
Evidentiary Gaps in AI Collusion Detection: When Price Patterns Are Not Enough¹

Sharon T. Itonyo

Independent Researcher · Chicago, Illinois

With Apart Research

Abstract

Parallel pricing can reflect coordination, shared market shocks, or independent responses to public guidance. Motivated by the pre-existing working paper *Beyond Metrics*, which distinguishes measurement from accountable inference, we studied whether richer audit evidence improves identification of planted coordination without justifying unsupported attribution. We generated four scripted two-seller market controls and trained price-only and evidence-aware logistic detectors on disjoint development episodes. At each of three evidence-visibility levels, evaluation used 160 positive and 480 negative held-out episodes. Under complete observations, the price-only detector identified 70/160 coordination episodes (43.8%) with 28/480 false positives (5.8%); the evidence-aware detector identified 159/160 (99.4%) with no false positives in the constructed test distribution. At 15% audit visibility, evidence-aware detection fell to 137/160 (85.6%) with 29/480 false positives (6.0%). A separately generated, intentionally difficult benign stress test modeled two sellers independently following an identical public price bulletin. Without a provenance feature distinguishing public advice from private coordination, the frozen evidence-aware detector flagged all 160 full-visibility cases. This stress result is conditional on the deliberately chosen generator, not a real-world error rate. A separate abstention rule covered only 48.3% of sparse-evidence cases, illustrating the cost of uncertainty. The study shows that useful detection scores cannot alone establish the source or legal meaning of apparent coordination. No deployed LLMs or real-market records were evaluated.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

Pricing similarity is ambiguous. Two automated sellers can raise prices together because they coordinated, because each independently responded to common costs, or because both followed a public external signal. A detector trained to identify matching price trajectories may be useful as a triage instrument yet inadequate as an attribution mechanism. This is especially relevant to AI-agent markets, where behavioral similarity need not reveal agreement, communication, or the origin of a recommendation [2–4].

The pre-existing SSRN working paper Beyond Metrics argues that measurement alone does not secure accountability: producing a signal differs from interpreting it, assigning responsibility and acting appropriately [1]. We take that distinction as conceptual motivation, not as data or a previously demonstrated empirical result. The new question is narrow: in a controlled synthetic market, which evidence views distinguish planted action coordination from innocent correlation, and what remains unknowable when apparently decisive signals share an alternative causal explanation?

We compare a price-only baseline with a small evidence-aware auditor, then evaluate a separately generated benign public-reference scenario. Although the causal mechanisms differ, a limited audit schema may classify public recommendations as if they were private price proposals. Our contribution is a reproducible controlled demonstration of measurable detection gains and an independently generated attribution stress test, not proof of legal collusion or emergent LLM coordination.

2. Related Work

Algorithmic collusion research has studied how language-model agents can sustain elevated prices or cooperate in simulated auctions [3, 5]. Recent work emphasizes that high prices or correlated behavior alone can fail to distinguish tacit market effects from deliberate coordinated action [4]. Antitrust analysis likewise distinguishes parallel conduct from additional evidence sometimes called plus factors [2]. In the sprint guidance, Track 2.3 specifically asks what should count as convincing evidence of collusion, beyond suspicious correlations [6].

The original Beyond Metrics manuscript was completed before this sprint and is cited here solely for its four-layer accountability framing [1]. Its text, tables and analyses are not presented as sprint findings. All scripted scenarios, feature extraction, model fits, stress tests, figures and measurements reported here were generated for this October 2026 exercise. No proprietary project software, private logs, architecture, datasets or unreleased work were used.

3. Methods

3.1 Synthetic market and ground-truth controls

We generated two rule-based pricing controllers for each of four case types. Each episode contained 40 price-setting rounds. A competitive reference process combined a common cost component, demand variation, seller-specific cost variation and pricing noise. This creates legitimate parallel behavior even without coordination. Exactly one case type has planted joint action; labels describe how the simulator generated prices, not legal findings.

Table 1. Four deliberately constructed worlds; these are scripted synthetic controllers, not LLMs.

Case type	Ground truth	Role in experiment
Coordinated action	Planted joint markup and price-target follow-through	Positive control
Independent pricing	No shared coordination process	Negative control
Common cost shock	Parallel price changes from shared exogenous cost	Confounding control
Talk without action	Price proposals present but unimplemented	Speech-only negative control

In the planted-coordination case, synthetic price proposals are followed by matching prices by construction. In the talk-only case, proposals are not implemented. The independently generated stress control, described in Section 4.3, instead represents a public price bulletin followed separately by two sellers. The feature extractor records recommended price and follow-through but not whether the source was public or private. The distinction is deliberately absent from the audit input.

3.2 Evidence visibility and disjoint evaluation

We masked each episode's cost/demand context and proposal-like records independently at 100%, 50% and 15% visibility, while leaving market prices visible. Training used 75 seeds per primary scenario, validation 35 different seeds, and the held-out standard test 160 additional seeds per type. Each mask condition therefore contains 640 standard episodes (160 planted positives; 480 negatives). Masks at different visibility levels reuse the same underlying episodes and are paired, not independent market samples. The separate public-guidance control uses an additional 160 unseen seeds and does not enter training or threshold selection.

3.3 Detectors, thresholds and abstention

Both classifiers use standardized numeric features and regularized logistic regression. The price-only baseline receives seven summary statistics of prices (including level and change correlations, price spread, volatility, direction agreement and high-price prevalence). The evidence-aware model adds observable cost-adjusted residual features, proposal-to-price follow-through and missingness indicators. Neither detector receives ground-truth case identity. These features are deliberately simple and predeclared in the accompanying clean-room Python script.

We selected an alert threshold separately for each model on the held-out development validation subset, maximizing F1 subject to a validation false-positive rate no greater than 12%, then froze both thresholds for testing (price-only 0.54; evidence-aware 0.32). We report detection/recall among planted coordination, false-positive rate among the three noncoordinating controls, precision and balanced accuracy. The classifier probabilities are uncalibrated scores; thresholds are not legal confidence standards.

For a distinct selective-decision analysis, the auditor abstains when its score lies strictly within 0.14 of its frozen alert threshold (absolute score distance < 0.14). The evidence-aware rule also abstains if zero recommendations are observed AND fewer than 20% of cost observations are available. We report coverage and the false-positive and false-negative rates conditional on a decision, including their denominators; selective accuracy alone can obscure the large number of unresolved cases.

4. Results

4.1 Detection with complete and missing evidence

Table 2. Fixed-threshold held-out test; 160 planted-coordination and 480 noncoordination cases per row. TPR = true-positive rate; FPR = false-positive rate.

Visible audit evidence	Price-only TPR / FPR	Evidence-aware TPR / FPR	Evidence-aware precision
100%	43.8% / 5.8%	99.4% / 0.0%	100.0%
50%	43.8% / 5.8%	98.1% / 1.0%	96.9%
15%	43.8% / 5.8%	85.6% / 6.0%	82.5%

At complete visibility the evidence-aware model flagged 159/160 planted-coordination episodes and 0/480 negatives, compared with 70/160 and 28/480 for the price-only baseline. At 15% visibility the evidence-aware detector flagged 137/160 positives and 29/480 negatives. For 100% visibility, the Wilson 95% interval on evidence-aware TPR was 96.5%–99.9% (159/160); the corresponding FPR interval was 0.0%–0.8% (0/480). Paired bootstrap 95% intervals for the evidence-aware MINUS price-only false-positive rate were: -5.8 percentage points [-8.1, -3.8] with complete evidence, -4.8 [-7.1, -2.7] at 50%, and +0.2 [-2.5, +2.9] at 15%. These intervals are conditional on the scripted generator and frozen test protocol, not external-validity estimates.

Table 3. Evidence-aware detector uncertainty on the standard held-out test: 95% Wilson score intervals for rates and paired 4,000-resample bootstrap intervals for the TPR gain over price-only detection; conditional on the synthetic generators.

Evidence	TPR [95% CI]	FPR [95% CI]	TPR gain, pp [95% CI]
100%	99.4% [96.5%, 99.9%]	0.0% [0.0%, 0.8%]	+55.6 pp [48.1, 63.1]
50%	98.1% [94.6%, 99.4%]	1.0% [0.4%, 2.4%]	+54.4 pp [46.9, 61.9]
15%	85.6% [79.4%, 90.2%]	6.0% [4.2%, 8.5%]	+41.9 pp [33.1, 51.2]

Held-out synthetic scripted-agent tests; 160 seeds per scenario

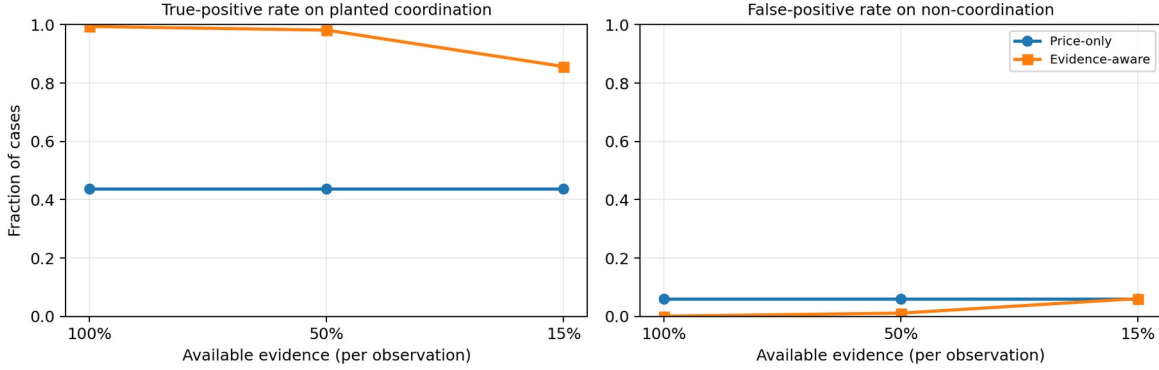


Figure 1. The price-only detector is unchanged by audit visibility. Evidence-aware gains erode when audit context and messages are missing. Test sets contain paired masks of the same underlying episodes.

4.2 Feature ablation and abstention

Evidence-type ablation illustrates how design choices drive the apparent gains. At complete visibility, prices plus cost context achieved a 76.3% TPR and 2.5% FPR; prices plus messages achieved 98.8% TPR and 1.3% FPR. At 15% visibility those TPRs fell to 53.8% and 78.1% (125/160), respectively. Planted messages are more often followed by matching action than in the talk-only control, so this discriminative feature is partly built into the generator. This should not be mistaken for a universal collusion signature.

Table 4. Selective decision rates after abstention (evidence-aware detector). FNR/FPR are conditional on cases receiving decisions, not all 640 observations.

Visibility	Decisions / coverage	Selective FNR	Selective FPR
100%	639/640 (99.8%)	1/160 (0.6%)	0/479 (0.0%)
50%	626/640 (97.8%)	2/159 (1.3%)	3/467 (0.6%)
15%	309/640 (48.3%)	5/130 (3.8%)	12/179 (6.7%)

The evidence-aware abstention rule made decisions on 639/640 (99.8%) full-evidence cases, 626/640 (97.8%) at 50% and 309/640 (48.3%) at 15%. At 15%, it abstained on 30 of 160 positives and 301 of 480 negatives. Among the 130 decided positives, 5 were missed (selective FNR 3.8%); among 179 decided negatives, 12 were falsely flagged (selective FPR 6.7%). Only 125/160 (78.1%) of all planted positives were flagged without abstention. This 78.1% numerically equals, by coincidence, the independent 125/160 TPR in the prices-plus-messages ablation; the calculations describe different procedures. Abstention does not recover absent provenance.

4.3 Source-provenance identifiability stress test

We replaced an initial relabeling-style counterfactual with a separately generated benign negative control (160 additional random seeds). In the new generator, each seller independently observes a publicly posted reference price built from common market factors and a public recommendation premium, then sets its own price using seller-specific costs and independent noise. Neither seller observes or coordinates with the other. Public bulletins are represented in the same anonymous price-target log field used for private proposals, because the auditor lacks signal-origin metadata. This is a deliberately difficult out-of-distribution stress test, not an independently validated population model.

Without retraining the model or changing thresholds, the evidence-aware detector flagged 160/160 public-reference cases at 100% visibility, 160/160 at 50%, and 123/160 (76.9%) at 15%. The Wilson 95% interval on the full-visibility flag rate was 97.7%–100%, reflecting repeated samples from THIS constructed stress generator only. Both 100% and 50% visibility conditions flagged all 160 cases. The public-reference generator is independent of the coordination generator, but intentionally produces strong shared-recommendation follow-through, so these numbers do not measure real-world benign false accusation prevalence.

The stress test exposes a missing audit field: recommendation provenance. An observer might see synchronized prices and target-follow-through, yet not know whether those targets were openly published, privately exchanged, or algorithmically generated from competitors' confidential data. In this simulator, the same detector calls the benign

public-guidance world coordinated. This supports auditing source and authorization separately from behavior; it does not establish that any real-world implementation of such provenance would resolve all ambiguity.

5. Discussion and Limitations

Within the standard planted-rule distribution, adding audit information strongly improves detection, but that comparison reflects purpose-built feature differences. Under sparse evidence, false positives rise and abstention becomes frequent. The separately sampled public-guidance hard negative is flagged despite being noncoordinated by construction, highlighting how a source-blind schema can support a confident but causally unjustified inference. The appropriate evaluation axes are detection, benign parallelism, provenance and selective coverage, not aggregate accuracy alone.

Limitations

These results concern scripted synthetic pricing controllers, not deployed LLM agents. Training and held-out standard tests use the same generator family; masked observations are paired, and planting agreement/action features makes the evidence-aware model unusually strong. The independently sampled public-guidance stress generator is intentionally difficult and outside the training distribution; its high flag rate is neither a real-world estimate nor proof of formal observational equivalence. The confidence intervals reflect repeated draws from specified synthetic mechanisms; no claims about legal agreement, causal attribution or deployable calibration follow. Further qualifications appear in Appendix A.

Future Work

This exploratory pilot identifies a practical audit failure mode rather than providing a market-ready legal or regulatory classifier. Future tests should use actual LLM-based market agents; diverse benign public-guidance generators; blinded source-attribution controls; recorded external shocks; calibrated selective risk thresholds; and preregistered evaluations. It remains open whether verified source metadata would improve attribution under authentic agent workflows and whether the gains survive changes in market structure.

6. Conclusion

Across controlled scripted markets, fuller audit evidence improved detection of planted coordination. But a newly and independently generated benign public-reference world was routinely flagged because the auditor could not identify the origin of price guidance. The contribution, motivated by the pre-existing Beyond Metrics framework, is an empirical reminder that measurement is not justified attribution. These toy results neither establish collusion by LLMs nor amount to a legal classification method.

Code and Data

All episodes and results come from a newly written, standalone Python simulator using synthetic data; no private market records, production systems, or proprietary external project implementation is included. A reproducibility ZIP with the exact evaluated script, seed, dependency record and CSV outputs is retained separately; no public repository or video is required for this sprint submission. The code can be shared with reviewers at the owner's discretion after terms review.

Author Contributions

Sharon T. Itonyo: research framing, study oversight, interpretation, author review and submission responsibility. AI-assisted software tools were used in simulation implementation, numerical checking and drafting as disclosed below.

References

- [1] Itonyo, S. T. (2026). Beyond Metrics: The Accountability Layer in Deployed AI Systems. SSRN working paper. <https://ssrn.com/abstract=6452578> (pre-existing research; conceptual inspiration only).

- [2] Berger, L. F., Weiner, M. L., & Jiang, W. (2025). Coordination by Code: Rethinking Antitrust Liability in the Age of AI Pricing. American Bar Association, Antitrust Law. https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/
- [3] Agrawal, K., Kunnavakkam, S., Srikanth, V., Teo, V., & Vazquez, J. (2025). Evaluating LLM Agent Collusion in Double Auctions. SPAR Research Library, working report. <https://library.sparai.org/reports/evaluating-llm-agent-collusion-in-double-auctions-wflyiz/>
- [4] Riemer, M., Tosato, T., Memarian, A., Puelma Touzel, M., Berseth, G., Rish, I., & Dumas, G. (2026). Position: Collusion Risks Among AI Reasoning Agents Justify Certification Requirements for Making Market Decisions. Proceedings of Machine Learning Research, 306, 171880-171907. <https://proceedings.mlr.press/v306/riemer26a.html>
- [5] Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024; revised 2026). Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
- [6] Apart Research. (2026). AI Collusion Research Sprint: Track 2.3 and guidelines. <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>

Appendix A. Limitations and Dual-Use / Ethical Considerations

A.1 Limitations

- The agents are entirely scripted numerical controllers, NOT LLMs or realistic autonomous agents. Deliberate coordination is planted by the experimental generator; its emergence is not established. The controls, label prevalence and evidence signatures were invented for methodology testing, not calibrated to any real market.
- Features such as a known competitive cost benchmark and a logged target price are unusually informative and may be unavailable, untrustworthy, or themselves regulated in practice. Message-follow-through is highly discriminative partly by construction; the high full-evidence accuracy is not a deployable performance estimate.
- The new public-guidance stress test has a SEPARATE data generator: its sellers independently follow an accessible external price bulletin using distinct local costs and noise. Although no coordination is planted, the audit feature schema strips bulletin provenance and preserves price-target follow-through. It intentionally mimics a strong risk feature, so the observed 100% full-visibility flag rate is conditional on this difficult synthetic generator, not a prevalence estimate or proof that all independent sellers look coordinated.
- Classifiers were evaluated on seed-disjoint episodes from the same scripted data generator, with masks of each episode paired across conditions. This prevents exact seed leakage but not structural similarity between train and test. The 25% positive prevalence is artificial. Precision changes with prevalence, and these model scores are not calibrated probabilities.
- No causal inference about legal agreement, social harm, market power, institutional incentives, future liability or actual optimal monitoring intensity follows. The four-layer framework in Beyond Metrics is prior research, not itself experimentally tested here. No ethics board, human-subject data or real customer data were involved.

A.2 Dual-Use and Ethical Considerations

- Communication-action features could expose blind spots that a malicious pricing agent might exploit by reducing logs or making proposals less explicit. The report deliberately omits detailed evasion optimization and real-world deployment instructions, and contains only synthetic code.
- An overly confident auditor may falsely flag lawful parallel behavior, with commercial and reputational consequences. We recommend provenance capture, explicit uncertainty, protected human review and meaningful appeal paths before any adverse action, rather than automatic sanctions based on classifier outputs.
- Price-setting systems and competition-law determinations have legal implications. This study is research methodology, not regulatory or legal advice; any operational screening would require independent validation, market expertise, permissioned datasets and institutional safeguards.

Appendix B. Reproducibility and Prior Work

Reproduction: Python 3 with NumPy, pandas, scikit-learn and Matplotlib; exact evaluated code and dependencies are included in a private reproducibility package. Run: `python evidence_audit_revised.py --out results`. Random root seed 2601025; each primary episode has 40 rounds; 75/35/160 case seeds in training/validation/test per scenario, each observed at 100%, 50%, 15%. Thresholds 0.54 (price-only) and 0.32 (evidence-aware) were frozen on development data before the independent public-guidance stress test. Per-rate intervals use 95% Wilson scores; paired bootstrap comparisons use 4,000 resamples of the same held-out episodes within each label group. All quantitative tables derive from the independent script output, including abstention denominators.

Prior work: Itonyo's Beyond Metrics SSRN working paper (2026) and the cited literature. New during this sprint: scripted simulator, standard positive/negative controls, price-only and evidence-aware model comparison, masking and abstention protocol, feature ablations, a replacement independently generated benign public-reference stress control, uncertainty intervals, fresh figures and the present report. Original relabeling-style stress outputs were superseded, not presented as evidence. No earlier unpublished proprietary implementation was reused.

The assessed script, raw CSVs and run log are in the author's separate reproducibility bundle; no public code URL or source-code fingerprint is claimed as accessible in this submission.

LLM Usage Statement

AI tools assisted with implementing and revising the synthetic simulator, numerical verification, methodological critique and report drafting. An AI-assisted review checked the original table arithmetic and identified stress-test and reporting weaknesses; the stress test was independently regenerated and the frozen detectors were rerun. The author is responsible for the study design, interpretation, factual validation and final submission. All figures and numbers reported here reflect recorded script outputs, not invented measurements.