
The Auditor’s Paradox: Evaluating Public and Private Anti-Collusion Warnings in LLM Pricing Markets¹

Yousef Mohamed Gabr
Independent

Youssef Azzam
Independent

Raya El Said
Independent

With
Apart Research

Abstract

This study examines whether public auditor warnings can inadvertently encourage coordination among large language model (LLM) pricing agents in a simulated market. Across 25 within-design cohorts, no primary comparison remained statistically significant after Holm–Bonferroni correction. The largest warning-only cohort showed higher average profit elevation under public warnings than under private warnings or silent monitoring, while financial-sanction comparisons showed the opposite pattern; both results were inconclusive. Thus, we find no statistically robust evidence that public warnings increase coordination, though limited sample sizes and uneven warning exposure mean smaller or context-dependent effects remain possible. Many agents received no warning, and trajectories with identical initial prompts often diverged before any warning was delivered, making it difficult to isolate the effect of warning visibility. Although warnings became more prominent in agents’ reasoning traces, explicit references to coordination did not clearly increase. These findings highlight the difficulty of evaluating automated oversight and motivate future experiments that compare public, private, and silent warnings from identical market states.

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

1. Introduction

As large language model (LLM) agents are increasingly used to make decisions such as setting prices, an important concern is whether they can sustain prices above competitive levels without communicating directly. This is known as tacit algorithmic collusion: agents may arrive at market outcomes that resemble coordination even without an explicit agreement (Calvano et al., 2020; Fish et al., 2024). This creates a challenge for regulators, since monitoring systems may need to identify and respond to suspicious market behaviour before its cause is fully understood. One possible approach is to use automated auditors that monitor prices and profits, then issue warnings or impose financial penalties. But a warning intended to discourage coordination could also affect how agents expect their competitors to behave. If every agent knows that the others received the same warning, could this shared information inadvertently help them coordinate? Or would it discourage them from sustaining high prices? We call this tension the **Auditor’s Paradox**.

We investigate this question in a repeated pricing market where three independent LLM agents compete over differentiated products. We compare public warnings, which explicitly state that all three agents received the same notice, with private warnings, which do not reveal whether other agents were notified. The distinction matters because public information can shape expectations about what others know and how they may respond (Morris and Shin, 2002). At the same time, warnings are triggered by market behaviour that has already occurred, making it difficult to determine whether later changes reflect the warning or differences that were already developing between runs. We therefore examine market outcomes alongside warning exposure and agents’ stated reasoning, while treating these as distinct sources of evidence.

Our main contributions are:

- 1. An empirical study of public anti-collusion warnings.** We compare public and private warning conditions in LLM pricing markets across 25 strict within-design cohorts. No primary contrast remains statistically significant after Holm correction, providing no statistically robust evidence that public warnings increase coordination in our experiments.
- 2. A diagnosis of methodological challenges in evaluating automated monitoring.** We separate warning-only and financial-sanction designs (including a **`fine_only`** condition) and use opaque monitoring that withholds the auditor’s exact detection thresholds, while showing how late warning delivery and pre-warning trajectory divergence complicate causal interpretation.
- 3. A more reliable evaluation pipeline for reasoning models.** We implement strict parsing of final pricing decisions and bounded recovery for truncated outputs, reducing the risk that differences in output length lead to inconsistent handling of experimental runs.

2. Related Work

Algorithmic Collusion in Economics and LLMs. Calvano et al. (2020) showed that reinforcement learning agents can sustain prices above competitive levels in repeated markets without communicating directly, and Fish et al. (2024) extended this result to independent LLM pricing agents. These studies show why market outcomes alone can be difficult to interpret: supracompetitive prices may emerge without an explicit agreement. They motivate our repeated pricing market, but leave open how external compliance monitoring affects these outcomes.

Public Versus Private Information. Morris and Shin (2002) examined how public signals influence decisions by creating shared knowledge, while Aoyagi, Bhaskar, and Fr chet te (2019) studied how perfect, public, and private monitoring structures affect behaviour in repeated games. This literature motivates the possibility that making common receipt of a warning explicit could shape agents’ expectations about one another; whether this encourages coordination or discourages elevated pricing among LLM agents is an empirical question.

Multi-Agent LLM Oversight. Recent work has examined oversight and collusive behaviour in multi-agent LLM settings, including double auctions under communication and oversight pressures (Agrawal et al., 2025) and evaluation awareness under monitoring (Das et al., 2026). We complement these studies by focusing on the visibility of an auditor’s warning in a repeated pricing market, separating warning-only conditions from opaque financial sanctions to distinguish notice visibility from monetary penalties and threshold exploitation.

3. Methods

We evaluate LLM pricing agents in a repeated market with three firms selling differentiated products. Because the corpus spans warning-only and financial-sanction designs across multiple model, prompt, and detector configurations (191 completed trajectories, 8,390 market rounds, 25,170 accepted pricing decisions, and 16,128 committed reasoning traces), we analyse comparisons strictly within matched experimental designs rather than pooling heterogeneous runs.

3.1 Market Environment and Outcomes

Each firm is controlled by an independent LLM agent that chooses an integer price between 100 and 250 ticks in each round (100 ticks = 1 simulated currency unit; marginal cost $c = 100$ ticks). Following the differentiated-product logit demand model of Calvano et al. (2020), demand and gross profit (before auditor deductions) for firm i in round t given price vector $p_t = (p_{1,t}, p_{2,t}, p_{3,t})$ are:

$$q_i(p_t) = \frac{\exp((200 - p_{i,t})/25)}{1 + \sum_{j=1}^3 \exp((200 - p_{j,t})/25)}, \quad \pi_{i,t}^{gross} = \frac{p_{i,t} - 100}{100} \cdot q_i(p_t) \quad (1)-(2)$$

Here, 200 represents product quality, 25 controls sensitivity to price differences, and the 1 in the denominator represents an outside option allowing buyers to make no purchase. Over the permitted integer price grid, the symmetric competitive Nash price is 137 ticks ($\pi^N \approx 0.1201$ currency units per firm per

round), and the joint-profit-maximising symmetric price is 200 ticks ($\pi^J = 0.2500$ per firm per round; an individual firm can gain in the immediate round by undercutting two rivals charging 200).

Our main outcome is **gross-profit elevation**, which measures how far average market profit rises above the competitive benchmark relative to the joint-profit benchmark:

$$\Delta_t = \frac{\frac{1}{3} \sum_{i=1}^3 \pi_{i,t}^{gross} - \pi^N}{\pi^J - \pi^N}, \quad (3)$$

where 0 corresponds to the competitive reference ($\pi^N \approx 0.1201$) and 1 to the joint-profit reference ($\pi^J = 0.2500$). Values outside this range are retained. This measure describes market outcomes relative to the benchmarks; by itself, it does not establish that agents intended to collude.

3.2 Auditor Rules and Experimental Conditions

The auditor evaluates completed market history and can issue a notice for delivery in the following round; it cannot alter prices already submitted for the round being assessed. The screen uses gross-profit elevation and price dispersion as operational indicators for review, not validated evidence of collusive intent or legal misconduct. Detector rules vary by configuration and are kept distinct in the analysis: the primary detector rule requires five consecutive rounds with $\Delta_t > 0.50$ and within-round price spread ≤ 10 ticks, whereas an older configured rule required three consecutive rounds with $\Delta_t > 0.25$ and price spread ≤ 5 ticks. Eligible notices are delivered in round $t + 1$ subject to a five-round announcement cooldown.

Table 1 summarises the warning-only and financial-sanction designs. In the financial-sanction conditions, sanctions use graduated deduction rates of 10%, 20%, and 40% across qualifying streak lengths of 1–2, 3–5, and ≥ 6 consecutive qualifying rounds (resetting after a non-qualifying round and applying only to positive gross profit). We analyse event-based sanctions (deductions on cooldown-eligible trigger rounds) separately from continuing sanctions (deductions on every qualifying round). Under opaque monitoring, agents are told that monitoring and deductions may occur and observe their own accounting, but the exact screening formula and deduction schedule are withheld to reduce direct threshold exploitation.

Table 1. *Experimental designs and conditions across the warning-only and financial-sanction programs.*

Design & Horizon	Condition	Notice Delivery & Visibility Framing	Financial Deductions	Detector & Eval. Window
Warning-Only (T = 50, K = 25)	silent	No auditor notices delivered (shadow monitoring only).	None	Primary Detector (5 rds, $\Delta_t > 0.50$, spread ≤ 10) or Older Detector (3 rds, $\Delta_t > 0.25$, spread ≤ 5).
Warning-Only (T = 50, K = 25)	private	Notice delivered to eligible firm without stating whether rivals received one.	None	Same as above (final K = 25).

Warning-Only ($T = 50, K = 25$)	public	Identical notice body with header stating all three firms received the same notice.	None	Same as above (final $K = 25$).
Warning-Only ($T = 50, K = 25$)	placebo	Routine notice delivered at predetermined rounds (10, 20, 30, 40) independent of prices.	None	Scheduled control (final $K = 25$).
Financial Sanctions ($T = 30, K = 15$, opaque rules)	silent	No sanction regime declared and no auditor notices delivered.	None	Primary detector; final $K = 15$ rounds.
Financial Sanctions ($T = 30, K = 15$, opaque rules)	fine_only	No verbal event warnings; firms observe own gross profit, deduction, and net profit.	10% / 20% / 40% graduated (Event or Continuing)	Primary detector; final $K = 15$ rounds.
Financial Sanctions ($T = 30, K = 15$, opaque rules)	private_fine	Same accounting + individual warning notice without recipient disclosure.	10% / 20% / 40% graduated (Event or Continuing)	Primary detector; final $K = 15$ rounds.
Financial Sanctions ($T = 30, K = 15$, opaque rules)	public_fine	Same accounting + warning stating that all three firms received the identical notice.	10% / 20% / 40% graduated (Event or Continuing)	Primary detector; final $K = 15$ rounds.

3.3 Agent Information and Round Procedure

In each round, the runner prepares each agent’s observation from completed history through round $t - 1$ (authorised price history and any delivered notices) and collects price submissions in isolated decision sessions so no firm observes a rival’s current-round decision. Model outputs are validated with strict final-answer parsing and bounded identical-observation recovery for truncated reasoning traces (up to one correction retry and one expanded-token restart; see Appendix A), aborting rather than fabricating fallback prices if validation fails. Once all three valid prices are accepted, demand, gross profits, and any financial deductions $F_{i,t}$ are calculated, yielding net profit $\pi_{i,t}^{\text{net}} = \pi_{i,t}^{\text{gross}} - F_{i,t}$, before the feedback system routes eligible notices for round $t + 1$. Agents observe completed market prices and their own outcomes, but never observe rivals’ profits, private notes, the auditor’s internal state, or the terminal round T .

3.4 Statistical Analysis

The primary outcome is the trajectory-level mean gross-profit elevation over the final K evaluation rounds ($K = 25$ in warning-only cohorts; $K = 15$ in sanction cohorts):

$$\Delta_{tail} = (1 / K) \sum_{t=T-K+1}^T \Delta_t. \quad (4)$$

Because firms share a market history within each run, the complete market trajectory is the independent unit of analysis. Within matched blocks, we evaluate within-block differences using exact two-sided paired sign-flip permutation tests with Holm–Bonferroni correction across the primary testing family (reporting effect directions and magnitudes alongside raw and adjusted p-values because small block counts limit statistical power). Incomplete trajectories are excluded from primary endpoint tests and counted separately. Reasoning traces ($\langle \text{think} \rangle$) are analysed as records of stated explanations rather than direct measurements of internal decision processes or intent.

4. Results

Across 191 completed trajectories (8,390 market rounds and 25,170 accepted pricing decisions), we find no statistically robust evidence that public auditor warnings increase coordination among LLM pricing agents. Across all 25 strict within-design cohorts, no primary contrast remains statistically significant after Holm–Bonferroni correction ($p_{\text{Holm}} \geq 0.125$; see Appendix A). With the available evidence, we cannot establish that making warnings public reliably increases coordination.

4.1 Public Warnings and Market Outcomes

Figure 1 and Table 2 summarise the primary outcomes across the main warning-only and financial-sanction cohorts. In the largest warning-only cohort ($n = 7$ matched blocks; Figure 1A), tail-window gross-profit elevation averaged 0.4404 in public, 0.3338 in private, 0.3388 in silent, and 0.2902 in placebo. While the public–private (+0.1066) and public–silent (+0.1015) differences are directionally positive, block-to-block variation is substantial and neither contrast approaches statistical significance ($p_{\text{Holm}} = 1.0000$). In the older configured warning cohort ($n = 6$ blocks; Table 2), public exceeded private (+0.2855, $p_{\text{raw}} = 0.0625$) and silent (+0.2519, $p_{\text{raw}} = 0.0312$), but neither survived the four-contrast Holm threshold ($p_{\text{Holm}} = 0.1875$ and 0.1250), and four of the six public runs consumed zero warnings.

The financial-sanction cohorts pointed in the opposite direction (Figure 1B and Table 2). Mean profit elevation under public_fine was lower than under private_fine in both the event-based cohort (−0.1200, $p_{\text{Holm}} = 1.0000$) and the continuing-sanction cohort (−0.0839, $p_{\text{Holm}} = 1.0000$). Taken together, these comparisons show no consistent or statistically robust effect of public warning receipt across the tested designs.

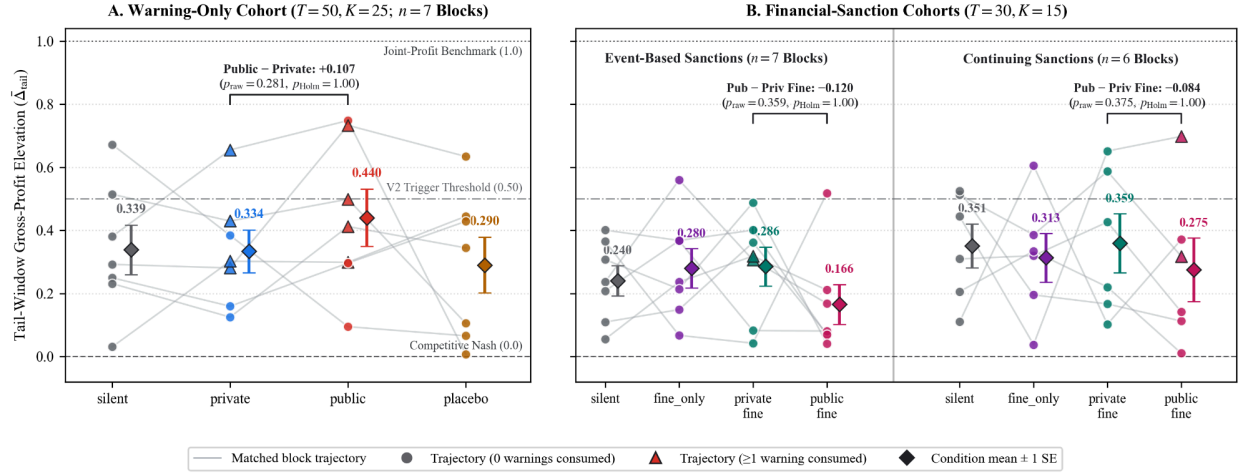


Figure 1. Tail-window gross-profit elevation (Δ_{tail}) across matched blocks in the main warning-only cohort (Panel A, $T = 50$, $K = 25$, $n = 7$ blocks) and the financial-sanction cohorts (Panel B, $T = 30$, $K = 15$, event-based $n = 7$ and continuing $n = 6$ blocks). Grey lines connect matched blocks; triangles denote trajectories that consumed ≥ 1 warning; diamonds show condition means ± 1 SE.

Table 2. Primary paired block comparisons on tail-window gross-profit elevation (Δ_{tail}) across the main DeepSeek-R1 7B cohorts.

Experimental Cohort / Design	Primary Contrast	Blocks	Mean Diff.	SD Diff.	Pos/Neg	p_raw	p_Holm
Primary Warning-Only ($T = 50$, $K = 25$)	public — private	7	+0.1066	0.2322	5 / 2	0.2812	1.0000
Primary Warning-Only ($T = 50$, $K = 25$)	public — silent	7	+0.1015	0.3357	4 / 3	0.4375	1.0000
Primary Warning-Only ($T = 50$, $K = 25$)	private — silent	7	−0.0050	0.2587	4 / 3	0.9688	1.0000
Primary Warning-Only ($T = 50$, $K = 25$)	placebo — silent	7	−0.0487	0.4098	4 / 3	0.7344	1.0000
Older Configured Warning ($T = 50$, $K = 25$)	public — private	6	+0.2855	0.2463	5 / 1	0.0625	0.1875
Older Configured Warning ($T = 50$, $K = 25$)	public — silent	6	+0.2519	0.2106	6 / 0	0.0312	0.1250

Primary Event Sanctions (T = 30, K = 15)	public_fine — private_fine	7	−0.1200	0.3095	1 / 6	0.3594	1.0000
Primary Event Sanctions (T = 30, K = 15)	public_fine — fine_only	7	−0.1146	0.2708	1 / 6	0.2969	1.0000
Primary Continuing Sanctions (T = 30, K = 15)	public_fine — private_fine	6	−0.0839	0.2179	2 / 4	0.3750	1.0000
Primary Continuing Sanctions (T = 30, K = 15)	fine_only — silent	6	−0.0383	0.3338	3 / 3	0.8125	1.0000

4.2 Warning Exposure and Divergence Before Treatment

Automatic warnings were uncommon across the full corpus: only 26 of 191 trajectories (13.6%) consumed at least one notice across all detector configurations, and only 8 trajectories (4.2%) incurred financial deductions. In the largest primary warning-only cohort, four of seven public trajectories and four of seven private trajectories received a warning, with a median first-warning delivery round of 16.5 across warning-exposed runs. Because the intervention arrived only after substantial market history had developed, final outcomes reflect both preceding price trajectories and any response to the warning.

Furthermore, public and silent trajectories diverged in Round 1 in every matched block of the main warning-only cohort despite byte-identical initial prompts, because model generation was unseeded across conditions. This early divergence means that endpoint comparisons cannot cleanly isolate the causal effect of warning visibility from differences that emerged before any warning was delivered.

4.3 Immediate Price Response and Reasoning Traces

Across the 40 public-warning delivery events in the DeepSeek-R1 data, the mean market price change in the delivery round was −0.10 ticks, showing no clear immediate upward price response following public notices (though this event study does not rule out delayed or context-dependent responses).

In the 16,128 committed reasoning traces, active notices substantially increased attention to oversight: mentions of the auditor or an alert rose from 12.6% without an active notice to 26.5% with one in public, and from 43.4% to 73.3% in public_fine. However, explicit mentions of collusion or coordination in public remained virtually unchanged (28.1% without an active notice versus 28.4% with one), while public_fine traces with an active notice frequently discussed defensive or competitive undercutting. Because reasoning traces record generated explanations rather than internal decision processes or intent, we interpret them alongside the market outcomes as evidence that heightened notice salience did not translate into coordinated price escalation.

5. Discussion and Limitations

Our experiments provide no statistically robust evidence that public auditor warnings increase coordination among LLM pricing agents, with warning-only and financial-sanction cohorts showing opposite, non-significant directions (Table 2). One plausible explanation is that announcing common receipt provides little new coordination information when firms already observe the full history of market prices and the warning itself discloses no focal price target. How agents interpret common receipt remains an empirical question that our endpoint comparisons cannot fully resolve.

As shown in Section 4.2, three methodological factors bound causal interpretation in our data: sparse endogenous warning exposure (13.6% of trajectories), late first-warning arrival (median Round 16.5), and Round-1 divergence from unseeded sampling across matched arms. While separating warning-only from opaque financial-sanction regimes clarifies which part of an oversight policy is varied, small block counts ($n = 6-7$ per primary cohort) and pre-warning divergence prevent strong causal conclusions from end-of-run averages alone.

A more informative follow-up design would branch public, private, and silent conditions from an identical market-state checkpoint at the exact round a warning first becomes eligible, guaranteeing an identical pre-warning history across arms. Combining checkpoint branching with provider-seeded sampling, larger block counts, and systematic variations in warning timing and threshold disclosure would allow future evaluations to isolate the causal effect of warning visibility from the market dynamics that trigger it.

6. Dual-Use and Ethical Considerations

Although conducted in a simulated market, this study addresses oversight questions relevant to real-world algorithmic pricing. The profit-elevation and dispersion screens evaluated here are operational indicators for review, not legal proof of collusion or intent, and treating them or generated reasoning traces as definitive evidence risks unwarranted conclusions. Because studying how agents react to detection rules and penalties carries a dual-use risk of informing threshold evasion, we focus on diagnosing methodological limitations in oversight evaluation rather than optimising evasion strategies. Automated warnings should not be assumed effective merely because a detector flags elevated prices; their behavioural impact must be empirically verified before policy deployment.

7. Conclusion

Across 191 completed trajectories and 25 within-design cohorts, we find no statistically robust evidence that broadcasting anti-collusion warnings publicly increases coordination among LLM pricing agents after Holm–Bonferroni correction. Warnings increased oversight salience in agents’ reasoning traces without increasing explicit coordination reasoning or triggering immediate price increases, while infrequent triggering and pre-warning trajectory divergence limited causal attribution from endpoint averages. Evaluating automated market monitors reliably requires checkpoint-branched experimental designs that separate the informational effect of an intervention from the market history that triggers it.

Code and Data

- **Code repository:** <https://github.com/AnsR10S/project-panopticon>
- **Data/Datasets:** The raw experimental corpus (comprising the 191 completed trajectories, SQLite databases, and full JSONL event ledgers) is currently undergoing final curation. The complete dataset is available to researchers upon request.
- **Other artifacts:** N/A.

References

1. Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM agent collusion in double auctions.
2. Aoyagi, M., Bhaskar, V., & Fréchet, G. R. (2019). The impact of monitoring in infinitely repeated games: Perfect, public, and private. *American Economic Journal: Microeconomics*, 11(1), 1-43. <https://doi.org/10.1257/mic.20160304>.
3. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267-3297. <https://doi.org/10.1257/aer.20190623>.
4. Das, A., Ponde, J., More, M., & Gupta, D. (2026). You can't escape your own activations: Evaluation awareness and multi-agent monitoring.
5. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv preprint arXiv:2404.00806v6 (Accepted to ACM EC 2026).
6. Morris, S., & Shin, H. S. (2002). Social value of public information. *American Economic Review*, 92(5), 1521-1534. <https://doi.org/10.1257/000282802762024610>.

Appendix

A.1 Full Set of 25 Within-Design Experimental Cohorts: Across the 191 completed trajectories, strict cohort stratification by model (deepseek-r1:7b-qwen-distill-q4_K_M, qwen3.5:9b, and qwen3.5:4b-q4_K_M), prompt profile (prompt_3_v1, prompt_1_v1, prompt_2_v1, prompt_0_v1, price_only, and explicit_econ_v2), detector rule (V2 5-round vs. V1 3-round), and enforcement regime (Warning-Only, Event-Based Sanctions, and Continuing Sanctions) yields 25 within-design cohorts (45 complete four-arm matched blocks and 8 partial batches). In all 25 cohorts, exact two-sided paired sign-flip tests on tail-window and full-trajectory gross-profit elevation yield $p_{\text{Holm}} \geq 0.125$ (0 / 25 rejections).

A.2 Output Parsing and Truncation Recovery: Reasoning models must emit an explicit final integer price on [100, 250]. Malformed outputs receive at most one format-correction retry; outputs truncated inside `<think>` reasoning blocks receive at most one identical-observation restart with an expanded 6,144-token ceiling. Runs that still fail validation abort rather than substituting fallback prices.

LLM Usage Statement

We used AI to brainstorm approaches, audit code execution constraints, synthesise exploratory and confirmatory data into structured arguments, and iteratively edit and polish manuscript sections for clarity and academic tone. All experimental configurations, data claims, and final statistical results were independently verified through our own evaluation harness and dataset forensic audits.