

Predicting False Collusion Alerts Under Incomplete Information

Evidence from a simulated market and an exploratory re-asking of LLM agents

Edimah SYNESIUS SONGO

Paris — 23–25 October 2026.

Research conducted at the [Apart Research AI Collusion Research Sprint](#).

Abstract

Collusion screens can mistake responses to shared information for collusion when auditors record less than sellers observe. In a simulated market with known ground truth, we predict and measure this failure. A standard conditional-independence screen wrongly flags 5.8% of non-colluding seller pairs with all five news levels, 51.7% with two recorded categories, and 100% with none. Model-based approximations closely predict these rates without fitting them. With two categories, increasing observations from 600 to 6,000 raises false positives to 100%. A stricter threshold delays the failure: with two categories, it still flags 99.7% of non-colluding pairs at 6,000 rounds. It can also reduce detection of planted collusion. In an exploratory LLM-market pilot, a news-only screen flags seven of nine pairs in games without a planted private channel. We also test re-asking: querying each AI seller again on its recorded prompt, with the planted private instruction removed wherever present, to obtain comparison answers. The pre-registered binary statistic misses planted pairs, while a rank statistic chosen after inspecting the games detects the planted pair in all three planted games before correction for multiple testing. We have not yet established whether re-asking gives reliable error rates when sellers respond to earlier rounds. We predict how often incomplete information causes false alerts, but do not claim to have validated a collusion detector. More data can strengthen evidence of dependence without strengthening evidence of collusion.

1 Introduction

AI agents can set similar prices for different reasons: they may collude, or they may simply respond to the same public information. Shared models and inputs make these explanations difficult to separate. Language-model agents have shown collusion in simulated markets [9], making reliable ways to screen for collusion an open AI-safety problem.

The auditor’s view is the shared information included in the collusion test. It may contain less detail than the sellers observe.

Consider two fuel stations that raise prices after the same crude-oil news. Their prices move together, but that alone does not show agreement. The standard test used here asks whether their prices still move together after accounting for the news included in the test. If they do, the screen flags the two stations as suspicious and worth investigating.

But suppose both stations raise prices independently in response to a news detail missing from the auditor’s view. The screen can then flag their shared response as possible collusion even though they are not colluding. We call the screen’s warning an alert, not a finding of collusion. A false positive is an alert about sellers that are responding independently to shared information. Figure 1 draws this case.

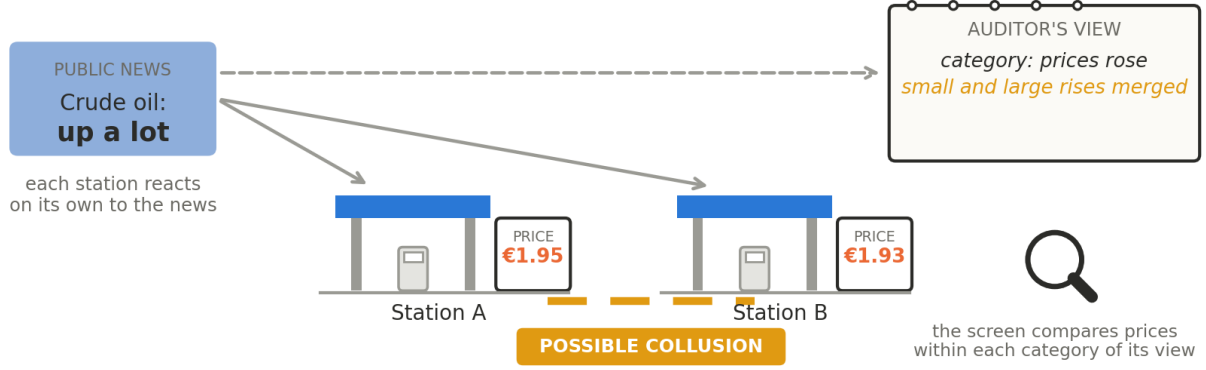
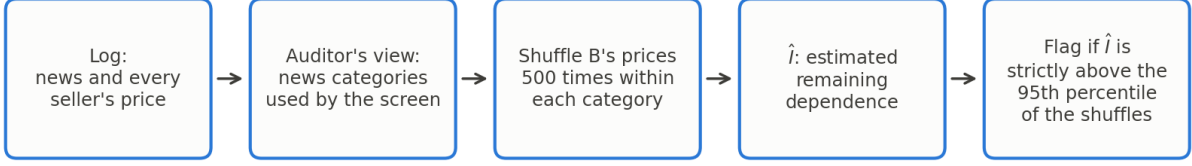


Figure 1: **The problem, in the paper’s running example.** Both stations respond independently to detailed crude-oil news. The auditor’s view merges small and large rises into one category, ‘prices rose’. Across rounds in that category, the stations’ prices can still move together, so the screen can raise a false collusion alert.

We predict how often the screen raises false alerts when the auditor misses shared information. In a simulated market where we know which sellers collude, we compare these predictions with measured false-positive rates and test whether more data or a stricter threshold fixes the problem. We then test re-asking: querying each AI seller again on its recorded prompt, with any private instruction removed, to obtain comparison answers. Our main findings come from the simulation. The LLM pilot tests an idea that needs further validation. Figure 2 contrasts the two ways of building comparison answers. Our main contributions are:

- measured false-alert rates under an incomplete auditor’s view (matched by theoretical predictions),
- evidence that more data can increase false alerts, while a stricter threshold can delay but does not prevent them,
- an exploratory test of re-asking agents and comparing answers to help distinguish collusion from parallel responses.

STANDARD TEST



RE-ASKING

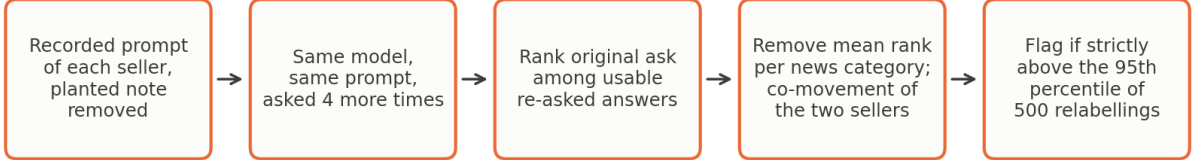


Figure 2: **Two ways to build comparison answers.** Top: the standard test shuffles one seller’s prices only among rounds with the same news category used by the screen, 500 times. It flags the pair when the estimated remaining dependence, $\hat{I}$, is strictly above the 95th percentile of the shuffled values. Bottom: re-asking sends each recorded prompt, with any planted private instruction removed, to the same model four more times. It ranks the original ask against usable re-asked answers, removes each seller’s mean rank within each news category, and measures the two sellers’ remaining co-movement. The pair is flagged when this statistic is strictly above the 95th percentile of 500 relabellings. Both methods look for dependence, not proof of an agreement.

2 Related Work

Bonjour et al. [3] use an information-based measure to detect collusion between agents. We examine how a screen using this measure behaves when the auditor misses shared information. The underlying concern is not new: shared shocks can make competing firms look collusive [2]. Our contribution is to quantify and predict this failure in a controlled market.

We use the auction environment of Agrawal et al. [1]. Work on secret communication between AI agents [11] motivates testing whether a screen can find private collusion without also flagging ordinary shared responses. Finding similar behaviour does not explain what caused it.

Re-asking draws on the model-X conditional-randomization idea: generate comparison answers from a specified conditional distribution [5]. Here, that distribution is supplied by querying the agent on its recorded prompt. We apply this existing idea to agent audits. That connection does not by itself show that our test has the stated error rate, especially when later prompts depend on earlier answers.

Reproducing prompts and querying the same model is more demanding than ordinary chatbot access, consistent with concerns about limited access in AI audits [6]. We test whether this access provides useful comparison answers, not whether it reveals every hidden input or detects every form of collusion.

3 Methods

3.1 Standard test measure

We measure how much one seller’s price tells us about the other’s after accounting for recorded news. This is **conditional mutual information**. Let A and B denote the two sellers’ prices and C the news category used by the screen. Write I for the population information measure and $\hat{I}$ for its estimate from observations. The estimate is $\hat{I}(A, B \mid C)$, where the vertical bar means “given

the recorded news”. Zero means no remaining information about one price in the other. Natural logarithms give the unit, nats.

To see which dependence could arise “by chance”, we shuffle seller B’s prices 500 times, only among rounds with the same news category used by the screen. This is a **stratified permutation test** [10]. We flag a pair when its observed statistic is strictly above the 95th percentile of the shuffled values. This is the implemented nominal 5% rule, not a guarantee of a 5% false-positive rate under an incomplete record.

3.2 Why missing news can produce false alerts

Return to the fuel stations. Suppose both see whether crude-oil prices rose a little or a lot, but the auditor’s view merges both into one category, “prices rose”. Both stations may raise their prices more after a large rise, independently of each other. Even within the auditor’s single category, their prices can therefore move together. The screen can mistake this shared response for evidence of collusion. This is an illustration of the missing-detail mechanism, not an additional experiment.

In the simulation, we know how sellers respond to each news level. We can therefore calculate how much dependence remains after categories are merged and use a large-sample approximation to predict how often the screen raises an alert. We compare those predictions with measured rates rather than fitting them to the results. Appendix A gives the approximation, assumptions and limits.

We also test a **stricter BIC-based rule** [12], which requires more evidence as the dataset grows. Let n be the number of rounds, K the number of populated recorded news categories, and $\log$ the natural logarithm. The rule flags when $2n\hat{I} > K \log n$. The left side measures evidence of remaining dependence. The right side is the threshold for an alert.

The problem is that missing information can leave real dependence even without collusion. With more rounds, evidence of that dependence grows roughly in proportion to n , while the stricter threshold grows only as $\log n$. Requiring more evidence can delay false alerts without fixing their cause.

3.3 Simulated market

Four sellers choose a low or high price after seeing one of five equally likely news levels. Let N denote the full news level, numbered 1 to 5, and let β denote how strongly prices respond to news. Let P denote probability. The logistic function σ turns any real-valued input x into a probability, using the natural exponential base e . It is defined by $\sigma(x) = 1/(1 + e^{-x})$. A seller chooses a high price with probability $P(\text{high}) = \sigma(\beta(N - 3))$, with $\beta = 1$. Choices respond to news without being deterministic. In planted-pair runs, two sellers use the same fair coin to set their prices on 20% of rounds.

The main experiment has 600 rounds and 200 repeats. The auditor keeps all five news levels, merges them into three or two categories, or records no news. “No news” is one constant category, so $K = 1$, not 0. We reuse the same simulated rounds across auditor’s views to isolate what changes when only the auditor’s view changes.

We also vary news sensitivity, unrecorded shared-input strength, the auditor’s view, and the round count. Sobol indices measure how much each input contributes to variation in the predicted false-positive rate, alone and through interactions (Appendix D).

3.4 LLM market

We use a double auction, where buyers post bids and sellers post asks and trades occur when they cross [1]. Three Claude Haiku sellers interact with three scripted buyers for 100 rounds. Seller chat is disabled. Each seller sees a public demand line and past asks. We analyse three games without a planted pair and three with one. A separate mixed-model comparison is in Appendix E.

The standard screen uses price moves, not levels: 1 if the standing ask rose from the previous round, 0 otherwise. This gives 99 moves per game. The auditor’s view has three demand categories and does not include the full history. Model and sampling settings are in Appendix E.

3.5 Re-asking

Re-asking asks whether the sellers’ original prices are unusually high or low together compared with what each seller would say again on the same non-private input.

We query the same model four more times per recorded prompt. The 900 prompts from the three games without a planted pair were reconstructed byte-for-byte. In planted games, the private coin instruction was removed from all 290 prompts carrying it.

We rank each original ask against its usable re-asked answers, counting ties halfway. We then remove each seller’s average rank within each news category and measure whether the two sellers’ remaining ranks move together. Figure 3 shows one round.

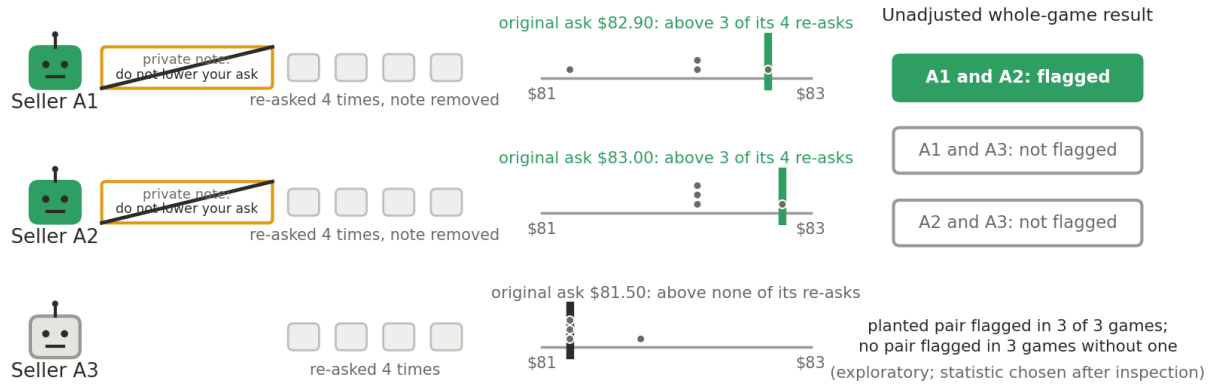


Figure 3: **Re-asking in one round of a planted game** (seed 1, round 10). Sellers A1 and A2 received the private note ‘do not lower your ask’, which is removed before re-asking. Dots show the four re-asked asks, and the bar shows the original ask. In this round, the planted sellers’ original asks sit above most of their re-asked asks. The test measures whether the two sellers’ ranks move together across usable rounds, after accounting for news categories. The decisions on the right are for the whole game, not this single round.

The comparison assumes the original answer could have been any one of the re-asked answers under the no-private-input null. This is called **exchangeability**: the label “original” should not identify a different kind of answer. We relabel answers 500 times and recompute the statistic. A pair is flagged when its covariance is strictly above the 95th percentile. Matching prompts supports this assumption locally, but does not establish it for an adaptive whole game (Appendix F).

The pre-registered statistic was the standard $\hat{I}$ on raised-or-not moves, compared with re-asked moves instead of shuffled ones (Figure 10a). It missed the planted pair because sellers could obey “do not lower your ask” by holding their price. We chose the rank statistic after inspecting those games, so its results are exploratory, not independent validation. Appendix F gives the full selection history.

We call a statistic pre-registered when it and its success criterion were committed to the repository before any result was computed (commit 806c69f). We then tried two variants on the same games. The first counted lowered asks and detected the planted pair in one of three games. The second, the rank statistic, detected it in three of three. The rank statistic was committed before it was computed (commit 143a047), but after the first two results were known. It is therefore the third version tried, and its result is exploratory.

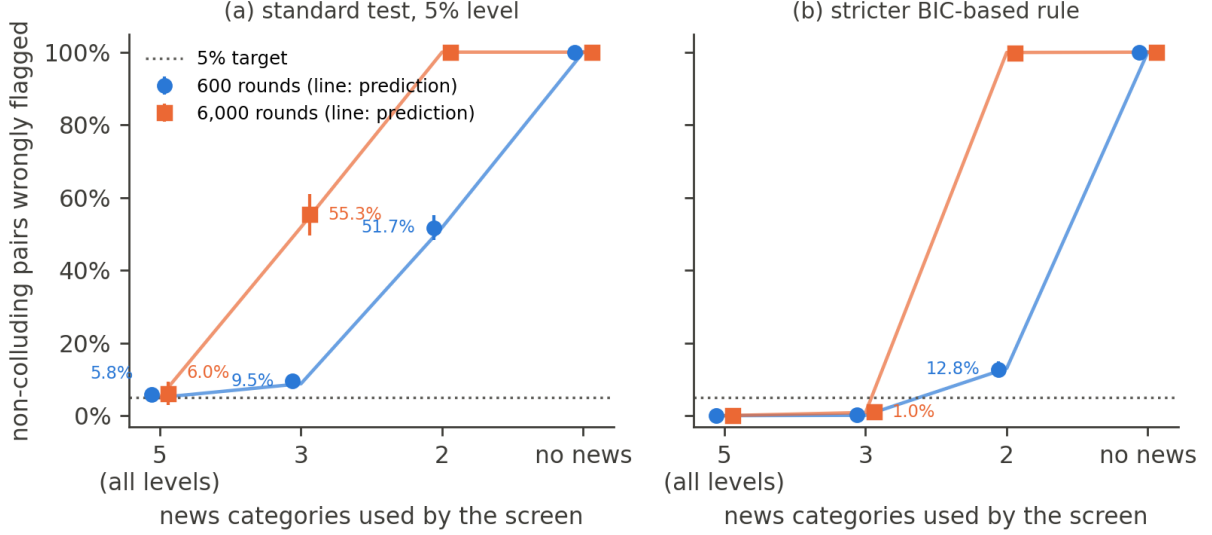


Figure 4: **Less detailed information produces more false positives.** Share of non-colluding seller pairs wrongly flagged (six pairs pooled per repeat), by the number of news categories used by the screen. (a) Standard test at the 5% level. (b) Stricter BIC-based rule, which flags when $2n\hat{I} > K \log n$. Blue circles: 600 rounds, 200 repeats; orange squares: 6,000 rounds, 50 repeats; printed numbers are the measured rates. Error bars: 95% whole-repeat bootstrap intervals (zero width at 0% or 100%, where every repeat agrees). Lines: model-based predictions, not fits. Dotted line: the standard test’s 5% target. In a separate condition where prices ignore news, the stricter rule detects a planted pair in 33.5% of repeats, against 98.5% for the standard test (all five news levels, 600 rounds).

3.6 Uncertainty

To show how much the simulated rates vary across repeated runs, we calculate 95% bootstrap intervals. We resample whole repeats, keeping their six seller pairs together, and take the middle 95% of 10,000 resulting rate estimates. Keeping related pairs together avoids treating them as extra independent evidence, which could make the error bars too narrow.

For LLM results, we report counts and game totals rather than treating pairs from the same game as independent trials. Exact binomial bounds [7] can be applied to independent games, but do not correct for choosing a statistic after inspecting the data. Appendix E explains this distinction.

4 Results

4.1 False positives increase with less detailed information

With all five news levels recorded, the screen wrongly flags 5.8% of non-colluding seller pairs as potentially collusive (95% interval: 4.5–7.1%), close to its 5% target (Figure 4a). With three recorded categories, it wrongly flags 9.5% (7.9–11.2%). With only two, it wrongly flags about half—51.7% (48.3–55.1%). With no news recorded, it wrongly flags every non-colluding pair tested. The model predicts rates of 5%, 8.8%, 51.8%, and 100%.

Sensitivity analysis shows that the auditor’s view and unrecorded shared inputs both matter, including through interactions. These rates are predictable from the known market model and the information used by the screen—not from that information alone.

4.2 Two traps: more data or a stricter rule

With two news categories, increasing rounds from 600 to 6,000 raises the standard test’s false-positive rate from 51.7% to 100% (Figure 4a). With three categories, more rounds raise it from 9.5% to 55.3% (prediction: about 52%). The stricter rule delays this rise (Figure 4b): with two categories it moves from 12.8% to 99.7%, and with three it stays at 1% at 6,000 rounds.

That protection has a cost: in a separate market where prices ignore news, with all five news levels and 600 rounds, the stricter rule detects a planted pair in 33.5% of repeats, versus 98.5% for the standard test. More data reveals the omitted dependence more reliably. A stricter threshold changes how much evidence is required, not what caused it.

4.3 LLM sellers

The screen marks seven of nine seller pairs as suspicious even though no private collusion channel was planted in those games (Table 1, first row). This does not prove the sellers were behaving competitively, but it shows why an alert alone cannot establish a private agreement. The news-only view omits history the sellers see. Across three same-model games, the standard test flags 7/9 pairs with no planted private channel. Across three mixed-model games, it flags 4/9. These small, dependent samples do not establish a model effect.

Rank-based re-asking flags none of nine pairs in three games without a planted channel. It detects the planted pair in all three planted games and flags none of the six other pairs. The pre-registered binary version detects no planted pair.

A post-hoc correction for testing three pairs within each game reduces rank-based planted-pair detection to two of three games, with no alerts in games without a planted channel. This unregistered correction only works as intended if the re-asking p-values are valid. We have not yet established that they are.

The standard screen also detects no planted pair and flags none of the other pairs in the planted games. The standard test flags 7/9 pairs in games without a planted pair but 0/6 non-planted pairs in planted games. With so few games, 7/9 is not a stable estimate of a general LLM-market false-positive rate.

Table 1: **Unadjusted LLM pair-test counts (no multiple-testing correction)**. Each game has three sellers and 100 rounds. Games without a planted pair have no planted private collusion channel. This does not establish competitive behaviour. Pairs within a game share sellers and history. Counts are descriptive; independent-pair confidence bounds are not used.

Method	Pairs flagged without a planted pair (3 games)	Planted pair detected (3 games)	Other pairs flagged in planted games
Standard test, news only	7/9	0/3	0/6
Re-asking, “raised or not” (pre-registered)	0/9	0/3	0/6
Re-asking, rank (chosen after inspecting planted games, exploratory)	0/9	3/3	0/6

Three games without a re-asking alert still leave substantial uncertainty about performance on new games. Appendix E explains the game-level bounds and why they cannot validate a statistic selected using these data.

5 Discussion and Limitations

An alert reflects the information used by the screen as well as what the agents did. Auditors should explain which shared inputs may be missing and assess false positives under explicit assumptions about them. Our simulation can calculate these rates because its generating model is known. A

real audit would need to justify that model. Preserving model versions, settings and re-asking inputs could make controlled re-asking testable in practice. An alert should open an investigation, not establish collusion.

Limitations

We do not yet know how well re-asking works on new games or collusion that develops without being planted.

Removing the private note can reveal that the note affected sellers’ answers. It does not by itself show that they made an agreement. A mismatch could also reflect changed settings or another unrecorded input. Conversely, subtle collusion or collusion through public history may escape detection. The absence of a planted private channel does not rule out tacit collusion through responses to public prices [4].

Re-asking requires answer variation and matching model versions, settings and state. Because later prompts contain earlier original asks, matching each prompt does not by itself justify the whole-game comparison. We have not yet established that the implemented test gives reliable error rates. Appendix F details the assumptions, response filtering, and possible sequential justification.

The simulated market uses binary prices and no memory. Its conclusions are not estimates for all real markets. Less detailed information can hide dependence as well as create it (Appendix A).

Future work

The priority is to establish reliable error rates for re-asking on adaptive agents. We would then test whether a richer auditor’s view—news only, then the order book, then the full prompt—reduces false alerts among LLM sellers. A pre-registered study with 20+ games and about 19 re-asks per prompt would test detection across providers and collusion that emerges with chat enabled. Finally, MTS-K fuel-price data would test the screening analysis in a real market, where collusion is not known in advance.

6 Conclusion

In this controlled market, missing information makes the screen wrongly flag non-colluding sellers as potentially collusive at predictable rates. Neither more data nor the stricter BIC-based rule removes the dependence caused by the missing information. In this small pilot, a re-asking statistic chosen after inspection separated planted from non-planted pairs before correction for multiple testing; after a post-hoc correction for the three pairs in each game, it detected the planted pair in two of three games. We do not yet know whether it gives reliable error rates. The next step is to check those rates and test a method chosen in advance on new games.

Code and Data

- **Repository:** [Edimah/hackathon-ai-collusion](https://github.com/Edimah/hackathon-ai-collusion).
- **Analysis commands:** `python sim.py --all`, `python figures.py`, `python sensitivity.py`, `python llm/analyse.py`, `python llm/replay.py --rank`, and `python llm/table1.py`. Reproducing LLM analyses that require original game logs also requires access to those logs.
- **Saved results:** `results/`. Cached re-asked answers: `results/replay_draws.jsonl`, 7,128 answers. Raw game logs are not in the repository.
- **Preexisting work:** the LLM auction environment is adapted from Agrawal et al., commit 2bb2a7e. The underlying information measures and testing tools also come from prior work cited above. Only the auction patch is shared because the upstream code has no licence.

- **Work done during the sprint:** the controlled missing-information study, model-based alert-rate calculations, simulation and sensitivity analyses, and LLM screening and re-asking experiments reported here. The contribution is their design, implementation and evaluation, not invention of the underlying information measure, BIC threshold, or conditional-randomization principle.
- **Other artifacts:** docs/gifs/record_dial.gif, an 8-second loop of the main result drawn by docs/gifs/fuel_story.py, which also draws Figures 1 and 3.

Author Contributions

E.S.S. designed the study, made the design decisions, wrote the conceptual code for the statistic, null, decision rule and BIC-based rule, and registered the LLM analyses and addendum. E.S.S. is responsible for the final report. AI drafting and implementation assistance is described below.

LLM Usage Statement

Claude Code implemented most of the simulation, LLM patch, re-asking and figure code under the author’s decisions. Claude assisted with design review, checking results and citations, and explaining the methods. AI assisted with drafting and simplifying report prose and incorporating audit findings. The author is responsible for the final text, claims, citations and results.

References

- [1] Agrawal et al. (2025). Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413.
- [2] Bajari, P., Ye, L. (2003). Deciding Between Competition and Collusion. *Review of Economics and Statistics* 85(4), 971–989.
- [3] Bonjour, T., Aggarwal, V., Bhargava, B. (2022). Information Theoretic Approach to Detect Collusion in Multi-Agent Games. UAI 2022, PMLR 180.
- [4] Calvano, E., Calzolari, G., Denicolò, V., Pastorello, S. (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110(10), 3267–3297.
- [5] Candès, E., Fan, Y., Janson, L., Lv, J. (2018). Panning for Gold: Model-X Knockoffs for High-Dimensional Controlled Variable Selection. *JRSS-B* 80(3), 551–577.
- [6] Casper, S. et al. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. ACM FAccT 2024. arXiv:2401.14446.
- [7] Clopper, C. J., Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 26(4), 404–413.
- [8] Cover, T. M., Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley.
- [9] Fish, S., Gonczarowski, Y., Shorrer, R. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
- [10] Kim, I., Neykov, M., Balakrishnan, S., Wasserman, L. (2022). Local Permutation Tests for Conditional Independence. *Annals of Statistics* 50(6). arXiv:2112.11666.
- [11] Motwani, S. R. et al. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. NeurIPS 2024. arXiv:2402.07510.
- [12] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 6(2), 461–464.
- [13] Sobol’, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* 55.

- [14] Saltelli, A. et al. (2010). Variance based sensitivity analysis of model output. *Computer Physics Communications* 181.
- [15] Jansen, M. J. W. (1999). Analysis of variance designs for model output. *Computer Physics Communications* 117.

A Why less detailed information can create dependence

A.1 Variables and assumptions

- In this appendix, A and B denote two sellers' random low/high prices in a generic round. C denotes recorded news.
- t : round index, taking integer values from 1 to the previously defined round count n .
- $A_t, B_t \in \{0, 1\}$: two simulated sellers' random prices, low or high.
- $N_t \in \{1, \dots, 5\}$: full public news, an ordered categorical variable.
- g : the fixed map that merges full news levels into recorded categories. Recorded news in round t is $C_t = g(N_t)$.
- K : number of populated recorded categories. No news means a constant category, $K = 1$.
- I : population conditional mutual information. Its estimate from counts is $\hat{I}$.

In the non-colluding news-only model, sellers use independent price randomness once full news is known: $A \perp B \mid N$. This assumption must include any other shared input before it can be applied to a richer model.

A.2 How omitted detail creates dependence

The information chain rule [8] gives:

$$I(A, B \mid C) = I(A, B \mid N, C) + I(A, N \mid C) - I(A, N \mid B, C).$$

The first term is zero when the information used by the screen is a function of news and sellers are independent given full news. Dependence can remain because that information loses part of the news.

Let q_A and q_B be the sellers' high-price probabilities as functions of full news. Thus $q_A(N) = P(A = 1 \mid N)$, with the same definition for seller B. Let c be a possible recorded category, and write Cov for covariance, which measures how two quantities vary together. Conditional independence gives:

$$\text{Cov}(A, B \mid C = c) = \text{Cov}(q_A(N), q_B(N) \mid C = c).$$

If both response functions increase with news, this covariance is nonnegative. It is strictly positive when a merged category contains news variation to which both respond strictly. Positive covariance establishes dependence. Merely omitting information is not sufficient: the omitted detail must create dependence, and the test needs enough data to detect it.

A.3 From counts to model predictions

Let a and b denote possible low/high price values and c a possible recorded category. Let $\hat{p}(a, b, c)$ denote their observed joint frequency, and let the corresponding expressions after a vertical bar denote conditional frequencies. The sum runs over all possible values of a , b , and c . The observed frequencies supply the estimated statistic:

$$\hat{I}(A, B \mid C) = \sum_{c,a,b} \hat{p}(a, b, c) \log \frac{\hat{p}(a, b \mid c)}{\hat{p}(a \mid c) \hat{p}(b \mid c)}.$$

Natural logs give nats. Zero-count joint cells contribute zero. Empty categories are omitted. Each category is weighted by its observed frequency.

For a prediction, let p without a hat denote a model probability rather than an observed frequency. Let u range over full news levels. The sum includes those levels that the previously defined map g sends to category c . For news-only non-colluding sellers:

$$p(a, b, c) = \sum_{u: g(u)=c} p(N = u) p(A = a \mid N = u) p(B = b \mid N = u).$$

For the shared-input model, let m_t denote the hidden shared random input in round t and let λ denote its nonnegative strength. Also integrate over that input. The reported theoretical calculation uses 80 Gauss–Hermite quadrature nodes to approximate this integral.

Here I is the population dependence calculated from the generating model. The notation $\chi_K^2(\delta)$ means a noncentral chi-square distribution with K degrees of freedom and noncentrality parameter δ , which describes the shift away from the no-dependence case. With binary prices and populated cells, the central chi-square distribution approximates $2n\hat{I}$ under independence. Under dependence, our working approximation is $2n\hat{I} \approx \chi_K^2(2nI)$. It uses noncentrality $2nI$ to predict alert rates without fitting them to measured rates. Let $q_{K,0.95}$ denote the central chi-square 95th percentile. Write $F_D(z)$ for the probability that a variable with distribution D is at most threshold z . The approximate chance of an alert is:

$$P(\text{alert}) \approx 1 - F_{\chi_K^2(2nI)}(q_{K,0.95}).$$

This is not the exact permutation distribution. Observed and predicted rates are 51.7% versus 51.8% for two categories at 600 rounds, 55.3% versus about 52% for three categories at 6,000 rounds, and 76.5% versus 78.6% for the shared-input pair with all five news levels at $\lambda = 1$. Agreement supports the approximation in these conditions, not universal accuracy.

A.4 The stricter threshold

The BIC-style rule [12] replaces the standard threshold with $K \log n$. Its approximate alert probability uses that threshold in the preceding formula. Fixed omitted dependence produces evidence growing approximately as $2nI$, eventually faster than $\log n$. Specified priors would be needed to interpret an exact Bayes factor.

A.5 Missing news can also hide dependence

For this counterexample only, let N be binary news and U a private binary input. They are independent fair coins. The exclusive-or operation $\oplus$ returns 1 when its two inputs differ and 0 when they match. Set the sellers' prices to $A = N \oplus U$ and $B = U$. Without recording N , A and B are independent. Given N , one determines the other. This counterexample shows that merging information can hide dependence as well as create it. It is not the planted-market mechanism.

B Simulation, decisions, and uncertainty

B.1 Generating the prices

News is independent across rounds and uniform on $\{1, \dots, 5\}$. Four sellers choose binary prices. Their base high-price probability is $\sigma(\beta(N_t - 3))$, with main setting $\beta = 1$.

Let κ be the probability that a planted pair uses a shared coin in a round. In planted conditions, sellers 2 and 3 have both prices overwritten by the same fair coin with probability $\kappa = 0.2$. For the separate shared-input condition, let m_t be a shared random input in round t and λ its nonnegative strength. The notation $\mathcal{N}(0, 1)$ means a normal distribution with mean 0 and variance 1. Sellers 1 and 4 use:

$$P(\text{high} \mid N_t, m_t) = \sigma(\beta(N_t - 3) + \lambda m_t), \quad m_t \sim \mathcal{N}(0, 1).$$

The shared input m_t is independent across rounds. The main shared-input setting is $\lambda = 1$.

The three-category map is $\{1, 2\}, \{3\}, \{4, 5\}$. The two-category map is $\{1, 2, 3\}, \{4, 5\}$. The no-news map is constant. The same generated rounds are reused across auditor’s views, making comparisons paired.

The generator uses `default_rng([seed, condition index])`. Main conditions use seeds 0–199. No-news calibration and 6,000-round runs use seeds 0–49. The 6,000-round extension tests non-colluding pairs only.

B.2 From observations to an alert

For each pair, compute $\hat{I}$ and 500 within-category shuffled values. Flag when the observed statistic is strictly above their 95th percentile, using linear interpolation between order statistics. The implementation does not use a +1 Monte Carlo p-value correction. Calibration was measured directly: 5.8% with the complete record and 6.3% when prices ignore news.

The false-positive rate is the share of tested non-colluding seller pairs that the screen wrongly marks as potentially collusive, counted across simulation repeats. Planted-pair detection and alerts on other pairs in planted markets are reported separately.

For pooled intervals, all six pairs stay together while whole repeats are resampled 10,000 times. The 2.5th and 97.5th percentiles form the interval. If every repeat has the same boundary outcome, this bootstrap returns a zero-width interval. It does not establish certainty about an unknown population rate.

B.3 Standard test versus the ranking rule

The ranking rule is the standard test plus a comparison with other agents, not a ranking alone. Here i and j identify sellers, and k identifies any other seller. Write $\hat{I}(i, j)$ for the estimated information between sellers i and j , with recorded-news conditioning left implicit. The net-influence score $\Gamma(i, j)$ compares their information with the strongest alternative partner for j . The symbol $\max$ means take the largest value across those other sellers. The rule flags pair (i, j) only when the standard threshold is crossed and both directional scores are positive [3]:

$$\Gamma(i, j) = \hat{I}(i, j) - \max_{k \notin \{i, j\}} \hat{I}(k, j).$$

The score Γ is distinct from the information estimate $\hat{I}$.

With prices ignoring news, standard-test calibration is 19/300 pair-tests, or 6.3%. The ranking rule flags 16/300, or 5.3%. For the shared-input pair, the standard test flags 153/200 repeats, or 76.5%, versus 148/200, or 74%, for the ranking rule. Planted-pair detection under the standard test is 197/200 without a news response and 200/200 with one. The ranking rule gives 195/200 and 200/200. Figure 5 shows every condition and pair.

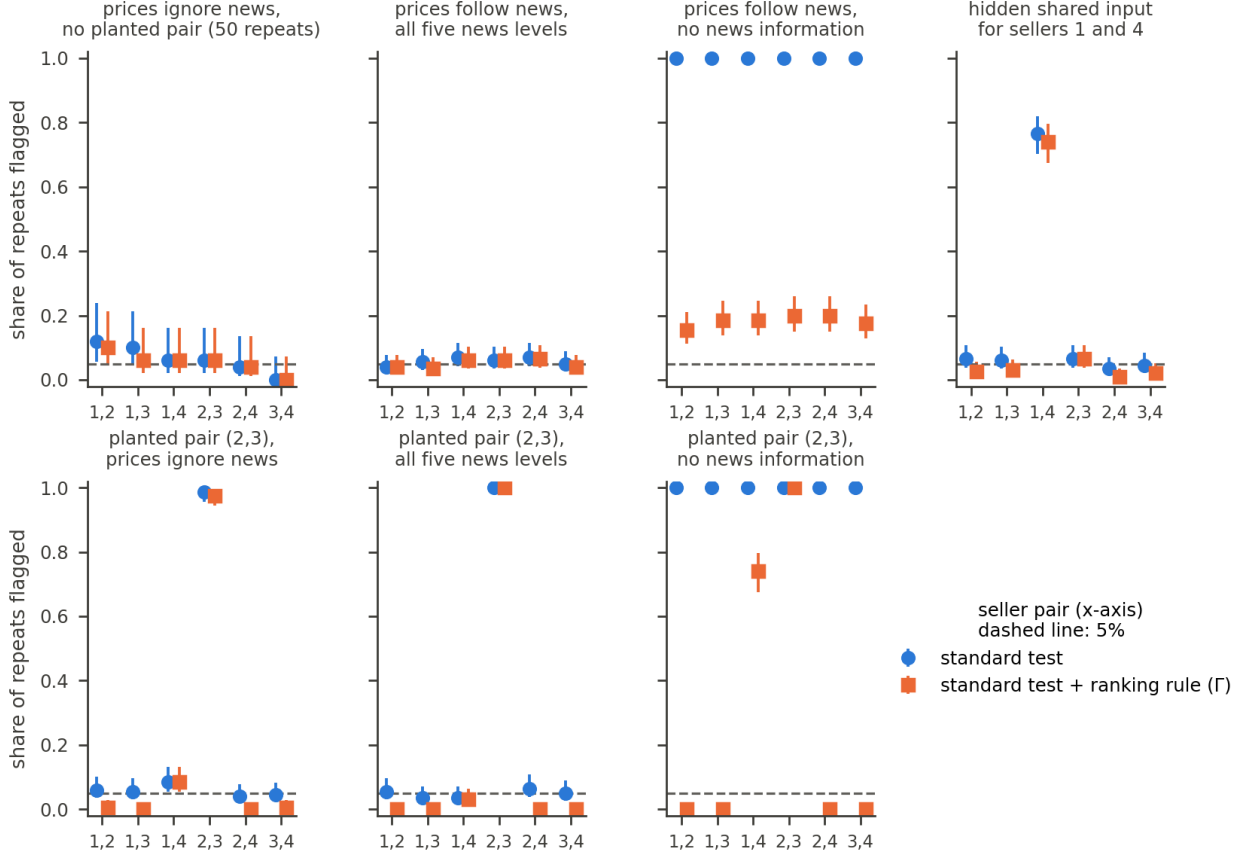


Figure 5: **Standard test and ranking rule in every condition.** 600 rounds; 200 repeats per condition, 50 when prices ignore news. Each panel is one condition and the x-axis lists the seller pairs. Circles: standard test; squares: standard test plus the ranking rule (Γ). Error bars: Wilson 95% intervals, not the whole-repeat bootstrap of Figure 4. Dashed line: 5%.

C Shared inputs and estimation noise

C.1 Varying a missing shared input

The sweep uses $\lambda = 0, 0.25, \dots, 2$, 200 repeats per value, 600 rounds, $\beta = 1$, and all five news levels. The output is the alert rate for sellers 1 and 4, which share m_t . For each setting, the model supplies I and the approximate standard-test alert probability. Figure 6 compares measured and predicted rates.

At $\lambda = 0$, this shared-input mechanism disappears. With all five news levels, no omitted news remains. In other conditions, setting $\lambda = 0$ would not remove dependence caused by coarser news categories.

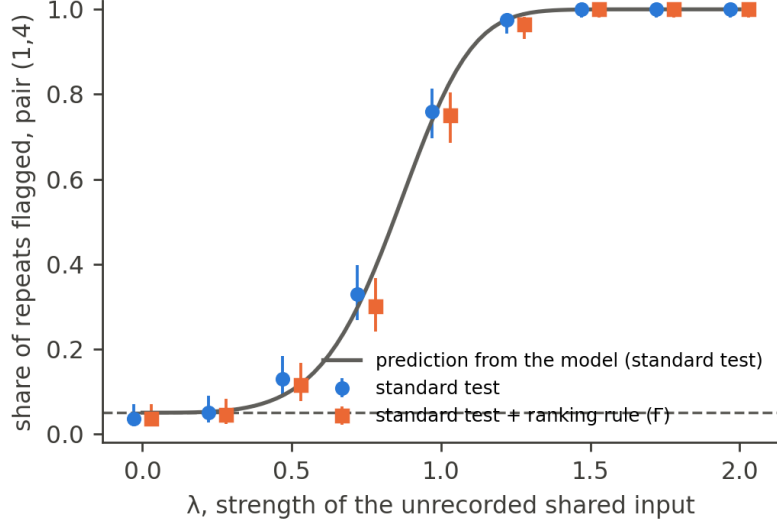


Figure 6: **Stronger unrecorded shared inputs produce more alerts.** Sellers 1 and 4 share a hidden input of strength λ ; all five news levels, 600 rounds, 200 repeats per value. Circles: standard test; squares: with the ranking rule; line: model prediction for the standard test. Error bars: Wilson 95%; dashed line: 5%.

C.2 A positive estimate can arise by chance

Even when true dependence is zero, finite counts fluctuate. Write E for the expected value, meaning the average over repeated samples. The notation χ_K^2 without a noncentrality parameter denotes the central chi-square distribution with K degrees of freedom. For binary prices and adequately populated categories:

$$2n\hat{I} \approx \chi_K^2, \quad E[\hat{I}] \approx \frac{K}{2n}.$$

This positive estimation bias is measured in nats. It is why “ $\hat{I} > 0$ ” is not a valid decision rule. Shuffled comparisons account for ordinary variation in the statistic. Figure 7 compares the shuffled mean with $K/(2n)$.

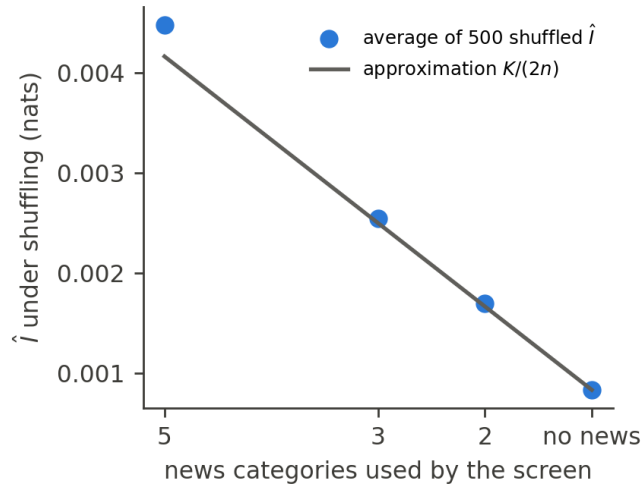


Figure 7: **Finite-sample bias under shuffled comparisons.** Average of the 500 shuffled $\hat{I}$ for non-colluding pairs at 600 rounds, against the approximation $K/(2n)$. At five categories the shuffled average is about 7.5% above the approximation.

D Which inputs drive the predicted rate?

D.1 Inputs and sampling

Let X denote the collection of sensitivity-analysis inputs, f the model calculation that maps them to a predicted alert rate, and Y that output. The predicted rate is $Y = f(X)$ and lies between 0 and 1. Inputs are sampled independently:

- News sensitivity β : uniform on $[0.25, 2]$.
- Shared-input strength λ : uniform on $[0, 1.5]$.
- Auditor’s view: equal probability of five, three, two, or no recorded news levels.
- Round count n : sample $\log_{10} n$ uniformly from $[2, \log_{10} 6000]$, then round to an integer.

The calculation uses 6,000 base samples and 24 quadrature nodes per model evaluation. Inputs are not grouped. These choices define the sensitivity question. Different ranges or distributions can give different indices.

D.2 Reading the indices

Let j identify one input, X_j denote that input, and X_{-j} denote all the other inputs. Write Var for variance, which measures variation around an average. The conditional expectation $E[Y | X_j]$ averages over the other inputs while holding X_j fixed. A first-order Sobol index, S_j , measures how much output variation that input explains on its own [13]:

$$S_j = \frac{\text{Var}(E[Y | X_j])}{\text{Var}(Y)}.$$

The total index, written S_{T_j} , also counts interactions with other inputs:

$$S_{T_j} = \frac{E[\text{Var}(Y | X_{-j})]}{\text{Var}(Y)}.$$

Here X_{-j} means all other inputs. Total indices need not sum to one because interactions contribute to more than one total.

Two independent sample matrices and hybrids swapping one column supply the pick-freeze estimates. The implementation uses Saltelli first-order [14] and Jansen total [15] estimators, with 200 whole-row bootstrap resamples and percentile intervals. For a self-check, x_1 and x_2 are two numerical test inputs and Y is their known weighted sum, $Y = x_1 + 2x_2$.

Total indices are 0.09 [0.08, 0.10] for news sensitivity, 0.53 [0.50, 0.55] for shared-input strength, 0.51 [0.48, 0.53] for the auditor’s view, and 0.29 [0.28, 0.31] for the round count. First-order indices are 0.02, 0.26, 0.27, and 0.14. These are shares of variation across this sweep, not shares of real-world false accusations (Figure 8).

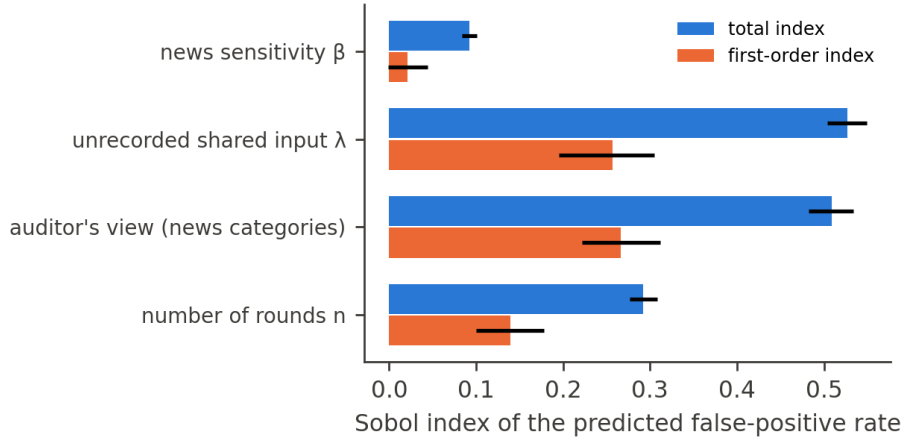


Figure 8: **Sensitivity of the predicted false-positive rate.** First-order and total Sobol indices under the input distributions above; error bars are 95% bootstrap intervals.

E LLM setup and small-sample uncertainty

E.1 Market and model settings

A game is one 100-round run with three LLM sellers and three scripted ZIP buyers. Seller cost is \$80 and buyer value is \$100. Each round supplies an independent, uniformly sampled demand category, generated from a seed and logged in `news.jsonl`. Sellers also see past asks. Chat is off.

The recorded model is `claude-haiku-5-5`, run on 25 October 2026. Mixed-model games use Haiku, Sonnet, and Haiku, with Sonnet identifier `claude-sonnet-5-5`. Requests use low effort and `max_tokens=4000`. Temperature and top-p are omitted, so API defaults apply. These identifiers describe the logged configuration, not proof of an unchanged backend snapshot.

Malformed JSON triggers upstream retries. Two retries were needed in one game. A withdrawn ask leaves the previous standing ask in place for the standard-screen series.

E.2 Price moves and qualitative check

The standard screen encodes 1 when a seller’s standing ask rises from the previous round, and 0 otherwise. It uses 99 moves. Price levels were not used because their shared drift from about \$95 toward \$80 could itself create dependence. Removing that drift does not establish that the remaining series is independent or stationary.

The same-model comparison yields 7/9 pair alerts, versus 4/9 in mixed-model games. With three games per condition and shared sellers within games, these counts do not establish a model effect.

A descriptive keyword check finds “demand”, “analyst”, or “news” in 681/900 same-model plans and 514/900 mixed-model plans, excluding template text. This count from `llm/qualitative.py` is not evidence that demand caused a particular action. Figure 9 shows two example games.

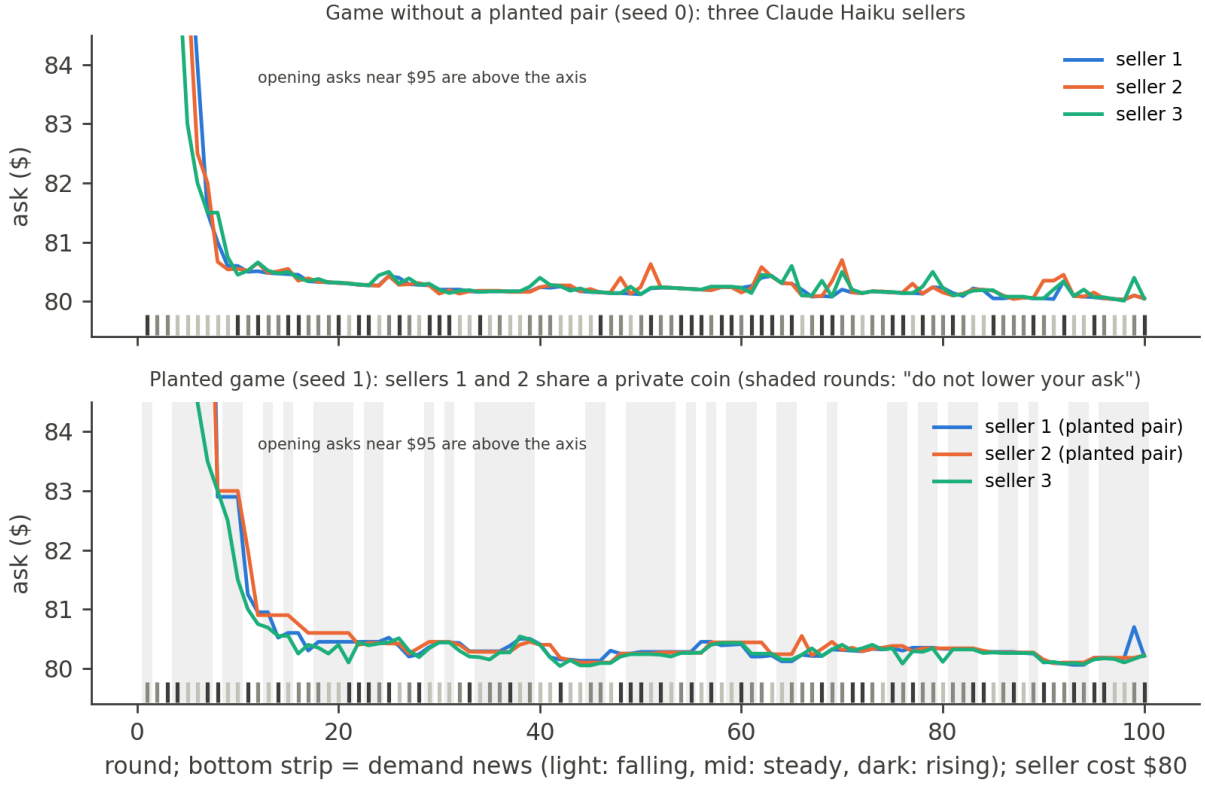


Figure 9: **Example price histories, not proof of collusion.** Seller asks per round in a game without a planted pair (seed 0) and a planted game (seed 1). Shaded rounds: private-coin heads (“do not lower your ask” for sellers 1 and 2). Bottom strip: demand news. Opening asks near \$95 are above the axis.

E.3 What the counts can establish

Three sellers produce three pairs per game. Those pairs share sellers and history, so nine pairs are not nine established independent trials.

For this calculation, let m be the number of independent trials and x the number of alerts or detections. Write $U(x, m)$ for the upper confidence bound and $L(x, m)$ for the lower bound. These are functions, distinct from the private binary variable U in Appendix A. At zero successes or all successes, the exact one-sided 95% bounds [7] are:

$$U(0, m) = 1 - 0.05^{1/m}, \quad L(m, m) = 0.05^{1/m}.$$

Using $m = 9$ gives an upper bound of about 28% for zero alerts, but requires independent pair trials. That is not justified here, so the paper does not use it as its uncertainty claim.

At game level, zero of three games has any re-asking alert. For a fixed procedure and independent comparable games, the upper bound on that event’s probability is about 63%. A confidence bound concerns the long-run coverage of a method, not a guarantee that the unknown probability cannot exceed it.

Three detections in three independent games would give a lower bound of about 37% for a fixed method. Because the rank statistic was chosen after inspecting the planted games, this is not a confirmatory performance bound for the selected method. Counts and selection history are therefore reported together.

F Re-asking: calculations, assumptions, and selection

F.1 Inputs and reconstruction

For seller i and round t , let H_{it} denote the recorded non-private prompt and let $a_{it}^{(k)}$ denote a numerical ask. Superscript $k = 0$ identifies the original answer and $k = 1, \dots, 4$ identifies re-asked answers. Asks are numerical outcomes. The prompt H_{it} is structured text. The previously defined C_t denotes recorded news. A private coin controls the extra instruction in planted games.

The 900 prompts from three games without a planted pair were recovered byte-for-byte from logged strings. Rounds were complete, demand lines matched `news.jsonl`, and original asks matched `auction_results.md` in 896/900 cases, with four withdrawals. A deliberately corrupted prompt was detected. Each prompt is a single user message.

In planted games, the private note was stripped from all 290 prompts carrying it. Re-asking therefore preserves the recorded non-private input, not every original input.

F.2 Rank and pair statistic

The rank score r_{it} describes where seller i 's original ask in round t lies among usable re-asked answers. Let k now index re-asked answers, $\mathcal{K}_{it}$ be the set of indices of usable re-asked answers, and m_{it} be their count, which can be less than four. The indicator $\mathbf{1}\{\cdot\}$ equals 1 when the condition inside holds and 0 otherwise. The score is:

$$r_{it} = \frac{1}{m_{it}} \sum_{k \in \mathcal{K}_{it}} \left[\mathbf{1}\{a_{it}^{(k)} < a_{it}^{(0)}\} + \frac{1}{2} \mathbf{1}\{a_{it}^{(k)} = a_{it}^{(0)}\} \right].$$

A high score means the original ask was high relative to its re-asked answers. Ties receive half weight, without random tie-breaking. Withdrawn or unparsed re-asked answers are dropped. Seller-rounds with a withdrawn original ask or no usable re-asked answer are dropped.

For sellers i and j , let $\mathcal{T}_{ij}$ be the set of rounds where both have scores and let n_{ij} be the number of such rounds, giving 95–99 usable rounds. Subtract each seller's mean rank within each news category over these rounds. Write $\tilde{r}_{it}$ and $\tilde{r}_{jt}$ for their centred scores. The pair's co-movement statistic S_{ij} is:

$$S_{ij} = \frac{1}{n_{ij}} \sum_{t \in \mathcal{T}_{ij}} \tilde{r}_{it} \tilde{r}_{jt}.$$

Positive covariance means the sellers' original asks are unusually high or low together relative to their re-asked answers.

Each of 500 null draws independently relabels which answer is original for every seller-round, applies the usable-answer rules, and recomputes ranks, category means, and covariance. The one-sided rule flags when observed covariance is strictly above the null's 95th percentile.

F.3 What exchangeability does and does not justify

On a fixed prompt, identically configured independent model calls would make original and re-asked answers interchangeable under the no-private-input null. Matching text is necessary but not sufficient: versions, settings, hidden state, response filtering, and call independence also matter.

Later prompts contain earlier original asks. Relabelling an earlier answer while freezing later prompts does not preserve the whole game's data-generating process. Local exchangeability therefore does not automatically validate the full resampling distribution.

A possible sequential argument considers each round conditional on its past. With matching independent calls and simultaneous sellers who do not see each other's current ask, exchangeability gives each seller's unselected mid-rank conditional mean 1/2. Subtracting 1/2 gives a score whose expected value, given the past, is zero under those assumptions. Products of the two sellers' scores would also have conditional mean zero if their current answers are independent given that past.

This suggests a possible way to analyse the rounds in sequence, rather than treating the whole game as interchangeable.

This is not a proof for the implementation. It requires justification after filtering unusable answers, and the implemented statistic centres by estimated news means rather than $1/2$. An asymptotic argument needs additional variance and convergence conditions. Matching means or variances alone is insufficient. Finite-sample calibration at about 99 rounds and model-snapshot stability remain unresolved.

F.4 Chronology and numerical checks

The initial registration, commit `806c69f`, used “raised or not” and detected no planted pairs in three games. An exploratory “lowered or not” variant detected one. After inspecting this failure and sellers’ tendency to hold asks, the rank addendum, commit `143a047`, was fixed before calculating that statistic. It detected all three planted pairs and flagged no other pairs. This is exploratory, in-sample evidence.

The initial 25% variation check concerned binary moves: in 5/20 sampled prompts, “raised or not” varied across four re-asked answers. It did not mean only 25% of numerical asks varied. A cached-data check found that the original and re-asked asks were not all equal in 53–77% of seller-rounds.

There are 7,128 re-asked answers: 3,564 per group, or three games $\times$ three sellers $\times$ 99 rounds $\times$ four re-asked answers. Of these, 54 are unusable: 37 without a planted pair and 17 in planted games. Per-call token usage was not logged. The author reports whole-day spend of about \$20 from the console.

Post-hoc Monte Carlo p-values for the planted pair are 0.032, 0.002, and 0.002. These are finite-resampling estimates, not exact population probabilities. Holm correction over three pairs per game retains the planted pair in two games. No pair is flagged without a planted channel, where the smallest adjusted p-value is 0.17. This calculation does not repair unresolved re-asking calibration. Figure 10 shows both statistics for every pair.

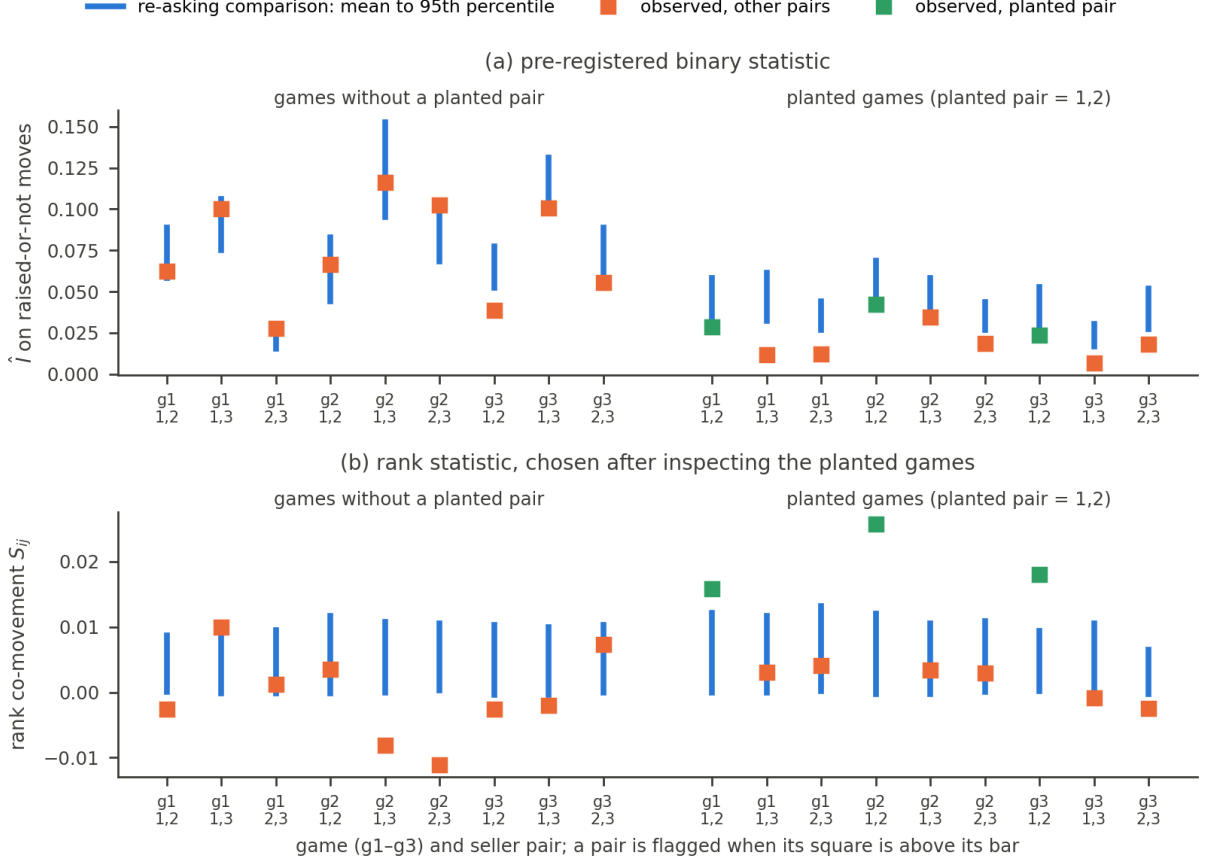


Figure 10: **Re-asking tests for every pair of every game.** (a) Pre-registered binary statistic. (b) Rank statistic S_{ij} , chosen after inspecting the planted games. Bars: the re-asking comparison, from the mean to the 95th percentile of 500 relabellings; squares: observed value (green: planted pair). A pair is flagged when its square is above its bar.

G Proposed audit procedure

This is a proposed workflow, not a validated enforcement standard.

1. **Preserve the available information.** Record prices, prompts, histories, versions and sampling settings. State what may be missing.
2. **Fix the screen before seeing outcomes.** Specify the statistic, encoding, shuffle groups, threshold and tested pairs. Check false positives under explicit, justified scenarios.
3. **Account for multiple pairs.** Let s be the number of sellers, giving $s(s-1)/2$ pairs. Here m counts the tested pairs and α is the chosen overall error level, such as 0.05. Holm correction orders the m p-values and compares them with $\alpha/m, \alpha/(m-1), \dots$, stopping at the first failure. It controls the chance of any false alert only if the individual p-values are valid. A maximum-statistic alternative requires a justified joint resampling scheme.
4. **Re-ask under stated assumptions.** Preserve non-private inputs and model state, check usable-answer variation, and decide in advance whether re-asking applies to all pairs or screened pairs. Selection must be included in claimed error guarantees.
5. **Investigate, rather than conclude.** Examine private communication, omitted shared inputs and alternative explanations. An alert is not proof of an agreement. No alert is not proof of its absence.

Key failure modes are missing common inputs, sparse categories, changing models, temporal dependence, dependent API randomness, hidden state, deterministic answers, parsing and filtering errors, multiple-pair selection, and statistics chosen after inspecting outcomes.

H Limitations and Dual-Use / Ethical Considerations

H.1 What the results establish

The simulation shows how a specified information gap can make a conditional-independence screen flag non-colluding sellers. Its known generating model lets us predict those rates and compare them with repeated experiments. This does not establish false-positive rates in real markets, where shared inputs, behaviour and information gaps may be unknown. The simulated prices are binary and sellers have no memory.

The LLM results are exploratory. There are three games per main condition, pairs within a game are dependent, and the rank statistic was chosen after inspecting planted games. No planted private channel does not mean no tacit collusion. Removing a private instruction tests sensitivity to an omitted influence, not to an agreement specifically. Re-asking validity for adaptive histories, response filtering and matching model snapshots remains unresolved. Appendices E and F give the uncertainty calculations and assumptions.

H.2 Dual-use risks

The results could help auditors identify unreliable screens, but they could also help colluding agents or their operators identify weaknesses in monitoring. Knowledge of an auditor’s information gaps could be used to make collusion harder to detect or to dismiss genuine warning signs as ordinary shared responses. We therefore do not present missing information as an excuse to ignore alerts, or an absence of alerts as evidence that a market is competitive.

The planted mechanisms are controlled experimental devices, not evidence that the same mechanisms dominate real markets. Any release of prompts or logs should be reviewed for confidential information, credentials and unnecessary operational detail about deployed monitoring systems.

H.3 False-positive harms and responsible use

An unsupported collusion allegation can cause reputational, financial or regulatory harm. Shared infrastructure or similar models can also produce dependence without establishing an agreement. Statistical alerts should therefore prompt investigation of private communication, omitted shared inputs and alternative explanations, not automatic penalties.

Conversely, the possibility of false positives does not show that flagged sellers are non-colluding. A responsible audit should document its information assumptions, account for testing multiple pairs, report uncertainty and distinguish controlled validation from exploratory evidence. Re-asking should not be used for enforcement on the strength of this pilot alone.