S_j is confidence-weighted reward, with implemented per-source reward $[1 - (p_{j,t} - Y_t^Q)^2/0.25]_{[0,1]}$, and N_j is effective validation mass. The prior strength uses

$$K_{\mathrm{type}} = \max(K_{\mathrm{EB,type}}, K_{\mathrm{sec,type}}),$$

combining empirical-Bayes shrinkage with a security lower bound. Source caps grow only with effective validation mass, while fixed type-aware cluster caps bound coordinated mass, preventing new or Sybil sources from obtaining large influence after a few easy successes.

Robust structured aggregation. Each evidence item receives a predictable weight from confidence, source reliability, context, and type priors. The weighted batch is clipped around audit and robust anchors, then source and cluster mass caps are applied. This converts raw strategic feedback volume into bounded effective adversarial mass.

Delayed specialist selector. FARR-SC computes candidate predictions $z_{m,t}$ and predictable support gates $s_{m,t}$ for its specialists. Let $\widehat{L}_{m,t-1}$ be mode m's confidence-weighted squared loss accumulated only over quality targets that matured before round t, and let V_{t-1} be matured validation mass. The implemented selector is

$$\eta_t = \mathrm{clip}\left(\eta_0 \sqrt{\frac{2\log|\mathcal{M}|}{\max(1, V_{t-1})}}, \eta_{\min}, \eta_{\max}\right), \qquad \pi_{m,t} \propto \omega_m s_{m,t} \exp(-\eta_t \widehat{L}_{m,t-1}).$$

Pending outcomes contribute no loss to the ledger. The current batch can change support $s_{m,t}$, while past matured losses determine the credit multiplying that support. Audit-only remains supported when feedback is unsafe.

Policy head: actions, calibration, and shielding. The platform maps reputation to a failure probability

$$p_t = \sigma(h(\mu_t, u_t) + b_t),$$

where b_t is updated only on matured failure labels F_t. Actions minimize estimated Bayes risk under an explicit platform cost model (Appendix E). The calibrated delayed-Hedge head is the single default operating point: every headline number reports the artifact model `farr_delayed_hedge_calibrated`. Removing calibration (the uncalibrated selector) and adding a conservative shield are the two directions a deployment can move along that frontier, toward blindspot recovery or toward clean guardrails respectively; both are quantified as ablations and operating points in Appendices F and I.

Appeal correction replay. When an appeal replaces a provisional outcome, FARR-SC subtracts the affected quality-target and failure-label contributions and inserts corrected values with adjudication confidence, updating source, specialist, and calibration ledgers. Each validation key contributes once, using its current adjudicated value.

4 THEORY

The mechanism has two mathematically distinct layers. At the evidence layer we characterize the minimax effect of capped strategic mass; at the adaptation layer we analyze the selector and calibrator against the sequential comparators they optimize. Full statements and proofs are in Appendix A. Let θ_t denote latent deploy-time reliability, $n_{\mathrm{eff},t}$ the honest effective sample size after weighting, and ε_t the normalized strategic mass that survives source and cluster caps.

Assumption 1 (Causal capped evidence)**.** *All persistent trust states, cap schedules, and prior strengths at round t are measurable with respect to past matured information. Item weights and support gates may additionally use the current model-visible batch, but are fixed before (Y_t^Q, F_t) is observed and never use strategic status or future outcomes. Honest scores are bounded, conditionally unbiased up to β_t, and sub-Gaussian with scale σ. Source caps are applied independently; a common cluster cap does not increase when strategic items are added; strategic status is never model-visible.*

Theorem 1 (Minimax-tight strategic-mass dependence)**.** *Let $\widehat{\theta}_t^0$ be a fixed-anchor bounded aggregate of the clean capped evidence and $\widehat{\theta}_t^a$ the same aggregate after adding source-disjoint strategic evidence of post-cap mass $\varepsilon_t \le 1/2$. Then*

$$\mathrm{TV}(P_t^a, P_t^0) \le \varepsilon_t, \qquad |\widehat{\theta}_t^a - \theta_t| \le \beta_t + \sigma\sqrt{\frac{2\log(2T/\delta)}{n_{\mathrm{eff},t}}} + b_{\mathrm{rob},t} + \varepsilon_t$$

simultaneously with probability at least $1-\delta$. Conversely, for every online estimator and every sequence $\varepsilon_t \in (0, 1/2]$, fresh audit-blindspot contexts exist for which

$$\sup_{\xi \in \{+,-\}^{\mathcal{V}_T}} \mathbb{E} \sum_{t \in \mathcal{V}_T} |\widehat{\theta}_t - \theta_t^{\xi_t}| \ge \sum_{t \in \mathcal{V}_T} \frac{\varepsilon_t}{2(1-\varepsilon_t)} \ge \frac{1}{2} \sum_{t \in \mathcal{V}_T} \varepsilon_t.$$

Thus the strategic-mass dependence is exactly $\Theta(\sum_t \varepsilon_t)$; it is not slack introduced by clipping analysis.

Corollary 1 (Decision-risk transfer)**.** *If calibrated failure probability is L_p-Lipschitz in the estimate and action risk is L_R-Lipschitz in that probability, a plug-in Bayes action has per-round excess risk at most $2L_R L_p|\widehat{\theta}_t - \theta_t|$. For the converse, suppose a neighborhood of $\theta^\star$ maps monotonically across an interior boundary $p^\star$ with lower slope $\lambda_p > 0$, two actions are the active Bayes-optimal pair on opposite sides, their risk gap is at least $\kappa_p|p - p^\star|$, and every other action remains separated there. Then a shifted hard subclass incurs $\Omega(\kappa_p \lambda_p \sum_t \varepsilon_t)$ excess risk. Appendix E verifies the cost-gap condition with $\kappa_p \ge 6$; the link condition holds locally whenever an active boundary has an unclipped interior preimage.*

The cap schedule turns attacker spending into a *per-round standing-mass* bound, $\varepsilon_t \le B_T L_{\mathrm{cap}}/(B_T L_{\mathrm{cap}} + W_{\min}) < 1$ (Appendix A.4); it does not claim that one-time warmup spending also bounds $\sum_t \varepsilon_t$, because a warmed source can be reused. Raw rating volume never enters

Table 1: Evidence sources and their roles. Execution provenance, mechanism tests, and transfer evaluation answer different questions.

Source	Role	Evaluation scope
Governed workflow	Audits, feedback, and appeals	Status-blinded controlled replay
Local-chain execution	Contract events and provenance	Local execution
LinuxArena extra8	Repeated-agent feedback utility	External task outcomes
cx-cmu trajectories	Repeated-agent robustness	External service trajectories
Chatbot Arena	Same-model-over-time feedback	Human preferences, limited actions
Synthetic trajectories	Quality error, sparsity, scale	Known simulated dynamics
Recruited-human pilot	Deployed website workflow	Controlled traffic, accelerated clock
Bitcoin OTC / Alpha	Frozen-head transfer and source AUC	Public trust ratings; fraud labels withheld from learning

either statement. On a matched-gain fixed-monitor robust path, Appendix 2 further shows that state propagation preserves this linear dependence under an explicit contraction factor.

Theorem 2 (Implementation-aligned adaptation)**.** *For every fixed specialist m, the frozen four-mode selector satisfies*

$$\mathrm{Reg}_{\mathrm{sel},T}(m) \le \frac{\log(1/\omega_m)}{\eta^{\downarrow}_{N+1}} + \frac{1}{8}\sum_{r=1}^{N} \eta^{\downarrow}_r + D^{\mathrm{sel}}_T + A^{\eta}_T + S_T(m) + C_{\mathcal{M}}(T),$$

where $N = |\mathcal{V}_T|$, $\eta^{\downarrow}$ is the nonincreasing envelope of the implemented learning rate, D^{sel}_T is stale-ledger error, A^{η}_T is learning-rate variation from replay, $S_T(m)$ is the exact fractional-support normalization penalty, and $C_{\mathcal{M}}(T)$ is corrected or corrupted mode-loss mass. The implemented calibration update separately has $O((B^2+1)\sqrt{V_T})$ surrogate regret against a fixed intercept on its realized base-logit sequence.

Before its numerical floor is active, the selector's first two terms recover the familiar $O(\sqrt{N \log |\mathcal{M}|})$ scale; after that point the bound explicitly prices continued adaptation rather than silently dropping the floor. Exact correction replay reconstructs both ledgers, so the same statements hold on the corrected sequence.

Together, Theorems 1 and 2 give a deliberately modular guarantee: the nonstandard strategic term has matching $\Theta(\sum_t \varepsilon_t)$ dependence up to universal constants, while adaptation is analyzed against the comparators the implementation actually optimizes. A full-policy result would add separate terms for hard activation changes and path-dependent ledger memory.

5 Experiments

Evaluation questions. *Q1: when does feedback help?* Figure 2 compares one frozen head with the selected non-FARR-SC baseline in each causal cell. *Q2: which mechanism components limit manipulation or enable recovery?* Figure 3 isolates source learning and post-cap strategic mass; correction and validation-corruption checks are reported below and in the appendix. *Q3: does the mechanism transfer beyond our traces?* The deployed-workflow pilot and public trust-network replays test transfer to recruited-human feedback and independently collected trust ratings.

Evidence stack. Table 1 separates algorithmic evidence from systems provenance. The causal grid covers six repeated-entity traces; the 20K trace tests scale; the recruited pilot exercises a deployed website workflow under controlled traffic; and Bitcoin OTC/Alpha test zero-tuning transfer under naturally occurring manipulation. Local-chain evidence establishes execution and provenance; algorithmic comparisons use the frozen causal replays.

Threat model and scope. Attackers manipulate structured feedback only; they know the mechanism, mint fresh sources (Sybil, whitewashing), coordinate clusters (collusion), spam appeals, and, in the adaptive mode, run sleeper sources that behave honestly through warmup before flipping. Attack volume scales with a budget parameter; corrupted validation is a bounded fraction of validation

Table 2: Causal regime summary at attack scale 1. Both blindspot regimes include the same combined adaptive feedback attack; they differ in whether the audit/outcome mismatch is injected or trace-native. Δ is paired selected-baseline loss minus the frozen default's loss; positive favors FARR-SC. The baseline is the lowest-mean non-FARR-SC variant over eight predeclared replay seeds; its interval is conditional on that selection and descriptively summarizes replay-seed variation.

Regime	Cells	Interval wins	Median Δ
Clean	6	0/6	-0.249
Injected audit blindspot	6	4/6	$+1.379$
Natural blindspot	6	4/6	$+0.697$
Broad combined attack	6	1/6	-0.296
Blindspot, 5–20% validation corruption	18	12/18	$+1.48$ to $+1.60$
20K injected blindspot	1	1/1	$+2.914$

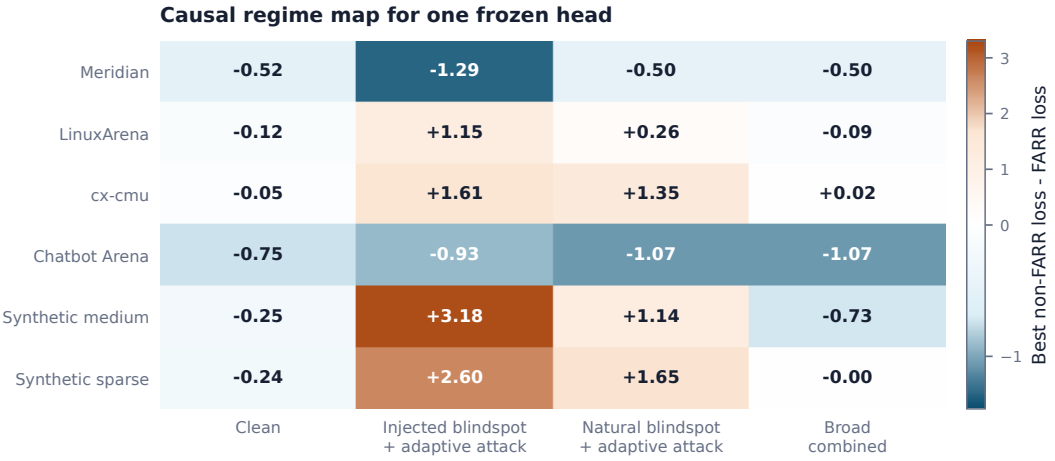


Figure 2: Causal regime map for the frozen default at attack scale 1. Cells show paired selected-baseline loss minus FARR-SC loss over the same eight seeds. Both blindspot columns add the same combined adaptive feedback attack. The gain appears in four of six cells in each blindspot regime, while every clean cell and both Meridian/Chatbot blindspot cells favor a baseline.

confidence mass; exact generators are in Appendix D. We evaluate robustness, not attribution: the guarantee bounds the damage of any strategic feedback within capped mass, malicious or merely wrong, while attribution and enforcement belong to the appeal/adjudication channel and are out of scope.

Baselines and metrics. We compare against audit-only, Beta reputation (Jøsang & Ismail, 2002), Elo-style (Elo, 1978) and Kalman (Kalman, 1960) trackers, PageRank-style trust (Page et al., 1999), EigenTrust (Kamvar et al., 2003), TrustRank (Gyöngyi et al., 2004), online matrix-factorization (Koren et al., 2009) and NCF-style (He et al., 2017) reputation models, UCB (Auer et al., 2002) and Thompson (Thompson, 1933) validation bandits, robust and capped feedback aggregators, RDR (our robust dynamic-reputation ablation baseline: the same clipping and caps but no delayed-validation selector), and ablations without source learning (Appendix F). All models use causal outcome maturity. The primary FARR-SC head and its hyperparameters are fixed across traces, scenarios, costs, and attacks; baseline configurations are fixed except for the predeclared UCB/Thompson fairness sweep. Each cell reports the lowest-mean non-FARR-SC variant over eight seeds; its interval is conditional on that selection and descriptive, without selection adjustment. Primary metrics are mean platform loss, failure-event non-delisting rate ($F_t = 1$, action list or review), non-failure-event delisting rate ($F_t = 0$), review rate, attack-induced loss delta, and synthetic true-quality absolute error. The artifact fields `badAgentAllowedRate` and `goodAgentDelistedRate` denote these event-level rates, not source-fraud detection rates.

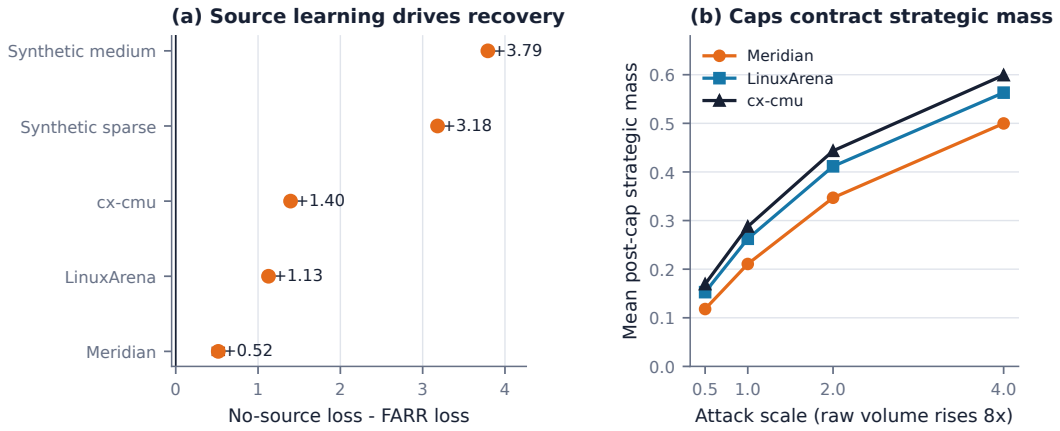


Figure 3: Mechanism evidence under causal timing. (a) A matched *non-adversarial* audit-blindspot ablation changes only source learning; all five intervals exclude zero. This differs from the attacked blindspot regimes in Figure 2. (b) Raw strategic volume increases eightfold from attack scale 0.5 to 4, while source/cluster caps keep mean post-cap normalized mass below 0.61.

Where feedback helps. The benefit is regime-dependent (Figure 2, Table 2). Both blindspot settings favor the fixed head on LinuxArena, cx-cmu, and the synthetic traces; Meridian and Chatbot Arena favor a baseline. Every clean cell favors a conservative baseline, and broad combined attack leaves one win, one overlapping interval, and four baseline wins. Thus audit/outcome mismatch alone does not establish that feedback will help. The predeclared tuned-bandit envelope preserves the same 2/3 win split on the three real/semi-real traces; this within-regime selection is descriptive. Seed-based intervals summarize the fixed eight replay seeds, not population-level uncertainty (Appendix F).

What carries the gain. In the matched non-adversarial blindspot ablation, removing source learning raises default-head loss by 0.52, 1.13, 1.40, 3.79, and 3.18 on Meridian, LinuxArena, cx-cmu, synthetic medium, and synthetic sparse respectively; every paired difference is positive on 8/8 seeds. Component tests separate roles: frequency aging and learned mode weights improve recovery; calibration selects a decision operating point; warmup growth contains broad attacks; and a concentrated-collusion diagnostic, unlike the primary matrix where source caps bind first, shows the cluster cap prevents an additional $0.42/0.03/0.20$ loss on Meridian/LinuxArena/cx-cmu (Appendix F). Across the strategic-mass grid there are no cap-order violations and the default's fitted loss slope is positive in all nine trace–attacker groups (eight intervals exclude zero), although 20 of 216 adjacent scale transitions violate pointwise monotonicity.

Correction, corrupted validation, and costs. Exact correction replay matches oracle source, specialist-loss, and calibration state at zero lag. With 20% provisional corruption and lag 64, final source error is 0, specialist-weight ℓ_1 error is 0.021, and calibration-intercept error is 0.006; all eight seeds return to the preregistered oracle neighborhood. Under 5–20% corrupted validation, the same four of six blindspot traces win at every level, while clean and broad-combined cells remain trade-offs. Across the 3×3 cost grid, the default wins all nine LinuxArena blindspot cells and no Meridian blindspot cells; clean winners change with costs. These checks support state consistency and regime stability, not a claim that every correction or cost matrix improves immediate loss (Appendix E).

Scale and frozen-parameter checks. On the eight-seed 20K trace, the fixed head beats the strongest baseline by 2.914 in the injected blindspot and by 0.0559 in broad combined attack, while losing clean traffic by 0.298. Across 54 preregistered neighboring-parameter cells, the frozen configuration is within 1% of the best tested setting in 32 cells and within 5% in 52; the median excess is 0.73%. These neighboring settings diagnose sensitivity and never replace the default row by row.

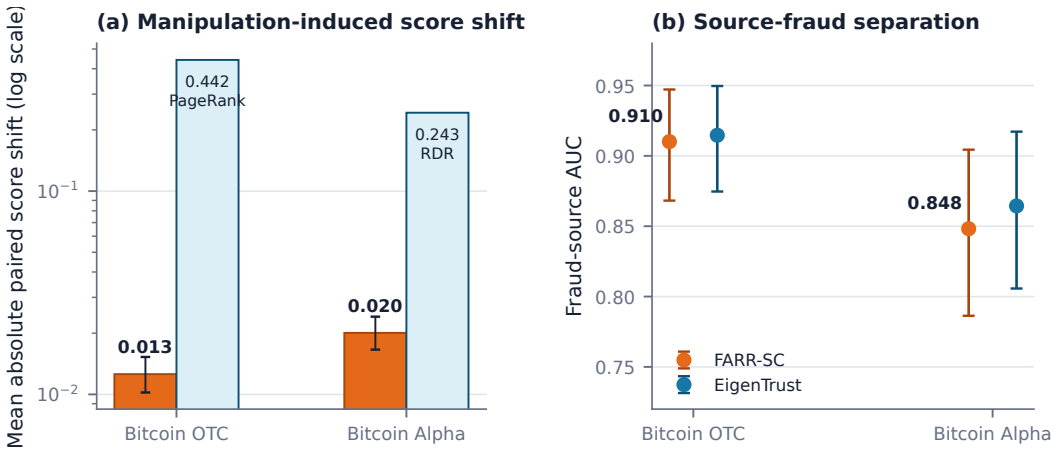


Figure 4: Zero-tuning external generalization on public networks with real fraud labels. (a) Mean absolute paired score movement between clean and attacked replays; bars use a log scale. (b) Source-fraud AUC from reliability learned without fraud labels. Intervals use entity-block and source-level bootstrap respectively.

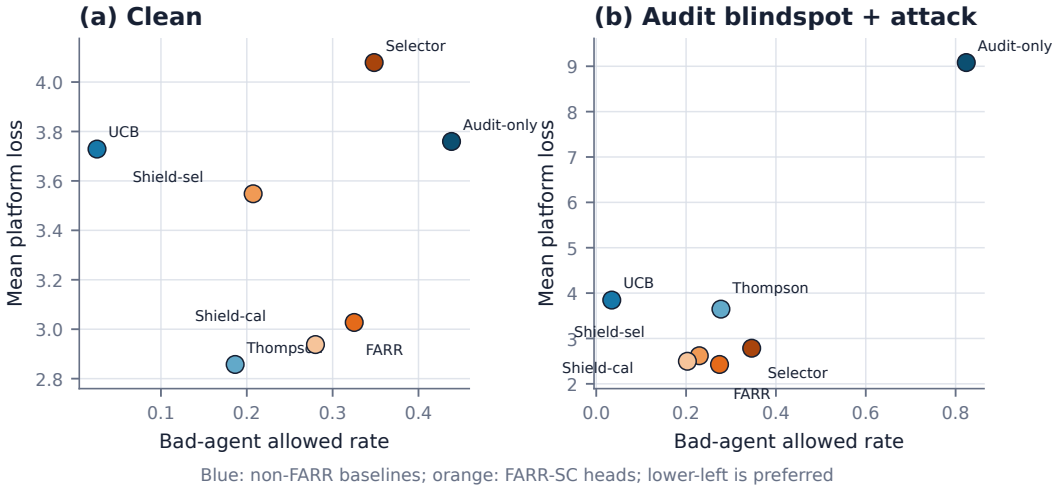


Figure 5: Causal operating frontier, macro-averaged over six traces and eight seeds; lower-left is preferred. Removing calibration and adding a shield expose fixed alternative operating points.

Recruited-volunteer market-pilot replay. The platform workflow is deployed as a live website, but launch-stage organic traffic was too sparse for a powered evaluation. We therefore recruited controlled traffic to exercise the deployed workflow. The 10,129-row export contains 7,062 recruited-human activity rows; conversion retains 4,603 native structured feedback items at 204 definitive audit anchors, while 1,001 platform shadow-score rows remain metadata only. The frozen head's overall loss is 1.6294 versus 1.7321 for the best non-FARR-SC baseline (EigenTrust). The 25 red-team source rows map to 23 anchors; on those cases the frozen head has the lowest point estimate, 0.477 [0.332, 0.684], but intervals overlap. This is a no-leakage consistency check on real feedback, not a powered victory claim (Appendix G).

External generalization to public manipulation data. We froze every parameter and replayed Bitcoin OTC (35,592 ratings) and Alpha (24,186); the "attacked" counterpart is an all-edges counterfactual, including fraud-source edges, not an observed adaptive attack. Fraud labels define the clean counterpart and assess source reliability; future consensus supplies validation and platform-loss targets. The default beats every non-FARR-SC baseline in all four platform-loss replays (Alpha clean/attacked $1.787/2.002$ versus RDR $3.809/3.786$; OTC $2.594/2.806$ versus RDR/PageRank

$3.727/3.827$). Its paired score shifts are 0.0126 [0.0102, 0.0153] and 0.0201 [0.0166, 0.0241], smallest among all models. For attacked-minus-clean contrasts, score shifts use all matched identifiers, but a post-run label-alignment correction retains only rows with unchanged delayed label/severity (21,425 OTC/14,201 Alpha rows; 301/263 identifiers excluded). It changes no replay output, model state, trace, or score. The retained loss deltas are -0.131 $[-0.211, -0.058]$ and -0.0096 $[-0.049, 0.021]$, with Alpha not best relative to its baseline (Figure 4). Without fraud-label supervision, source AUC is 0.910 [0.868, 0.947] on OTC, compatible with EigenTrust's 0.915, and 0.848 [0.786, 0.904] on Alpha, below EigenTrust's 0.864. The no-source ablation has lower loss in all four cells. Source-fraud separation and decision-loss improvement are therefore distinct findings: the former transfers to these networks, while a benefit from source learning on decision loss is established by the blindspot ablation, not by these four replays (Appendix J).

Operating-point trade-off. Low decision loss need not mean fewer failure events retained. On the public networks, the default does not delist 0.987–0.999 of events labeled as failures by delayed senior consensus (list or review), while delisting 0.8–1.6% of non-failure events. Robust non-FARR-SC baselines retain 0.29–0.46 and delist 28.5–31.8%, respectively. These are event-level decisions under the consensus target, not rates of blocking fraud-labeled users. Figure 5 shows the loss–retention trade-off across fixed operating points on the six-trace grid; deployments with different harm costs require their own policy calibration.

6 Related work

Dynamic truth discovery is the closest estimation problem. Li et al. (2015) update truth and source reliability incrementally with temporal smoothing and decay; Li et al. (2017) adapt the frequency of source-weight assessment over streams; and Yao et al. (2018) couple truth estimation with seasonal time-series prediction. These formulations learn source weights from agreement with inferred truths. FARR-SC instead scores stored source predictions against later validation, ties capacity growth to matured validation mass, and replaces credit when outcomes are corrected. The distinction concerns the signal that earns source credit, not whether truth or reliability can change over time.

Entity trackers and graph propagation provide complementary reputation primitives (Jøsang & Ismail, 2002; Kamvar et al., 2003; Glickman, 1999; Herbrich et al., 2007). Robust and Sybil-resistant aggregation motivates clipping and caps (Huber, 1964; Yu et al., 2006; Diakonikolas et al., 2019); review- and graph-fraud methods identify manipulation (Rayana & Akoglu, 2015; Hooi et al., 2016; Kumar et al., 2018; 2016). FARR-SC bounds the influence of reports without requiring a malicious-source classification. Its adaptation and policy layers use delayed experts, sleeping specialists, and calibration (Joulani et al., 2013; Freund et al., 1997; Kleinberg et al., 2010; Foster & Vohra, 1998; Zadrozny & Elkan, 2002). Agent benchmarks and preference platforms measure task performance or comparative preference (Liu et al., 2024; Jimenez et al., 2024; Chiang et al., 2024). Our deployment setting adds repeated cost-sensitive actions with delayed and correctable outcomes, connecting to online mechanism design and truthful elicitation (Parkes & Singh, 2004; Jurca & Faltings, 2009; Dekel et al., 2010). Table 3 compares the learning signals and objectives of these adjacent methods.

7 Limitations

Deployment. Launch-stage organic traffic is too sparse for a powered prospective study; the controlled pilot, incomplete-action traces, and non-agent public networks remain the main gap. *Costs.* Decision loss uses a sensitivity-tested matrix and must be recalibrated per platform. *Theory.* Strategic-mass dependence is minimax tight for capped evidence, whereas selector/calibration guarantees use realized sequential comparators; full-policy closure must price activation, ledger memory, and a joint decision surrogate.

8 Conclusion

FARR-SC separates decision evidence from validation-earned source credit; strategic-mass bounds quantify its cost and causal replays locate gains in selected blindspots. Clean baselines and public-network ablations expose the resulting trade-offs.

ETHICS STATEMENT

All recruited pilot participants were informed before agreeing to participate that the activity was a controlled evaluation of the deployed platform workflow, what data would be collected, and how those data would be used. Participants assigned to red-team tasks were additionally informed of that role; every participant was compensated. The public artifact contains only de-identified or internal identifiers and aggregate replay results. It excludes names, contact fields, raw prompts, credentials, payment identifiers, private adjudication notes, and raw tool input/output. Controlled red-team probes are explicitly separated from ordinary feedback and are never presented as organic marketplace attacks. The authors followed the applicable ethical and institutional requirements and take responsibility for the study conduct, analysis, and release boundary.

REPRODUCIBILITY STATEMENT

All quantitative results use one deterministic harness and causal validation maturity. The package contains 1,245 completed condition/seed tasks with zero failures: 680 mechanism and correctness tasks, plus 565 confirmatory tasks spanning a 384-task six-trace grid, tuned-baseline fairness, eight-seed 20K scale, cost sensitivity, the recruited pilot, and public-network replays. Paired statistics match trace, seed, scenario, attack scale, attacker mode, corruption, costs, and track before comparison. Human-pilot uncertainty uses 20,000 case bootstraps; public-network uncertainty uses 5,000 entity-block and 20,000 source bootstraps rather than repeated deterministic seeds. The release audit normalized converter-specific score and threshold conventions and reran every affected branch end to end; no pre-correction and corrected performance rows are mixed. The supplement includes code, frozen protocols, sanitized traces, per-task status, summaries, raw per-source outputs, figure source, and a claim-to-artifact registry. Local-chain legacy material is retained only for non-overlapping execution provenance; overlapping immediate-update results are excluded. Algorithms, costs, and attacks are specified in Appendices B, E, and D.

AI USE STATEMENT

We used generative AI tools for information retrieval and preliminary discovery of potentially relevant prior work, and to generate synthetic datasets for controlled experiments. The authors independently formulated the research questions, specified the synthetic-data procedures, designed and executed the experiments, developed and reviewed the code, created the figures, interpreted the results, and wrote the final manuscript. The authors checked the generated data and independently verified all citations, reported numbers, derivations, and final claims, and take responsibility for the paper, code, and artifacts; no language model is an author.

REFERENCES

Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. In *International AAAI Conference on Web and Social Media*, 2013.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 2002.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, 2011.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica.

Chatbot arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning*, 2024.

Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8), 2010.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2), 2019.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco, 1978.

Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. Poisoning attacks to graph-based recommender systems. In *Annual Computer Security Applications Conference*, 2018.

Michal Feldman, Christos Papadimitriou, John Chuang, and Ion Stoica. Free-riding and whitewashing in peer-to-peer systems. In *ACM SIGCOMM Workshop on Practice and Theory of Incentives in Networked Systems*, 2004.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2), 1998.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the ACM Symposium on Theory of Computing*, 1997.

Mark E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C*, 1999.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, 2019.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with TrustRank. In *Proceedings of the International Conference on Very Large Data Bases*, 2004.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the International Conference on World Wide Web*, 2017.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, 2007.

Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: Bounding graph fraud in the face of camouflage. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 1964.

Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. CatchSync: Catching synchronized behavior in large directed graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

Audun Jøsang and Roslan Ismail. The beta reputation system. In *Proceedings of the Bled Electronic Commerce Conference*, 2002.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *International Conference on Machine Learning*, 2013.

Radu Jurca and Boi Faltings. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34, 2009.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 1960.

Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the International Conference on World Wide Web*, 2003.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80, 2010.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

Srijan Kumar, Francesca Spezzano, V.S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *IEEE International Conference on Data Mining*, 2016.

Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and V.S. Subrahmanian. REV2: Fraudulent user prediction in rating platforms. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2018.

Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. In *Advances in Neural Information Processing Systems*, 2016.

Tianyi Li, Yu Gu, Xiangmin Zhou, Qian Ma, and Ge Yu. An effective and efficient truth discovery framework over data streams. In *Proceedings of the International Conference on Extending Database Technology*, pp. 180–191, 2017. doi: 10.5441/002/edbt.2017.17.

Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 675–684, 2015. doi: 10.1145/2783258.2783277.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2), 1994.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024.

Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, 62(12), 2016.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the ACM Symposium on Theory of Computing*, 2018.

Dina Mayzlin, Yaniv Dover, and Judith Chevalier. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8), 2014.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

David C. Parkes and Satinder Singh. An MDP-based approach to online mechanism design. In *Advances in Neural Information Processing Systems*, 2004.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, 2018.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Communications of the ACM*, 43(12), 2000.

Shai Shalev-Shwartz. *Online Learning and Online Convex Optimization*. Foundations and Trends in Machine Learning, 2012.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4), 1933.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, 2024.

Liuyi Yao, Lu Su, Qi Li, Yaliang Li, Fenglong Ma, Jing Gao, and Aidong Zhang. Online truth discovery on time series data. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 162–170, 2018. doi: 10.1137/1.9781611975321.19.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 2018.

Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. Sybilguard: Defending against sybil attacks via social networks. In *ACM SIGCOMM*, 2006.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

A Theory Details

This appendix proves Theorems 1 and 2. All scores are normalized to $[0, 1]$.

A.1 Online setting

Let $(\mathcal{F}_t)_{t\geq 0}$ be the platform filtration immediately before choosing an action at round t. The platform receives a bounded structured evidence batch

$$B_t = \{e = (j_e, g_e, \tau_e, x_e, a_e)\},$$

where j_e is a source, g_e is a source cluster, $x_e \in [0, 1]$ is a structured score, and $a_e \in [0, 1]$ is evidence confidence. Let $\mathcal{F}_t^0$ denote clean pre-action history. The latent deploy-time reliability is the clean conditional quality mean $\theta_t := \mathbb{E}[Y_t^Q \mid \mathcal{F}_t^0, i_t, \text{current context}] \in [0, 1]$. Validation may arrive after a delay. Each matured outcome has a quality target $Y_t^Q \in [0, 1]$, a failure label $F_t \in \{0, 1\}$, and confidence $v_t \in [0, 1]$. The first two targets are kept distinct because reputation/specialist learning and failure calibration solve different supervised problems. Some validation mass may be corrupted.

For mode m, let $z_{m,t} \in [0, 1]$ be its prediction and let $z_t = \sum_m \pi_{m,t} z_{m,t}$. The selector weight $\pi_{m,t}$ must be predictable.

A.2 ROBUST EVIDENCE

For honest evidence $e \in H_t$, assume

$$x_e = \theta_t + b_e + \xi_e, \qquad |b_e| \le \beta_t, \qquad \mathbb{E}[\xi_e \mid \mathcal{F}_t] = 0,$$

and the weighted noise is conditionally σ^2-sub-Gaussian. Let

$$n_{\mathrm{eff},t} = \frac{1}{\sum_{e \in H_t} \alpha_{e,t}^2}$$

be the honest effective sample size after normalizing honest weights.

Let P_t^0 be the normalized clean distribution after source and cluster caps, and let P_t^a be the distribution obtained by adding source-disjoint strategic items and rerunning the same caps. Writing $A_{\mathrm{ret},t}$ for retained strategic weight and Z_t^a for total attacked weight, define $\varepsilon_t := A_{\mathrm{ret},t}/Z_t^a \in [0,1)$; strict inequality follows from the positive clean mass needed to define P_t^0. The cluster cap used for a clean cluster is unchanged when strategic items join it.

Lemma 1 (Cap-to-total-variation contraction)**.** *If the normalized post-cap strategic mass in P_t^a is ε_t, then*

$$\mathrm{TV}(P_t^a, P_t^0) \le \varepsilon_t.$$

Consequently, every fixed $f\colon B_t \to [0,1]$ satisfies $|\mathbb{E}_{P_t^a} f - \mathbb{E}_{P_t^0} f| \le \varepsilon_t$.

Proof. After independent source caps, fix a cluster and write its honest and strategic masses as H and A. If its common cap is C, the clean and attacked cluster multipliers are $c_0 = \min(1, C/H)$ and $c_a = \min(1, C/(H+A))$. The honest mass displaced by adding strategic items is $D = H(c_0 - c_a)$. A direct case split on whether H and $H + A$ exceed C gives $0 \le D \le c_a A$, the strategic mass retained in that cluster. Summing over clusters gives total displaced honest mass at most retained strategic mass A_{ret}. If Z_a is attacked total mass, clean total mass is $Z_0 = Z_a - A_{\mathrm{ret}} + D \le Z_a$. The two normalized distributions therefore share at least $(Z_a - A_{\mathrm{ret}})/Z_a = 1 - \varepsilon_t$ mass, proving the TV bound. The final claim is the dual characterization of total variation for functions in $[0,1]$. □

Let $\widehat{\theta}_t^0 = \mathbb{E}_{P_t^0} f_t$ and $\widehat{\theta}_t^a = \mathbb{E}_{P_t^a} f_t$, where the clipping map and anchor in f_t are fixed from past matured information. Lemma 1 gives $|\widehat{\theta}_t^a - \widehat{\theta}_t^0| \le \varepsilon_t$. The honest weighted noise is sub-Gaussian with variance proxy $\sigma^2 \sum_e \alpha_{e,t}^2$; the deterministic score bias contributes β_t, and clipping/anchoring contributes $b_{\mathrm{rob},t}$. A union bound over t proves the upper half of Theorem 1.

Corollary 2 (Contractive robust-state propagation)**.** *Consider coupled clean and attacked executions of the fixed-monitor robust specialist with the same event schedule and update gain $\gamma_t \in [\gamma_{\min}, \gamma_{\max}]$. If $d_t = |\mu_t^a - \mu_t^0|$, then*

$$d_t \le \rho d_{t-1} + \gamma_{\max} \varepsilon_t, \qquad \rho = 1 - 0.72\gamma_{\min} < 1,$$

and consequently

$$\sum_{t=1}^{T} d_t \le \frac{d_0}{1-\rho} + \frac{\gamma_{\max}}{1-\rho} \sum_{t=1}^{T} \varepsilon_t.$$

For the frozen gain range, $\gamma_{\min} = 0.07$ and hence $\rho \le 0.9496$.

Proof. The monitor anchor is 0.72 monitor score plus 0.28 prior state. Holding the monitor and gain fixed, clipping is nonexpansive in its anchor, and Lemma 1 adds at most ε_t. The convex state update therefore contracts the prior-state difference by at most $1 - 0.72\gamma_t$ and adds at most $\gamma_t \varepsilon_t$. Frequency-aware propagation is itself nonexpansive under the shared event schedule. Iterating the recursion and summing its geometric kernel proves the result. □

A.3 SOURCE RELIABILITY

For source j, let $p_{j,s} \in [0,1]$ be its bounded prediction on validated item s. The implementation stores validation reward

$$r_{j,s} = \left[1 - \frac{(p_{j,s} - Y_s^Q)^2}{0.25}\right]_{[0,1]},$$

confidence v_s, and effective validation mass

$$N_j(T) = \sum_{s \le T : j_s = j} v_s, \qquad S_j(T) = \sum_{s \le T : j_s = j} v_s r_{j,s}.$$

The source estimate is

$$q_j(T) = \frac{K_{\text{type}} q_{\text{type}} + S_j(T)}{K_{\text{type}} + N_j(T)}.$$

Theorem 3 (Source reliability under corrupted validation)**.** *Assume honest validation rewards for source j have conditional mean ρ_j, and let $C_j(T)$ be the total confidence mass of corrupted validation attributed to source j. With probability at least $1 - \delta$,*

$$|q_j(T) - \rho_j| \le \frac{K_{\text{type}}}{K_{\text{type}} + N_j(T)} |q_{\text{type}} - \rho_j| + \frac{N_j(T)}{K_{\text{type}} + N_j(T)} \sqrt{\frac{\log(2/\delta)}{2N_j(T)}} + \frac{C_j(T)}{K_{\text{type}} + N_j(T)}.$$

Proof. Write $q_j - \rho_j$ as a prior-bias term plus $(S_j - N_j\rho_j)/(K_{\text{type}} + N_j)$. Split the latter into honest and corrupted validation mass. The honest part is bounded by Hoeffding or Freedman for bounded martingale differences. Since $r_{j,s}, \rho_j \in [0, 1]$, the corrupted part is bounded by its total confidence mass $C_j(T)$. □

Corollary 3 (Anti-whitewashing warmup)**.** *If a source cap grows as*

$$\text{cap}_j(N) = \text{cap}_{\text{cold}} + (\text{cap}_{\max} - \text{cap}_{\text{cold}}) \frac{N}{N + K_{\text{cap}}},$$

then realizing fraction s of the available increase from cap_{cold} to $\text{cap}_{\max}$ requires

$$N \ge \frac{K_{\text{cap}} s}{1 - s}.$$

A.4 From caps and budget to adversarial mass

Theorem 1 conditions on the post-cap strategic mass ε_t. Here we bound that per-round standing mass from the mechanism. Suppose (i) each source's per-batch weight is capped by the warmup schedule of Corollary 3 and cluster caps only tighten it; (ii) honest evidence (at least the audit/monitor anchor) retains predictable normalized weight at least $W_{\min} > 0$ in every batch; (iii) creating a source that can submit structured feedback costs at least c_{src}, and one unit of effective validation mass N_j costs at least $c_{\min}$, because validation mass accrues only on matured, validated interactions; and (iv) the attacker's total budget through horizon T is B_T.

Theorem 4 (Budget-implied adversarial mass)**.** *Under (i)–(iv), at every round t the capped adversarial weight satisfies*

$$A_t \;\le\; \frac{B_T}{c_{\text{src}}} \text{cap}_{\text{cold}} \;+\; (\text{cap}_{\max} - \text{cap}_{\text{cold}}) \frac{B_T}{c_{\min} K_{\text{cap}}},$$

and hence the normalized adversarial mass obeys

$$\varepsilon_t \;\le\; \frac{A_t}{A_t + W_{\min}} \;\le\; \frac{B_T\, L_{\text{cap}}}{B_T\, L_{\text{cap}} + W_{\min}} < 1, \qquad L_{\text{cap}} := \frac{\text{cap}_{\text{cold}}}{c_{\text{src}}} + \frac{\text{cap}_{\max} - \text{cap}_{\text{cold}}}{c_{\min} K_{\text{cap}}}.$$

Proof. Let the attacker operate k sources with warm validation masses $N_1, \ldots, N_k$ at round t. Per-batch adversarial weight is at most the sum of source caps, and by Corollary 3's schedule $\text{cap}_j(N_j) \le \text{cap}_{\text{cold}} + (\text{cap}_{\max} - \text{cap}_{\text{cold}}) N_j / K_{\text{cap}}$, so $A_t \le k\, \text{cap}_{\text{cold}} + (\text{cap}_{\max} - \text{cap}_{\text{cold}}) \sum_j N_j / K_{\text{cap}}$. The budget constraints give $k \le B_T / c_{\text{src}}$ and $\sum_j N_j \le B_T / c_{\min}$. Normalizing against honest weight at least $W_{\min}$ yields the bound on ε_t. □

The bound explains the attack generators of Appendix D. Whitewashing pays c_{src} per fresh source and every such source is pinned at cap_{cold}; the sleeper attacker pays $c_{\min}$ per unit of warm mass because its warmup phase must survive delayed validation, i.e., perform genuinely acceptable work; and raw review volume is free but never appears in the bound. K_{cap} is the exchange rate between attacker spending and standing reputation influence. This is a bound on each ε_t, not on $\sum_t \varepsilon_t$: a source purchased once can be reused across rounds unless the deployment charges per use or expires stale validation.

A.5 DELAYED SPECIALIST SELECTOR

The implementation runs four candidates, $\mathcal{M} = \{\text{audit}, \text{robust}, \text{feedback}, \text{blindspot}\}$. Let

$$G_{m,t-1} = \sum_{r<t: r \text{ matured before } t} v_r s_{m,r} \ell(z_{m,r}, Y_r^Q), \qquad \ell(z,y) = (z-y)^2.$$

It first forms

$$q_{m,t} = \frac{\omega_m e^{-\eta_t G_{m,t-1}}}{\sum_k \omega_k e^{-\eta_t G_{k,t-1}}}, \quad d_t = \sum_m q_{m,t} s_{m,t}, \quad \pi_{m,t} = \frac{q_{m,t} s_{m,t}}{d_t}, \quad z_t = \sum_m \pi_{m,t} z_{m,t},$$

with the clipped learning rate displayed in the main text. A corrected label subtracts its earlier ledger contribution before adding the replacement, so each validation key contributes exactly once.

For a matured round, set $g_{m,t} = v_t s_{m,t} \ell(z_{m,t}, Y_t^Q) \in [0,1]$. Corrections may reduce total matured mass and make the clipped η_t increase, so let $\eta_t^{\downarrow} = \min_{r \le t} \eta_r$ be its nonincreasing envelope. Let $q_t^{\downarrow}$ be delayed FTRL formed with this envelope and let $\bar{q}_t$ be the corresponding immediate-feedback distribution. Define

$$A_T^{\eta} = \sum_{t \in \mathcal{V}_T} |\langle q_t - q_t^{\downarrow}, g_t \rangle|, \quad D_T^{\text{sel}} = \sum_{t \in \mathcal{V}_T} |\langle q_t^{\downarrow} - \bar{q}_t, g_t \rangle|,$$

and, for fixed specialist m,

$$S_T(m) = \sum_{t \in \mathcal{V}_T} v_t (s_{m,t} - d_t)_+ \ell(z_t, Y_t^Q).$$

Theorem 5 (Confidence-weighted fixed-specialist regret)**.** *For every fixed $m \in \mathcal{M}$,*

$$\sum_{t \in \mathcal{V}_T} v_t s_{m,t} [\ell(z_t, Y_t^Q) - \ell(z_{m,t}, Y_t^Q)] \le \frac{\log(1/\omega_m)}{\eta_{N+1}^{\downarrow}} + \frac{1}{8} \sum_{r=1}^{N} \eta_r^{\downarrow} + A_T^{\eta} + D_T^{\text{sel}} + S_T(m) + C_{\mathcal{M}}(T),$$

where $N = |\mathcal{V}_T|$ and $C_{\mathcal{M}}(T)$ is the total variation introduced into the mode-loss ledger by provisional, corrected, or corrupted validation.

Proof. Convexity gives $v_t d_t \ell(z_t, Y_t^Q) \le \langle q_t, g_t \rangle$. Hence

$$v_t s_{m,t} [\ell(z_t, Y_t^Q) - \ell(z_{m,t}, Y_t^Q)] \le \langle q_t, g_t \rangle - g_{m,t} + v_t (s_{m,t} - d_t)_+ \ell(z_t, Y_t^Q).$$

Summing reduces the first difference to delayed entropy-FTRL on bounded losses. Add and subtract $q_t^{\downarrow}$ and $\bar{q}_t$ to obtain A_T^{η} and D_T^{sel}. The standard nonincreasing-rate entropy-FTRL potential gives

$$\sum_t \langle \bar{q}_t, g_t \rangle - \sum_t g_{m,t} \le \frac{\log(1/\omega_m)}{\eta_{N+1}^{\downarrow}} + \frac{1}{8} \sum_{r=1}^{N} \eta_r^{\downarrow}.$$

Ledger replacement or corruption changes bounded losses by at most its total variation, yielding $C_{\mathcal{M}}(T)$; the remaining positive term is exactly $S_T(m)$. □

The valid comparator is a fixed specialist, not the squared loss of the best deterministic prediction mixture. Before the implementation floor is reached, the entropy-FTRL terms have the usual square-root scale. Once $\eta_{\min} = 0.02$ is active, the generic frozen-implementation bound contains $O(\log(1/\omega_m)/\eta_{\min} + \eta_{\min} N)$; the floor is an engineering choice for continued adaptation and cannot be omitted from an implementation-level theorem. Standard delay bounds may be substituted for D_T^{sel} when a homogeneous delay model is appropriate (Joulani et al., 2013).

A.6 CALIBRATION AND PLATFORM DECISIONS

FARR-SC uses clipped $p_t = \mathrm{clip}(\sigma(h_t + b_t), 0.02, 0.98)$ and projected online gradient descent

$$b_{t+1} = \Pi_{[-B,B]}\left(b_t - \alpha_t v_t (p_t - F_t)\right).$$

For the realized base logit h_t, define the convex surrogate

$$\phi_t(b) = v_t \int_0^b [\mathrm{clip}(\sigma(h_t + u), 0.02, 0.98) - F_t], du.$$

Its derivative is exactly the implemented gradient and has magnitude at most v_t.

Theorem 6 (Realized-path calibration regret)**.** *For every $b^\star \in [-B, B]$ and the implementation's nonincreasing step sizes,*

$$\sum_{t \in \mathcal{V}_T} [\phi_t(b_t) - \phi_t(b^\star)] \le \frac{(2B)^2}{2\alpha_N} + \frac{1}{2} \sum_{t \in \mathcal{V}_T} \alpha_t v_t^2 = O((B^2 + 1)\sqrt{V_T}).$$

Proof. Apply the projected-OGD potential inequality on $[-B, B]$ and sum it with the nonincreasing step schedule. Exact correction replay rebuilds the ordered ledger and reruns the same update, so the statement applies to the final corrected sequence. □

If $a_t = \arg\min_a R(a, p_t)$ and $a_t^\star = \arg\min_a R(a, p_t^\star)$, Lipschitz risk gives

$$R(a_t, p_t^\star) - R(a_t^\star, p_t^\star) \le 2L_R |p_t - p_t^\star|.$$

Theorem 6 compares intercepts on the same realized base-logit sequence; it does not compare a counterfactual specialist trajectory and a different intercept jointly. Accordingly, the decision inequality is used as an explicit plug-in transfer, not to manufacture a joint platform-regret comparator.

A.7 MATCHED LOWER BOUND AND THE BOUNDARY OF GLOBAL CLOSURE

The nonstandard statistical term is capped strategic mass. The next construction matches Lemma 1 on an audit-blindspot subclass: all model-visible fields are identical across the two worlds, and strategic status is evaluation-only.

Theorem 7 (Hidden-subset strategic-mass lower bound)**.** *For every online estimator $\mathcal{A}$ and every $\varepsilon_t \in (0, 1/2]$, there are two admissible worlds with identical complete attacked batches and post-cap strategic mass ε_t, but clean means separated by $\varepsilon_t/(1 - \varepsilon_t)$. Across fresh contexts,*

$$\sup_{\xi \in \{+,-\}^{\mathcal{V}_T}} \mathbb{E}_{\xi,\mathcal{A}} \sum_{t \in \mathcal{V}_T} |\widehat{\theta}_t - \theta_t^{\xi_t}| \ge \sum_{t \in \mathcal{V}_T} \frac{\varepsilon_t}{2(1 - \varepsilon_t)} \ge \frac{1}{2} \sum_{t \in \mathcal{V}_T} \varepsilon_t.$$

Proof. Fix a round and write $\epsilon = \varepsilon_t$. Present the learner with the same post-cap attacked batch in both worlds: score-zero and score-one groups each have normalized mass $1/2$, and all source IDs, clusters, confidences, timestamps, and matured histories are identical. In W^+, a hidden strategic subset of mass ϵ belongs to the score-zero group; in W^-, it belongs to the score-one group. Strategic status is used only to define the clean counterfactual and is not model-visible. Choose source and cluster masses below their caps so this is a valid post-cap batch. Removing the hidden subset gives

$$\theta_t^+ = \frac{1}{2(1 - \epsilon)}, \qquad \theta_t^- = \frac{1/2 - \epsilon}{1 - \epsilon}, \qquad \theta_t^+ - \theta_t^- = \frac{\epsilon}{1 - \epsilon}.$$

In each world, let the delayed quality target Y_t^Q be Bernoulli with the corresponding mean $\theta_t^\pm$; it matures only after the current prediction and is independent across the fresh contexts. Thus these clean means agree with the definition $\theta_t = \mathbb{E}[Y_t^Q \mid \mathcal{F}_t^0, i_t, \text{current context}]$. The learner has the same conditional output distribution in both worlds. Any common prediction has average absolute error at least half the separation. Fresh source contexts and independent signs prevent earlier validation from revealing the next world. Summing and then fixing a sign sequence that attains at least the average proves the claim and completes Theorem 1. □

Corollary 4 (Platform-risk lower bound under an active action margin)**.** *Suppose an interior failure-probability boundary $p^\star$ has preimage $\theta^\star$ and a neighborhood U such that: (i) $p(\theta)$ crosses $p^\star$ monotonically and $|p(\theta)-p^\star| \geq \lambda_p|\theta-\theta^\star|$ on U; (ii) two actions are the unique Bayes-optimal pair on opposite sides and their risk gap is at least $\kappa_p|p-p^\star|$; and (iii) every other action is separated from this active pair by a fixed positive margin on $p(U)$. Then every online platform mechanism incurs expected excess platform risk*

$$\Omega\left(\kappa_p\lambda_p \sum_{t\in\mathcal{V}_T} \varepsilon_t\right)$$

on an admissible audit-blindspot subclass.

Proof. Choose fixed $r > 0$ so that $[\theta^\star - r, \theta^\star + r] \subset U$ and the third-action margin dominates the local pairwise gap throughout this interval. Replace the score-zero and score-one atoms in the proof by $\theta^\star - r$ and $\theta^\star + r$. The two clean means become $\theta^\star \pm r\varepsilon_t/(1-\varepsilon_t)$ while the attacked batch remains identical. The learner therefore has the same action distribution in both worlds. Whichever action it chooses, one world charges at least the active pairwise gap; a third action costs at least its fixed margin. Averaging gives $\Omega(\kappa_p\lambda_p r\sum_t \varepsilon_t)$. Since the active cost boundaries and r are fixed, this has the stated rate. The plug-in upper transfer in Corollary 1 follows by adding and subtracting $R(a(\widehat{p}), \widehat{p})$ and using Bayes optimality at $\widehat{p}$ plus Lipschitzness twice. □

Why a best-mixture comparator is invalid. Take two always-supported specialists with $z_{1,t} = 0$, $z_{2,t} = 1$, and $Y_t^Q = 1/2$ at every round. Their losses are equal, so exponential weights preserve any unequal prior ratio and predict a fixed $p \neq 1/2$. The selector incurs $T(p-1/2)^2$, while the fixed deterministic mixture $(1/2, 1/2)$ predicts $1/2$ and incurs zero. Thus the frozen selector has $\Omega(T)$ regret to the best prediction mixture even without delay or attack. Theorem 5 uses the fixed specialist comparator that the update actually optimizes.

The matched result is intentionally evidence-level. Extending it to coupled executions of the full policy requires explicit terms for activation changes and for mode-loss differences retained by the cumulative ledger; neither effect is a function of instantaneous post-cap mass alone.

B Algorithm Details

Algorithm 1 gives the per-event online update; Algorithm 2 gives the delayed-validation and appeal-replay update. Persistent trust states, cap schedules, and prior strengths use past matured information. Item weights and support gates may also use the current model-visible batch, before its quality and failure outcomes are observed, as specified in Assumption 1.

Algorithm 1 FARR-SC online platform reputation update

Require: agent i_t, evidence batch B_t, states (μ_i, u_i), q_j, mode losses $\widehat{L}_m$, intercept b

$A_i \leftarrow$ expected-update age since last event; revert μ_i toward the prior and grow u_i with A_i
for all $e \in B_t$ **do**
  $q_{j_e} \leftarrow$ EB-shrunk source reliability; $w_e \leftarrow q_{\tau_e}\, q_{j_e}\, a_e \cdot$ (context, staleness)
end for
apply source and cluster caps to $\{w_e\}$ ▷ bounds effective adversarial mass ε_t
$z_{\text{audit}}, z_{\text{robust}}, z_{\text{feedback}}, z_{\text{blindspot}} \leftarrow$ specialist estimates from (B_t, w, μ_i), clipped around audit/robust anchors
$s_{m,t} \leftarrow$ predictable support gates; $\pi_{m,t} \propto \omega_m s_{m,t} \exp(-\eta_t \widehat{L}_m)$; $z_t \leftarrow \sum_m \pi_{m,t} z_{m,t}$
$\mu_i \leftarrow \text{clip}(\mu_i + \gamma_t(z_t - \mu_i))$ with state gain γ_t from u_i and effective sample size; update u_i
$p_t \leftarrow \sigma(h(\mu_i, u_i) + b)$; choose $a_t \leftarrow \arg\min_a R(a, p_t)$ under the platform cost model
store pending record: source predictions, mode predictions $z_{m,t}$, base logit h_t, weights, validation keys

Algorithm 2 Delayed validation and appeal correction replay

Require: matured validation or appeal for past record r: quality target Y_r^Q, failure label F_r, confidence v_r

1: **for all** stored source predictions $p_{j,r}$ **do**
2: upsert keyed reward $[1-(p_{j,r}-Y_r^Q)^2/0.25]_{[0,1]}$ with mass v_r; recompute q_j
3: **end for**
4: **for all** stored mode predictions $z_{m,r}$ **do**
5: upsert the loss keyed by (r,m) with $v_r s_{m,r}\ell(z_{m,r},Y_r^Q)$; update matured mass and η_t
6: **end for**
7: upsert the keyed calibration record (h_r,F_r,v_r) using stored h_r, not current reputation
8: replay the ordered calibration ledger from the affected checkpoint; append once when the key is new

C Positioning Against Prior Reputation Families

Table 3: Learning signals and objectives in representative adjacent formulations. The comparison concerns the cited methods and the baseline implementations used here, not every possible extension of a method family.

Method / family	Learning signal	Objective and relation to FARR-SC
Beta / Elo / Glicko / TrueSkill / Kalman	Entity outcomes or measurements	Track entity state or skill; separate from learning report-source credit
EigenTrust / TrustRank / PageRank	Endorsement links and graph structure	Propagate trust rather than score reports against delayed outcomes
Matrix factorization / NCF	Observed ratings and interaction structure	Predict ratings from latent representations
Dynamic truth discovery (Li et al., 2015; 2017; Yao et al., 2018)	Discrepancy from inferred truths, with temporal structure	Jointly estimate evolving truth and source weights
Robust / capped aggregation	Bounded current observations	Control per-round influence without a validation-credit ledger
UCB / Thompson baselines	Matured action losses; reports unused here	Learn conservative decisions from delayed validation
FARR-SC	Matured quality and failure targets, with keyed corrections	Learn source credit and specialist predictions; calibrate risk and choose a minimum-risk action

For FARR-SC, specialist learning minimizes quality-prediction losses and calibration follows a failure-label surrogate; the action rule minimizes estimated Bayes risk. These are distinct updates, not a jointly optimized end-to-end platform-loss objective. The standing-mass budget bound is stated separately in Appendix A.4.

Additional adjacent work. The component choices also draw on Byzantine-robust aggregation and corruption-robust learning (Blanchard et al., 2017; Yin et al., 2018; Lykouris et al., 2018; Gupta et al., 2019), and on the classical online-learning and delayed-bandit literature (Cesa-Bianchi & Lugosi, 2006; Littlestone & Warmuth, 1994; Shalev-Shwartz, 2012; Vernade et al., 2017; Pike-Burke et al., 2018; Chapelle & Li, 2011; Platt, 1999). Electronic-market manipulation, whitewashing, promotional review, graph-fraud, and recommender poisoning establish the surrounding attack surface (Resnick et al., 2000; Feldman et al., 2004; Luca & Zervas, 2016; Mayzlin et al., 2014; Akoglu et al., 2013; Jiang et al., 2014; Li et al., 2016; Fang et al., 2018). WebArena, OSWorld, and LLM-as-judge studies broaden agent evaluation beyond static tests (Zhou et al., 2024; Xie et al., 2024; Zheng et al., 2023). These lines motivate the component tools; our analysis concerns their use under delayed-validation credit and bounded current influence.

D Threat Model and Attack Implementations

Attacker goal, knowledge, and capabilities. The attacker manipulates only structured feedback; audits, monitors, and the validation channel are outside its control except for the bounded corrupted-validation fraction. The attacker knows the mechanism, targets the worst-case direction (positive manipulation only on agents that later fail, negative manipulation only on safe agents), and can create fresh sources and clusters at will. All attack volumes scale linearly with the attack-budget parameter s (counts below are at $s = 1$ and are multiplied by s).

Attack generators. *Sybil*: 10–18 fresh sources per event with extreme scores (96 ± 3 positive, 4 ± 3 negative) at confidence 0.34, in per-source or shared clusters. *Collusion*: 4 sources sharing the developer's cluster, scores 94 ± 3, confidence 0.48. *Whitewashing*: 10 never-seen-before sources per event, scores 97 ± 2, confidence 0.26, testing the warmup caps of Corollary 3. *Appeal spam*: 8 appeal-type claims per event, scores 98 ± 2, confidence 0.22. *Adaptive (sleeper)*: 4 sleeper sources submit honest-looking feedback for a two-event warmup phase to accumulate source reliability, then flip to coordinated manipulation (86 ± 5 or 18 ± 5 at confidence 0.66), joined by 4 additional adaptive Sybils; this is a mechanism-aware attacker that exploits the reliability-learning loop itself. *Injected blindspot*: replay-facing audit status is blinded to pending and observed audit scores on later-failing events can be perturbed to false-safe values, so recovery must come from feedback. *Natural blindspot*: no injection; the stress uses only events where the trace's own audit score disagrees with the later outcome.

Robustness, not attribution. FARR-SC never decides whether a source is malicious. Honest-but-unreliable sources and adversarial sources are treated identically: their influence is bounded by learned reliability, warmup caps, cluster caps, and clipping, and Theorem 1 bounds the perturbation of capped evidence from any strategic feedback within the capped effective mass regardless of intent. Proving that a real-world source acted maliciously is an attribution and enforcement problem for platform governance, whose interface in our formulation is the appeal/adjudication channel; it is not required for, and is not claimed by, the robustness guarantee.

E Implemented Actions and Cost Model

The replay experiments use the action set $\{\text{list}, \text{review}, \text{delist}\}$ plus a continuous deposit multiplier; richer actions (restrict, slash, compensate) appear in the platform traces but are not separately optimized by the learner. The failure link is $h(\mu, u) = (62 - \mu)/11 + (u - 18)/20$ on the 0–100 score scale, clamped to $[0.02, 0.98]$ after the sigmoid. Actions minimize expected risk $R(a, p)$ with context-severity factor $s \in \{1.35\ (\text{DeFi}), 1.25\ (\text{devops}), 1.1\ (\text{chatbot}), 1.05\ (\text{research}), 1.0\}$: listing risks $10sp$, review risks $4sp + 1.5$, delisting risks $8(1 - p) + 0.4p$. Realized platform loss charges $10 \cdot \text{severity}$ for listing a failing agent, $4 \cdot \text{severity} + 1.5$ for reviewing one, 1.5 for reviewing a safe agent, 8 for delisting a safe agent, 0.4 for delisting a failing agent, and a deposit term $0.35\,|d - d^\star|$ against an oracle deposit $d^\star \in \{2.7, 1.1\}$. All models are evaluated under the same cost model.

The action-margin condition in Corollary 4 is non-vacuous for this matrix. Writing context severity as $s \geq 1$, pairwise risk-gap slopes in p are $6s$ for list/review, $4s + 7.6$ for review/delist, and $10s + 7.6$ for list/delist. The lower envelope switches only at the active list/review and review/delist boundaries; the third action is strictly separated in a sufficiently small neighborhood of either switch. Hence $\kappa_p \geq 6$. At any active boundary with an unclipped interior preimage, the logistic failure link is smooth and strictly monotone; on normalized score its local magnitude is $\lambda_p = p^\star(1 - p^\star)/0.11 > 0$ while uncertainty and the realized intercept are held fixed.

Cost-matrix sensitivity. To test whether the regime conclusions depend on these constants, we reran the grid over a 3×3 cost matrix (harm cost $\in \{5, 10, 20\}$, false-positive cost $\in \{4, 8, 16\}$, review cost fixed) on Meridian and LinuxArena, clean and audit-blindspot scenarios, eight seeds (144 reports; 36 trace–scenario–cost groups). Against the selected non-FARR-SC baseline in each cell, the frozen default wins all nine LinuxArena blindspot cells and no Meridian blindspot cell. These within-cell comparisons are descriptive. It wins two of nine LinuxArena clean cells and no Meridian clean cell. The experiment therefore preserves the trace-level blindspot boundary while showing that clean operation is cost-sensitive; it does not support a universal cost-robust ordering.

F Additional Experimental Notes

Representative causal slices. Table 4 supplies exact losses and paired intervals for the regime map in Figure 2. All blindspot rows below use injected audit misses with the combined adaptive feedback attack. Chatbot Arena provides a negative transfer case: the default loses 0.928 on this blindspot setting and 1.068 on broad combined attack at scale 1.

Table 4: Representative causal slices with eight predeclared seeds and the same frozen default. The baseline is the lowest-mean non-FARR-SC variant in each cell; the paired interval is conditional on this within-cell selection and is descriptive. Δ is baseline loss minus default loss, with a paired 95% bootstrap interval; positive favors FARR-SC. Blindspot rows include adaptive attack.

Trace / setting	Baseline	Its loss	Default loss	Paired Δ [95% CI]
Meridian clean	Thompson	1.6704	2.1950	-0.525 $[-0.525, -0.525]$
Meridian blindspot	Thompson	2.5829	3.8747	-1.292 $[-1.378, -1.196]$
Chatbot Arena blindspot	Thompson	3.7676	4.6951	-0.928 $[-0.932, -0.923]$
LinuxArena blindspot	Thompson	2.1928	1.0409	$+1.152$ $[1.147, 1.157]$
cx-cmu blindspot	UCB	2.7148	1.1082	$+1.607$ $[1.603, 1.610]$
Synthetic-medium blindspot	Thompson	4.8055	1.6236	$+3.182$ $[3.166, 3.199]$
Synthetic-sparse blindspot	UCB	4.8175	2.2193	$+2.598$ $[2.579, 2.622]$
Synthetic-20K blindspot	RDR	4.9117	1.9976	$+2.914$ $[2.899, 2.929]$

Local-chain marketplace execution. The executable marketplace artifact contains 1270 local-chain transactions, 1504 decoded contract events, 164 audit windows, 203 replay events, 24 same-agent-over-time marketplace entities, 328 on-chain structured reviews, 656 structured replay feedback signals, 39 appeals, and 41 slashes. It is controlled local execution, not live deployment.

Causal experiment grid. The retained suite contains 1,245 completed tasks and zero failures. The mechanism-diagnostic suite contributes 680 selector-alignment, matched-ablation, component, cluster-cap, strategic-mass, and correction cells. The confirmatory suite contributes 565 tasks: a 384-task six-trace grid, 24 tuned-bandit fairness tasks, eight 20K tasks, 144 cost tasks, one pilot replay, and four public-network replays. The confirmatory grid contains 46,136 causal rows; two suite-level causal checks cover the pilot and public-network replays. PageRank, EigenTrust, TrustRank, matrix factorization, NCF, UCB, Thompson, RDR, and tracker baselines share the same traces and seeds. RDR retains source learning, clipping, and caps but omits delayed mode selection and action calibration.

Component ablation. Table 5 is a matched causal ablation at attack scale 1 without validation corruption. Its audit-blindspot rows contain no feedback attack, unlike the main grid's combined blindspot setting. It changes only source learning under the calibrated head. Clean and broad-combined effects are mixed, but every blindspot difference is positive on 8/8 seeds.

Table 5: Matched source-learning ablation on non-adversarial audit blindspots (eight seeds; lower loss is better). Benefit is no-source loss minus the frozen default's loss.

Trace	Frozen default	No source learning	Paired benefit
Meridian	2.9955	3.5122	$+0.5167$
LinuxArena	1.0071	2.1341	$+1.1270$
cx-cmu	1.1976	2.5925	$+1.3950$
Synthetic-medium	0.9646	4.7566	$+3.7920$
Synthetic-sparse	2.0290	5.2098	$+3.1808$

Component, strategic-mass, and correction checks. Fixing mode weights increases blindspot loss by $0.267/0.697/0.695$ on Meridian/LinuxArena/cx-cmu; removing frequency aging increases it by $0.176/0.843/0.553$. Removing warmup growth is nearly neutral in the standard blindspot but

raises Meridian broad combined loss by 0.850. Source caps bind before cluster caps in the primary grid, so we add a concentrated-collusion diagnostic: disabling the cluster cap raises loss by $0.418/0.032/0.202$. The 288-task strategic-mass grid has no cap-order or cumulative-mass monotonicity violations; all nine fitted loss slopes are positive, with eight intervals excluding zero, while 20 of 216 adjacent loss transitions are not monotone. Exact correction replay matches oracle state at lag zero; at 20% provisional corruption and lag 64, final source, specialist-weight, and calibration errors are 0, 0.0214, and 0.0061. These checks identify component roles and state recovery, not monotone immediate-loss improvement.

Audit-blindspot protocol. The injected audit-blindspot stress test is designed to ask whether feedback can correct a plausible audit miss, not to represent all deployment failures. In this stress test, replay-facing audit status is blinded to pending. For events with later failure labels, the observed audit can be perturbed into a false-safe high score; for safe events, a smaller false-penalty perturbation is possible. The combined setting then adds adaptive positive and negative feedback attacks. Natural-blindspot traces are different: they use audit/outcome mismatches already present in the trace rather than injected audit flips. We therefore report clean, injected blindspot, natural blindspot, and combined attack settings separately instead of treating the injected stress test as a deployment frequency estimate.

Bandit baselines and policy heads. UCB and Thompson validation bandits are intentionally strong conservative baselines. They often win clean bad-agent guardrails by refusing feedback exposure, but this can delist many good agents or miss the platform-loss opportunity in audit-blindspot regimes. The shielded FARR-SC policy head is therefore not a new evidence estimator: it is an operating point that lets a conservative bandit raise risk when blindspot-mode support is low. We do not claim shielding monotonically improves FARR-SC; it trades some blindspot recovery for cleaner guardrail behavior.

Market-pilot package. The deployed website workflow produced a 10,129-row accelerated export containing 7,062 recruited-human activity rows. The converter retains 4,603 native structured feedback items at 204 definitive validation anchors; 1,001 shadow-score rows are metadata only. Twenty-five volunteer red-team source rows map to 23 replay anchors. This is controlled launch-stage traffic, not a powered organic-traffic sample. Its no-leakage replay is described in Appendix G.

Interpretation of audit-only. Audit-only has zero feedback attack surface because it ignores feedback. We therefore treat it as a valid no-feedback operating point. The causal grid shows that conservative UCB and Thompson bandits are also strong no-feedback or low-feedback operating points on clean settings. FARR-SC is useful when the platform must use structured feedback to correct blindspots while bounding manipulation.

Relation to governed-agent platform artifacts. The governed-agent platform materials are used as an existing platform artifact and replay source. They are not claimed as a new contribution of this algorithm paper. The new contributions are the delayed-validation learning formulation, FARR-SC, the regret/manipulation analysis, and the platform-loss attack protocols.

G Market-Pilot No-Leakage Replay

Conversion rules. The recruited-volunteer market-pilot export (10,129 rows) is converted at the agent-validation-anchor grain into a status-blinded replay trace. Workspace-run events become bounded usage signals; feedback events become bounded structured reviews and probes; validation events provide delayed labels only when the outcome is definitive (pass, seller-fix-required, restrict, delist), and 76 ambiguous observe-only validations are excluded. Platform-side reputation scores, recommended actions, raw prompts and tool contents, and private adjudication notes are stripped; the 1,001 shadow scoring rows are metadata only. Synthetic fallback feedback is disabled, so replay feedback consists solely of 204 monitor observations and 4,603 native recruited-human structured inputs. The resulting trace has 204 audit anchors over 61 agents with 41 positive delayed labels (20.1%) and includes all 25 volunteer red-team events (22 validations, 3 feedback probes).

Table 6: Status-blinded clean replay of the recruited-volunteer market pilot (single accelerated pilot trace; lower loss is better). Representative subset of the 14 evaluated models.

Model	Loss/event	Bad allowed	Good delisted	Review rate
FARR-SC without source learning	**1.6074**	0.4634	0.0552	0.2696
FARR-SC frozen default	1.6294	0.4634	0.0613	0.2598
EigenTrust	1.7321	0.3415	0.0859	0.2745
PageRank trust	1.7621	0.3415	0.0859	0.2892
Beta reputation	1.8109	0.4146	0.0798	0.2990
FARR-SC selector	1.8261	0.3902	0.0736	0.4167
Thompson bandit	1.9895	0.2683	0.1288	0.2892
Audit-only	3.7020	0.1707	0.4356	0.1225
UCB bandit	5.8967	0.0244	0.7853	0.1765

Interpretation and caveats. The frozen default beats every non-FARR-SC model on overall loss, although its no-source ablation is slightly lower. On the 23 red-team anchors, the default has the lowest point estimate, 0.477 [0.332, 0.684], versus 0.555–0.588 for the strongest graph/rating baselines, but intervals overlap. The trace lacks delayed corrections and cannot support powered win/loss statistics. We use it as no-leakage consistency evidence on recruited-human feedback, not as a superiority claim.

H Natural-Blindspot Results with Modern Baselines

Table 7 extends the natural-blindspot setting (audit/outcome mismatches already present in the trace; no injected flips) to the full causal baseline set. The frozen default wins four of six scale-1 cells and three of six scale-2 cells. LinuxArena, cx-cmu, and synthetic sparse remain positive at both scales; Meridian and Chatbot remain negative; synthetic medium flips at scale 2. This is a trace and attack-strength boundary, not a selected-head table.

Table 7: Natural blindspot plus combined adaptive attack at scale 1 (eight seeds). Δ is the lowest-mean non-FARR-SC loss minus frozen-default loss within each cell; this descriptive interval is conditional on that within-cell selection.

Trace	Best baseline	Baseline loss	Default loss	Δ
Meridian	Thompson	1.9837	2.4826	-0.4990
LinuxArena	Thompson	2.1781	1.9226	$+0.2555$
cx-cmu	Thompson	2.6860	1.3379	$+1.3481$
Chatbot Arena	Audit-only	3.4821	4.5500	-1.0679
Synthetic medium	RDR	3.4906	2.3513	$+1.1394$
Synthetic sparse	RDR	3.6021	1.9487	$+1.6534$

I Shielded Policy Head Trade-off

Table 8 reports the mean effect of shielding across the six grid traces (shielded minus unshielded). Shielding consistently lowers bad-agent allowance but raises good-agent delisting. For the calibrated head it slightly improves clean loss, is neutral on broad combined loss, and raises blindspot loss. It is therefore a priced operating point, not a monotone upgrade.

J External Generalization Protocol and Results

Frozen protocol and post-run integrity correction. The conversion and evaluation protocol below was frozen before any external run. After replay, we found that removing fraud-source edges

Table 8: Equal-weight macro shielded-minus-base deltas across six traces and eight seeds. Negative is better for loss and bad-allowed; positive good-delisted is worse.

Scenario	Base head	Loss Δ	Bad-allowed Δ	Good-delisted Δ
Clean	Selector	-0.5309	-0.1409	$+0.0583$
Clean	Calibrated	-0.0894	-0.0451	$+0.0372$
Combined	Selector	-0.5064	-0.1653	$+0.0976$
Combined	Calibrated	$+0.0010$	-0.0495	$+0.0732$
Audit blindspot	Selector	-0.1656	-0.1166	$+0.1484$
Audit blindspot	Calibrated	$+0.0707$	-0.0715	$+0.1317$

can change the delayed label or severity for a minority of matched audit identifiers. The correction was recorded after replay and before interpreting the oral-v2 loss results: score shifts use every matched identifier, whereas paired loss deltas use only matched identifiers with identical delayed label and severity. It changes no replay output, model state, trace, or score and was adopted to remove a measurement confound, not to select a favorable model. The datasets are the public Bitcoin OTC and Bitcoin Alpha signed trust networks (Kumar et al., 2016) with fraudulent-user ground truth from REV2 (Kumar et al., 2018). Each trust-rating row defines one replay decision event for the rated user. The harness encodes it at an audit anchor: the current rating enters as structured feedback from the rater, while a past-only rolling robust consensus of earlier ratings supplies the audit/monitor anchor. Ratings in $[-10, 10]$ are rescaled to $[0, 100]$ and timestamps are preserved; no labels touch either input. Delayed validation is future senior-rater consensus: a rater is senior if it had at least 5 prior ratings before the event, the validation window is the next 8 ratings with at least 2 senior ratings required (events without enough future validation are excluded, not imputed), and a failure label fires when future senior consensus falls below 45 on the 0–100 scale. This matured failure/severity outcome determines both F_t and the released harness's quality target Y_t^Q; neither is a fraud label. Fraudster labels (platform-verified ground truth) are used only to (i) construct the clean counterfactual by removing fraud-source edges (1,603 edges on each network; the equality is a verified coincidence) and (ii) evaluate learned source reliability post hoc; they never appear in model-visible fields. All fourteen models run with the main-grid hyperparameters and one deterministic causal replay per trace. Uncertainty is not formed from repeated deterministic seeds: score and loss differences use 5,000 rated-entity block bootstraps, and source AUC uses 20,000 source bootstraps. For the attacked-minus-clean loss contrast, the correction retains 21,425 OTC and 14,201 Alpha common-label/severity event rows and excludes 301 and 263 changed-label/severity audit identifiers, respectively; score shifts still use all matched identifiers.

Table 9: External generalization: mean platform loss under the deterministic frozen replay. "Attacked" denotes a retained-fraud-edge counterfactual (all edges, including labeled fraudsters); "clean" removes fraud-source edges. The frozen default is best among non-FARR-SC methods in every column, while the no-source ablation is lower in all four columns.

	Bitcoin Alpha		Bitcoin OTC	
Model	attacked	clean	attacked	clean
FARR-SC without source learning	**1.8772**	**1.6748**	**2.7828**	**2.5319**
FARR-SC frozen default	2.0024	1.7873	2.8061	2.5941
RDR	3.7861	3.8090	3.8467	3.7265
PageRank trust	3.8361	3.8321	3.8272	3.7813
EigenTrust	3.8887	3.8943	3.8976	3.8293
UCB bandit	7.1836	7.3085	6.7240	6.7841

Table 10: Learned source reliability versus ground-truth fraudsters (AUC of $1-q_j$). FARR-SC's q_j is learned only from delayed validation and never observes fraud labels or performs graph propagation.

Model	Bitcoin OTC	Bitcoin Alpha
RDR	**0.9340**	**0.8652**
EigenTrust	0.9146	0.8644
FARR-SC frozen default	0.9101	0.8482
TrustRank	0.8820	0.8166
PageRank trust	0.5026	0.3957

Transfer and operating point. The frozen default has the smallest mean absolute paired score shift on both networks, but the common-label platform-loss contrast is strongest only on OTC: -0.131 with a 95% entity-block bootstrap interval $[-0.211, -0.058]$. Alpha's corresponding -0.010 interval $[-0.049, 0.021]$ contains zero and is not the best baseline contrast. Source separation is strong but not maximal: RDR has the highest point AUC on both networks, and EigenTrust is modestly higher than the default on Alpha. The low failure base rate selects a permissive operating point (failure-event non-delisting 0.987–0.999, non-failure-event delisting 0.008–0.016). These event-level labels come from future senior consensus; fraud labels are used separately for source AUC. The attacked-minus-clean loss contrast uses the post-run label-aligned estimand described above. The experiment supports score stability and source learning without fraud-label supervision, not aggressive fraud blocking or a real-data attack-budget bound. Specialist fraud detectors solve a different task and are therefore not included as detection baselines.