

PAIR Auditing Pairs for AI Agent Collusion

CHUNG Ho Man Owen

The Hong Kong University of Science and Technology

AI Collusion Research Sprint | 23 October 2026

Primary track Detection and Audit of Collusion

Abstract

Collusion can leave evidence across two agents while each individual record appears innocuous. A monitor that reports only per-agent coverage can therefore overstate its ability to collect actionable evidence. We study how audit correlation changes detection under an exact record budget and identical marginal coverage for every agent. PAIR applies a standard max-min linear program to the joint inclusion probabilities of a specified interaction graph. For eight agents and two audits per round, four feasible disjoint pairs obtain 25% joint coverage, compared with 3.57% under uniform fixed-size sampling. The same policy reduces coverage when either record alone suffices. We derive the exact reversal: under equal marginals and a mixture of joint-only and single-record evidence, pairing helps precisely when the joint-only fraction exceeds one half. In 40 shared-controller bandit runs per condition, the higher coverage deters attacks in a deliberately selected incentive regime. However, allowing migration to unmodeled pairs or revealing the realized audit draw defeats the specialized policy. A uniform hedge restores a measurable coverage floor but does not guarantee deterrence. These are synthetic allocation results, not evidence of emergent LLM collusion or a validated semantic detector. The contribution is a reproducible evidence-aware benchmark and explicit limits on when correlated auditing is beneficial.

1. Introduction

An execution agent and an approving agent may produce records that become suspicious only when compared. If a financial institution can inspect two of eight records, auditing every agent 25% of the time does not determine how often it sees both sides of a transaction. The missing quantity is joint coverage. We isolate that allocation problem without assuming that price correlation, shared language or communication alone proves collusion.

We contribute an equal-budget and equal-marginal benchmark, an exact result connecting audit correlation to evidence type, and adversarial tests of partner migration and audit leakage. The financial setting motivates the two-record threat model; the experiments are abstract simulations rather than a validated financial workflow. Our central recommendation is conditional: choose audit correlation after establishing what records a detector actually needs.

2. Related Work

Randomized audit allocation is established in Stackelberg security games [1]; our linear program is an application of that machinery. Fish et al. study LLM pricing behavior [2], while Hartline et al. develop regret-based pricing audits [3]. Casheekar and Tangirala analyze audit incentives and learning with shared policies [4]. We instead hold individual audit marginals fixed and vary their joint distribution, distinguishing evidence requiring both records from evidence requiring either. Recent work also cautions against inferring a punishment mechanism from price impulse responses alone [5]. We make no claim that correlated audits, max-min optimization or the probability identities below are new mathematical techniques.

3. Methods

3.1 Threat model and evidence

There are n agents, with one auditable record per agent per round. The auditor samples a subset S of exactly k records. A preexisting coalition chooses one eligible pair $e = \{i, j\}$, or abstains, before learning the realized S . The schedule distribution is public. No cross-round evidence pooling is available: the required records concern the same round. A known interaction graph G specifies feasible pairs; expanding it to all pairs is an adversarial migration test.

In the primary AND evidence model, the detector identifies a harmful transaction only when both records are audited. Conditional sensitivity is s and specificity is assumed perfect. In the OR control, either record suffices. These are stipulated visibility models, not estimated properties of an LLM monitor. A toy XOR-sharing construction checks the AND plumbing: each single bit is uniform, but their XOR reveals the harmful label. It does not demonstrate practical detector accuracy.

3.2 Policies and equal treatment

Uniform samples each k -subset with equal probability. PAIR chooses probabilities x over the same subsets to maximize the minimum probability of collecting sufficient evidence on G . Every agent has marginal inclusion $m = k/n$. Thus the comparison changes audit dependence while keeping individual exposure, total records and maximum per-round load identical. A 20% hedge mixes 80% PAIR with 20% Uniform.

$$\begin{aligned} &\text{maximize } t \quad \text{subject to } q(e) \geq t \text{ for every eligible pair } e \\ &\quad \Sigma x(S) = 1, \quad \Pr(i \in S) = k/n \text{ for all } i, \quad x(S) \geq 0 \end{aligned}$$

For AND evidence, $q(e) = \Pr(i \text{ and } j \text{ are in } S)$. For a mixture with joint-only fraction ρ , the objective uses $d(e) = \rho \cdot q(e) + (1 - \rho) \cdot \Pr(i \text{ or } j \text{ are in } S)$. The evidence-aware variant solves the same LP with $d(e)$. All probabilities are computed by finite enumeration. SciPy HiGHS solves the small LPs; enumerating all subsets is not a scalable solution for large deployments.

3.3 Adaptive coalition and protocol

One shared bandit controller represents an already coordinated coalition. Its arms are eligible pairs plus abstention. This deliberately gives the coalition centralized pair selection; it does not simulate independent agents discovering collusion. An attack earns benefit 1 and incurs penalty 8 if detected. The controller uses sample-average epsilon-greedy learning, with epsilon declining linearly from 0.20 to 0.02 across 20,000 rounds. Evaluation freezes estimates and removes exploration for 5,000 rounds. Forty independent controller replicates are retained per condition.

The realized schedule is hidden unless a Bernoulli leakage event reveals it before action. In leaked rounds a stipulated safe fallback selects a pair lacking complete coverage. These rounds do not update hidden-context estimates. Sensitivity 0.5 tests detector degradation. Benefit and penalty intentionally place Uniform and PAIR on opposite sides of a deterrence threshold; the resulting learning gap is a mechanism demonstration, not an unbiased estimate of real-world effect size.

3.4 Analysis and reproducibility

The primary contrast is PAIR minus Uniform in undetected harmful transactions per evaluation round for known pairs and hidden draws. Confidence intervals use 10,000 paired percentile bootstrap resamples over controller replicates. They cover simulation randomness only. Secondary analyses include five fixed graph families, 20 random graphs, n from 4 to 12, feasible budgets from 2 to 6, migration, leakage and detector sensitivity. All seeds, raw endpoints, schedules and configuration are included. The local analysis plan was not externally preregistered; the evidence-type and penalty sweeps are exploratory analytical additions.

4. Results

4.1 Identical individual coverage can conceal a sevenfold gap

For $n = 8$, $k = 2$ and four disjoint eligible pairs, Uniform chooses one of 28 pairs each round. PAIR chooses one of the four eligible pairs. Every agent is audited with probability $1/4$ under both schedules, but eligible-pair coverage rises from $1/28$ to $1/4$ (Figure 1). A fixed-attack positive control with 200,000 trials yields detection rates of 3.61%, 25.14% and 20.82% for Uniform, PAIR and the hedge, consistent with exact values of 3.57%, 25.00% and 20.71%.

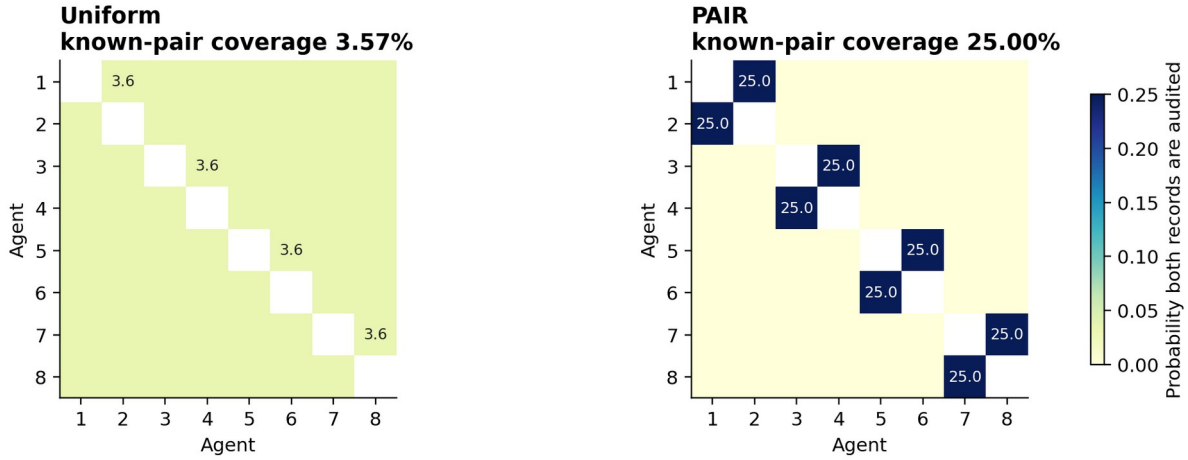


Figure 1. Exact joint inclusion probabilities for two fixed-budget schedules. Every off-diagonal cell under Uniform is 3.57%; the printed numbers mark the four eligible pairs. Every agent has 25% marginal coverage under either schedule. Cells show joint coverage, not detector accuracy.

4.2 Evidence type reverses the recommendation

Let q be pair coverage and m the common individual marginal. Inclusion-exclusion gives OR coverage $2m - q$. With joint-only fraction ρ , detection is $(1 - \rho) \cdot 2m + (2\rho - 1) \cdot q$. Therefore an increase Δq changes detection by $(2\rho - 1) \cdot \Delta q$. Pairing helps for $\rho > 1/2$, hurts for $\rho < 1/2$, and has no effect at $\rho = 1/2$, assuming equal conditional sensitivity across evidence types. For OR evidence in the matching example, PAIR reduces coverage from 46.43% to 25.00%.

Evidence model	Uniform	PAIR	Change
Both records required	3.57%	25.00%	+21.43 pp
Either record sufficient	46.43%	25.00%	-21.43 pp
Equal mixture	25.00%	25.00%	0.00 pp

Table 1. Exact detection coverage under three stipulated evidence models. All rows use the same record budget and equal individual audit marginals.

These are exact coverage probabilities, not statistical estimates. The result is a design test: measure whether useful evidence is predominantly joint before increasing audit correlation. With different sensitivities for the two evidence types, the threshold becomes a comparison of their sensitivity-weighted frequencies rather than a universal 50%.

4.3 Topology and adversarial adaptation bound the benefit

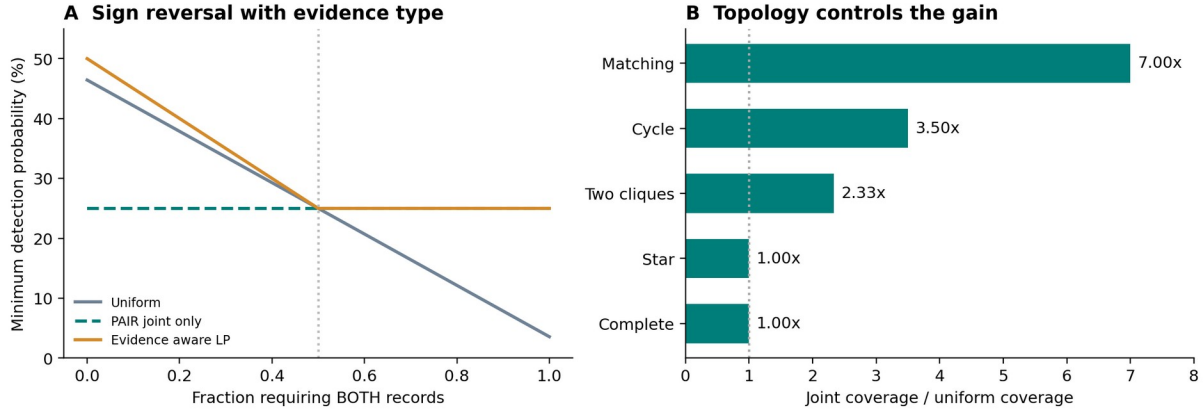


Figure 2. A: Exact detection under mixed evidence; the evidence-aware LP chooses its correlation accordingly. B: Gain in worst eligible-pair coverage at $n = 8$ and $k = 2$. A star has no gain under equal marginal exposure because the central agent is the bottleneck.

Undetected harm %	Uniform	PAIR	20% hedge
Known pairs hidden	96.45 [96.37, 96.52]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Any pair allowed	96.47 [96.41, 96.53]	100.00 [100.00, 100.00]	99.28 [99.24, 99.32]
Half of draws leaked	98.24 [98.18, 98.30]	50.09 [49.84, 50.34]	50.09 [49.84, 50.34]
All draws leaked	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]
Sensitivity 0.5	98.23 [98.17, 98.29]	28.52 [15.41, 41.67]	89.58 [89.44, 89.72]

Table 2. Mean undetected harmful transactions per 100 evaluation rounds, with 95% controller-seed bootstrap intervals. Each cell uses 40 controller runs and 5,000 evaluation rounds per run. Zero intervals reflect identical simulated endpoints; they do not imply certainty in real deployments.

In the primary condition PAIR reduces undetected harm by 96.45 percentage points (95% paired interval [96.37, 96.52]). All PAIR controllers abstain during evaluation. Expected attack payoff explains this: $1-8q$ equals 0.714 under Uniform and -1 under PAIR. The demonstration depends on those incentives.

The adverse tests are consequential. Migration to any pair exposes zero-coverage edges under pure PAIR; its undetected harm becomes 100%. The hedge restores only 0.71% minimum off-graph coverage, insufficient for deterrence at penalty 8. Full leakage enables safe attacks under every schedule. At sensitivity 0.5, PAIR lies exactly on the zero-payoff boundary; run variability is large. The hedge remains profitable to attack. Coverage gains alone do not establish robust deterrence.

5. Discussion and Limitations

An audit dashboard that reports only a 25% inspection rate for every agent cannot distinguish the policies in this study. A useful dashboard should also report coverage of the evidence sets required by its detector. In a financial workflow, those sets could be linked approval records, paired execution decisions or corresponding internal messages. Their evidentiary value must be established on labeled workflow data before optimizing a schedule.

PAIR is useful when harmful behavior requires jointly inspecting a small, stable set of possible relationships. It is less attractive when a single record is informative, when feasible relationships are uncertain, or when attackers see the current allocation. A uniform hedge provides an explicit safety floor, but the floor can be too small to change incentives. The complete-graph control shows why: with exactly k audits the sum of all pair inclusion probabilities is always $k(k-1)/2$, so no schedule improves the minimum over every possible pair beyond Uniform.

Limits of the evidence

The main empirical result is synthetic. One shared controller selects pairs; there are no independent foundation models, natural-language messages, spontaneous coalitions or real transactions. The detector is an assumed observation model. The XOR control has perfect specificity by construction and is not an empirical safety claim. We do not measure a trade-off between benign task completion and false accusations. Audits consume a constant record budget; real records may differ in token length, review time, evidentiary quality and privacy cost.

The known graph is supplied rather than inferred. The LP enumerates all k -subsets, limiting scale. The analysis ignores evidence stored across rounds, changing identities, more than two colluders, asymmetric inspection costs, delayed feedback and a coalition manipulating graph discovery. Source review identified closely related security-game and multiagent audit research; a comprehensive novelty assessment remains open. The proposed contribution is the evidence-type benchmark and its controlled failure tests, not a general solution to AI collusion.

Future work

The highest-value extension is to test whether the AND assumption holds in an actual financial agent workflow. Collect matched harmful and legitimate trajectories, label which subsets allow a blinded monitor to recover the violation, and estimate evidence-type prevalence and conditional sensitivity. Freeze the graph and monitor on development data, then compare schedules on held-out trajectories with a fixed token or reviewer-time budget. Only after that test should deterrence be evaluated with separate LLM agents. The included extension protocol specifies acceptance criteria without claiming these experiments have been run.

6. Conclusion

At identical individual exposure and exact audit cost, changing who is audited together can alter two-record evidence collection by a factor of seven in a small known graph. The same change reverses sign for single-record evidence, and specialized coverage can fail under migration or leakage. Effective collusion auditing therefore needs an evidence model, an interaction model and protected audit timing alongside a monitoring budget.

Code and Data

Code repository: <https://github.com/owenchung030915-hkust/PAIR>. It contains the source, unit tests, exact schedules, controller endpoints, figures, and this report. No external model API was called. The experiments were run on 23 October 2026.

References

- [1] Sinha, A., Fang, F., An, B., Kiekintveld, C., and Tambe, M. (2018). Stackelberg Security Games: Looking Beyond a Decade of Success. *IJCAI*, 5494-5501. <https://www.ijcai.org/proceedings/2018/0775.pdf>
- [2] Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. (2024; revised 2026). Algorithmic Collusion by Large Language Models. *arXiv:2404.00806v6*. <https://arxiv.org/abs/2404.00806>
- [3] Hartline, J. D., Wang, C., and Zhang, C. (2025). Regulation of Algorithmic Collusion, Refined: Testing Pessimistic Calibrated Regret. *ACM CS&Law*. *arXiv:2501.09740*. <https://arxiv.org/abs/2501.09740>
- [4] Casheekar, A. M., and Tangirala, H. (2026). Learning to Report Unsafe Tasks in a Multi-Agent Game. *arXiv:2610.09002v2*. <https://arxiv.org/abs/2610.09002>
- [5] Lopes de Moraes, K., and De-Losso, R. (2026). Algorithmic Collusion Without Deterrence: Auditing the Punishment Mechanism in Q-Learning Pricing Agents. *arXiv:2610.21867v2*. <https://arxiv.org/abs/2610.21867>
- [6] Apart Research (2026). AI Collusion Research Sprint. Official tracks, guidelines, rubric and report template. Accessed 23 October 2026. <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>

Appendix A Limitations and Dual Use Ethical Considerations

All experiments use synthetic records and abstract rewards. No financial system was contacted, no transaction was executed, no sanctions were imposed, and no private user communications were collected. The shared controller is a deliberately adversarial model organism, not a claim about any deployed product or provider. Coalition choice and harm labels are explicitly planted.

The migration and leakage tests describe ways a specialized monitoring policy can fail. Those insights could help an adversary evade a predictable monitor. We include them because omitting the failure modes would encourage unsafe deployment. The code contains no financial exploitation tooling. Real audit draws should be sampled after actions commit, or protected from agent access until commitment, and should not share an agent-visible random seed.

Audit allocation does not determine guilt. A real deployment needs validated evidence, access controls, data minimization and review before any consequential action. Concentrating joint inspection may expose more linked personal or commercial information even when marginal record counts remain equal. Equal marginals are an experimental control and do not certify equal privacy impact.

Appendix B Exact Results and Replication

B.1 Matching bound

Let $n = 2r$, $k = 2$ and let G be a perfect matching with r edges. Since $\Pr(i \text{ is audited}) = 2/n$, joint coverage of any edge cannot exceed $2/n$. Sampling each matching edge with probability $1/r = 2/n$ attains that upper bound simultaneously and preserves every marginal. Uniform pair coverage is $1 / \text{choose}(n, 2)$. The gain is $n-1$. For $n = 8$ the exact gain is 7. This elementary bound describes a special case and is not a claim of a new general optimization theorem.

B.2 No universal pair coverage improvement

For every size- k subset, exactly $\text{choose}(k, 2)$ pairs are jointly included. Taking expectations gives $\sum_e q(e) = \text{choose}(k, 2)$. Therefore $\min_e q(e)$ cannot exceed $\text{choose}(k, 2) / \text{choose}(n, 2) = k(k-1)/(n(n-1))$. Uniform sampling attains this value. If the coalition can choose any pair and the defender has no additional information, specialized scheduling cannot improve worst-case joint coverage.

B.3 Evidence and sensitivity

With equal marginals m , $\Pr(i \text{ or } j \text{ audited}) = 2m - q(e)$. If a fraction ρ of cases require both records and the rest require either, detection is $\rho \cdot s_{\text{AND}} \cdot q(e) + (1 - \rho) \cdot s_{\text{OR}} \cdot (2m - q(e))$. Increasing $q(e)$ helps exactly when $\rho \cdot s_{\text{AND}} > (1 - \rho) \cdot s_{\text{OR}}$. The half threshold in the main text assumes $s_{\text{AND}} = s_{\text{OR}}$. The prevalence ρ must describe the relevant attack population, not a monitor-selected sample that overrepresents easy detections. False positives and heterogeneous evidence costs require additional constraints.

B.4 Deterrence and the hedge

For attack benefit b , detection penalty F and conditional sensitivity s , expected attack reward is $b - F \cdot s \cdot q(e)$. Under perfect specificity and hidden draws, all eligible attacks are strictly deterred if $F \cdot s \cdot \min_e q(e) > b$. At equality, zero payoff gives no strict preference for abstention. Mixing Uniform with weight λ guarantees at least $\lambda \cdot k(k-1)/(n(n-1))$ coverage for every pair, irrespective of graph omissions. This is a coverage floor, not a deterrence guarantee.

B.5 Reproduction commands

```
python -m pip install -r requirements.txt
python -m unittest discover -s tests -v
python src/run_experiments.py
python src/make_figures.py
```

The completed numerical run used Python 3.12.14, NumPy 2.3.5 and SciPy 1.17.0. It ran in 18.3 seconds on the execution host. This timing is host-specific and excludes figure creation and document rendering. See `results/run_manifest.json` for source hashes, analysis settings and exact environment versions.

The 15 adaptive cells contain 600 controller runs, 12 million training rounds and 3 million evaluation rounds. The reported intervals resample the 40 controller endpoints in each condition, not the millions of dependent within-run decisions. Topology values are exact for the supplied graphs. Repeated identical graph families are not counted as independent empirical samples.

Source template: official Apart Google Docs template, downloaded as DOCX on 23 October 2026. Its section structure and page settings are retained; title, prose, figures and comparison tables have been populated and reformatted for readability.