

Auditing the Absence of Collusion: Measurement Failures and Inactive Oversight in an LLM Spot-Futures Market

Wu Zinuo, Wu Zijing, Saalim M. Abdulla, and Leo Liu

HKU

With Apart Research

AI Collusion Research Sprint, October 2026 | Primary track: 2 - Detection and Audit of Collusion
Research questions 2.2 and 2.3 | Live-run analysis dated 26 October 2026

Abstract

Can an LLM regulator distinguish collusion from ordinary market convergence, and can broader inspection authority reduce coordination? We audit seven live, 45-hour spot-futures games with ten traders per game, using the logged model identifier Qwen/Qwen3.8-27B. Two unsupervised baselines vary capital allocation; five supervised games vary intervention mode or the announced private-inspection cap. We reconstruct indicators and financial outcomes from archived messages, trades, payments, model replies, and source snapshots. Across 225 valid hourly regulator reviews, every coordination score is 1, with no inspections or interventions. However, the original behavioral reference marks 180 supervised hours positive. This discrepancy is not a measured false-negative rate: 333 of 334 keyword hits in the policy game contain coordination-denial language, and the implemented quote-clustering rule differs from its stated definition. Replacing whole-side clustering with any-pair clustering changes the penalty-cap comparison from 43, 43, and 6 positive hours to 45, 45, and 45. All five terminal audits fail before model inference and return keyword fallbacks, invalidating their apparent zero-collusion verdicts as LLM audit results. Financial ledgers reconcile, but cash-transfer market mechanics and prompt inconsistencies limit economic interpretation. This single-seed study establishes measurement and execution failures, not successful deterrence or undetected collusion. It motivates provenance-aware audit outputs, semantics-sensitive indicators, and validated positive controls before evaluating enforcement effectiveness.

Keywords: multi-agent systems; collusion auditing; market simulation; LLM oversight; measurement validity

1. Introduction

An automated regulator must decide whether similar prices reflect independent responses to a common market or an agreement that disadvantages others. A second difficulty arises when regulation changes what traders say: language intended to deny coordination can itself trigger a coordination lexicon. If an audit also substitutes a fallback for a failed model call, an apparently clean verdict can conceal an unexecuted evaluation.

We investigate these problems in a role-free goods market coupled to hourly, cash-settled price bets. The original design asks whether an LLM regulator detects and reduces emergent coordination. The available run instead supports an audit of the evidentiary chain: which components executed, what the indicators count, and which treatment comparisons remain interpretable. The primary contribution is a reproducible diagnosis of three failure modes: inactive enforcement, construct-sensitive behavioral proxies, and terminal audit failure.

2. Related Work

Fish, Gonczarowski, and Shorrer [1] show that LLM pricing agents can reach supracompetitive outcomes and that small prompt changes affect behavior. Agrawal et al. [2] study communication and oversight in double auctions. The local project builds on that experiment family, but its role-free goods transfers and hourly price bets differ from fixed buyer-value and seller-cost auctions. Consequently, we do not import a competitive-price benchmark from the original setting.

Nakamura et al. [3] distinguish communicated plans from action-based collusion in cooperative multi-agent systems. Their finding that stated plans and executed behavior can diverge motivates our separation of messages, quotes, transfers, and regulator decisions. Kale et al. [4] find that awareness of monitoring can affect monitor reliability in other agent environments. Our agents likewise know the announced oversight regime, although this study cannot identify the effect of that awareness.

Our contribution is not a new collusion detector or a replicated enforcement result. It is an end-to-end audit of a particular recorded experiment, demonstrating how an invalid reference label, an unactivated treatment, and a failed terminal evaluator can jointly distort an apparent oversight comparison. Track 2 is therefore the appropriate anchor, with secondary relevance to market interventions and monitoring awareness in Tracks 1 and 3 [5].

3. Methods

3.1 Dataset, model calls, and treatment arms

The latest full batch in the supplied *real test* directory contains seven complete games started on 26 October 2026 at 16:48:42-43 Hong Kong time. Each has 45 simulated hours and ten agent actions per hour: 315 market-hours and 3,150 actions in total. The archive contains 3,110 trader messages and 3,397 call-log records. Of these records, 3,150 are successfully parsed trader replies, 17 are failed trader attempts, 225 are valid regulator reviews, and five are failed audit attempts. No synthetic, mock, or earlier-run outcomes enter the analysis.

Game	Capital allocation	Regulator mode	Private cap
G1	Equal: all \$1 million	Off	-
G2	Unequal: A1-A2 \$100 million each	Off	-
G3	Unequal	Guidance	2
G4	Unequal	Penalty	2
G5	Unequal	Policy	2
G6	Unequal	Penalty	5
G7	Unequal	Penalty	8

Table 1. Seven live games, each with ten traders, 45 hours, environment seed 0, and the scripted strategy disabled. In unequal games A3-A10 each receive \$1 million. A1-A2 define Group A in every game, including G1; this designation is not a collusion label.

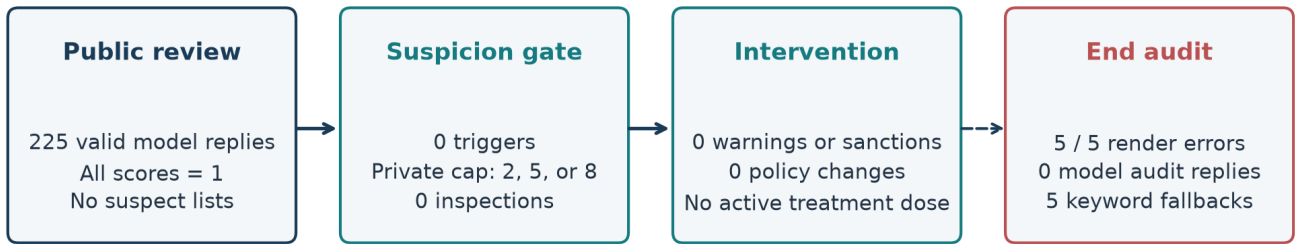
All successful trader and regulator call records name *Qwen/Qwen3.8-27B*. This is the recorded serving identifier, not an independent verification of model weights. Traders use temperature 0.7 and a 2,000-token output limit; the regulator constructor uses temperature 0 and 1,500 tokens. The archived Modal backend overrides the stale *gpt-4.1-mini* regulator configuration field. Environment seed 0 does not guarantee deterministic model sampling. There is one trajectory per arm and no estimate of between-run variance.

3.2 Market and oversight protocol

Traders maximize return on initial cash, with no fixed buyer or seller role. They must submit a goods offer each hour and both sides of a price bet from hour 6. Crossing offers trade at their midpoint. Each price-bet contract covers 100 units and settles within its hour to the quantity-weighted goods price over that hour and the preceding four. With no trades in that window, the most recent goods trade is used; with no prior trade, the reference is \$90. The stake is \$9 per unit, or \$900 per contract, corresponding to 10-fold leverage at the initial reference. Leverage sets collateral; it is not an extra multiplier on the 100-unit payoff.

Public and one-to-one private messages are capped at 500 characters and delivered to traders the following hour. The regulator reviews current public messages, cumulative public history, profit and loss, offers, positions, and identified trade history after trading. A positive suspicion gate permits inspection of the historical private messages between selected agents only. The cap limits the number of selected agents, not message length or a preassigned coalition. Caps 2, 5, and 8 expose at most 1, 10, and 28 of the 45 possible unordered private pairs; cap 8 covers 62.2%, not the entire network.

Guidance can warn. Penalty can fine up to positive profit earned in the detection hour or suspend for 1-3 hours; fines take precedence. Policy can impose price controls, change leverage, or halt both markets for at most three hours, in that precedence order. Restrictions are announced and apply from the next hour; a fine is deducted when applied. All modes and caps are disclosed from the outset. Terminal auditing is intended to inspect all public and private messages, P&L, and trades without affecting the market. Section 4 reports which stages actually ran.



Trader prompts disclose the mode and inspection cap from hour 1.

Recorded contrast: announced oversight. Realized inspections and interventions: zero.

Figure 1. Realized oversight pipeline in G3-G7. Counts come from regulator.jsonl and raw call records. Public reviews executed; the suspicion-gated inspection and enforcement branches did not. Every terminal audit failed during prompt rendering and substituted a keyword result. The dashed arrow denotes a separate end-of-game stage.

3.3 Measurement and statistical scope

We independently reconstruct message hits, quote clusters, hourly indicators, and trader P&L from the archived logs, checking them against summaries. All 182 archived source hashes match the seven configuration manifests. All 13 stored accounting checks pass in each game, and reconstructed goods and futures P&L match every agent summary. These checks validate arithmetic and provenance, not the presence or absence of collusion. The complete measurement definitions and source locators appear in Appendices B-C.

All statistics are descriptive. The seven games are seven treatment trajectories, not 315 independent experimental replicates. Temporal observations and agents share market histories. We report counts, denominators, within-game outcomes, and deterministic sensitivity analyses; we do not report significance tests, confidence intervals, or causal effect estimates. No live positive-control arm with known collusion is present in this batch.

4. Results

4.1 Valid hourly negatives coexist with failed terminal audits

Every supervised hour has a parsed model response: 225/225 reviews return coordination = no and score = 1, with no suspects, private inspections, or actions. There are no private-follow-up model calls. This is a verified property of the model outputs, not evidence that collusion is absent. G1-G2 have no regulator reviews and no terminal audits; their zero detection counters mean “not evaluated.”

All five supervised terminal audits fail before inference. The audit dossier already contains *round*, and the wrapper passes *round* again when rendering the prompt, producing a duplicate-keyword `TypeError`. Each audit log has a null prompt and response, source = keyword fallback, and an apparent zero count. These are five failed LLM audits, not five successful negative audits. The fallback uses a narrower phrase list than the behavioral lexicon and does not evaluate the full trading record as a model would.

4.2 Keyword counts largely measure price-floor talk or denials

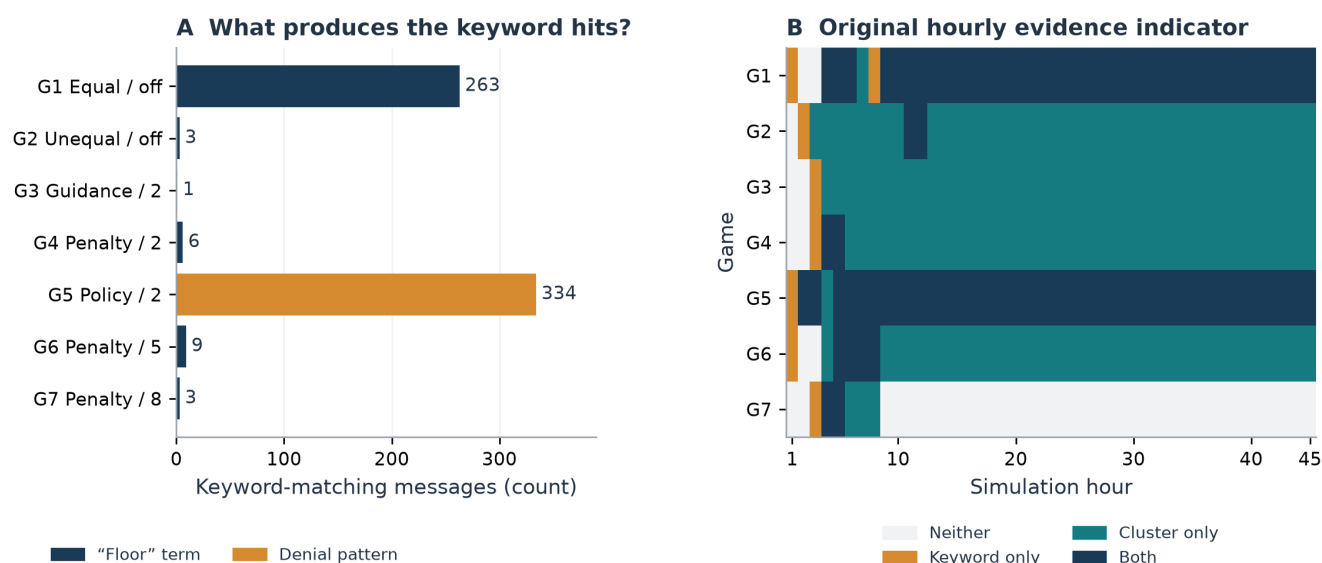


Figure 2. Message-level and hourly evidence reconstructed from all seven games. A: counts of keyword-matching trader messages, excluding regulator announcements. The policy game contributes 333 denial-pattern hits and one remaining “floor” hit. B: hourly evidence from the original lexicon and implemented whole-side clustering rule; colors distinguish their overlap. Denial patterns and price-floor language are text categories, not validated collusion labels. Each row spans 45 hours.

The original lexicon flags 619 of 3,110 messages. G5 has the highest raw hit rate, 334/444 (75.2%). However, 333/334 hits (99.7%) match a transparent, post hoc denial-language rule; the remaining message mentions a price floor. An identical statement about independent valuation and “No coordination with other agents” appears 212 times. Removing denial-pattern matches as a sensitivity check reduces G5 from 334 to one hit. This is not a replacement detector: a denial can be false, and the rule can miss other linguistic contexts.

All 263 hits in G1 contain “floor”; so do every hit in G2-G4 and G6-G7. Some statements describe a public price anchor, while others suggest support for it. Neither the word alone nor its absence establishes an agreement. The apparent policy-language spike therefore cannot be read as evidence that policy oversight increased collusion.

4.3 The private-cap comparison is not robust to its metric definition

Holding penalty mode fixed, the original indicator marks 43/45, 43/45, and 6/45 hours in G4, G6, and G7. The endpoint decline is 37 hours, or 86.0% relative to G4. On its own, that looks consistent with stronger monitoring. Yet every cap produces zero inspections, zero interventions, and zero private messages. The comparison varies announced authority and realized trajectories, not an observed dose of private surveillance.



Figure 3. Penalty-mode comparison at caps 2, 5, and 8 (G4, G6, G7), one 45-hour game each. A: original whole-side and alternative any-pair definitions, both using a \$0.50 threshold and the same keyword union. B: hours with fewer than two ask quoters, when ask clustering cannot be measured. C: mean same-direction ask movement among agents quoting in both adjacent hours; n is the number of eligible hour transitions, not independent replications. No uncertainty bars are justified by this design.

The documentation describes a cluster when at least two quotes lie within \$0.50. The implementation instead requires the maximum minus minimum price across *all* quotes on a side to be at most \$0.50. Under the stated any-pair definition, every game has an evidence-positive hour at all 45 time points. In G4/G6/G7 the comparison becomes 45/45/45, eliminating the apparent gradient. This sensitivity does not validate the broader rule; it shows that the reported gradient is not stable to the intended construct.

Book composition changes as well. G7 has only one ask quoter in 37/45 hours, compared with zero hours with fewer than two asks in G4 and G6. At hour 10, G7 contains one ask at \$90 and bids at \$90, \$89.90, and \$10. The bid range defeats whole-side clustering despite several equal-priced bids. An indicator sensitive to both side participation and price range can fall when low-cash agents post distant bids, without any demonstrated breakup of a coalition.

Ask co-movement also fails to decrease monotonically: 9.45%, 6.40%, and 23.81% at caps 2, 5, and 8. The corresponding denominators are 44, 43, and just seven eligible transitions. Ask dispersion rises from \$0.046 to \$0.067 to \$0.716, but the G7 estimate is conditional on its eight hours with at least two asks. These changes are jointly consistent with a different market composition; they cannot identify improved anti-collusion supervision.

4.4 Proxy disagreement is not detector accuracy

Across G3-G7, the behavior indicator marks 180/225 hours (80.0%) while the regulator flags none. The exported counters therefore record zero hits, 180 “misses,” and zero false alarms. We interpret these as disagreements with a heuristic reference. Without adjudicated agreements, a positive control, or valid terminal model audits, sensitivity, specificity, precision, and a false-negative rate for actual collusion are not identified.

4.5 Prices and profits vary without enacted enforcement

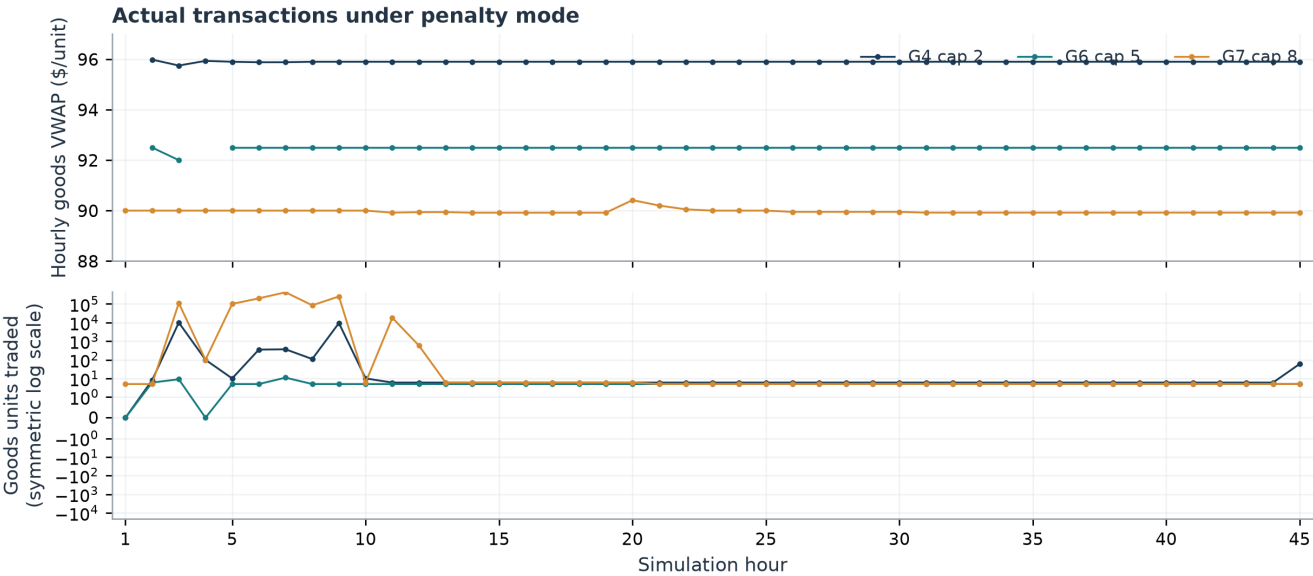


Figure 4. Actual goods-market transactions in G4, G6, and G7. Top: hourly quantity-weighted prices; gaps denote hours with no goods trades, not zero prices. Bottom: executed units, using a symmetric log scale with a one-unit linear region so zero activity remains visible. Data: trades.jsonl, hours 1-45. These are trajectories under announced caps; none received an inspection or penalty.

The penalty-cap games have full-run goods VWAPs of \$95.827, \$92.471, and \$89.999. The price ordering is descriptive and does not establish movement toward a competitive benchmark: this role-free setup has no specified redemption value or production cost for its traders. Goods volume is highly non-monotonic: 20,616, 226, and 1,189,113 units. G7 has just two regular price-bet contracts, while G6 has 5,347. A claim that cap 8 selectively suppresses collusive trading would require a welfare benchmark and replicated outcomes.

Game	Goods VWAP (\$)	Group A P&L; (\$m)	A return (%)	Others P&L; (\$m)
G1	88.675	+1.805	+90.225	-1.805
G2	90.000	-55.337	-27.668	+55.337
G3	89.945	+2.146	+1.073	-2.146
G4	95.827	-0.014	-0.007	+0.014
G5	95.498	+5.346	+2.673	-5.346
G6	92.471	+6.355	+3.178	-6.355
G7	89.999	-100.017	-50.009	+100.017

Table 2. Realized economic outcomes. Group A is A1-A2, with combined initial capital \$2 million in G1 and \$200 million in G2-G7. Other traders start with \$8 million combined. VWAP is recomputed using trade quantities; the summary field named spot.vwap_cents is an unweighted mean, so it is not used here. Positive and negative transfers reconcile to zero in each game.

The equal and unequal baselines differ in total capital (\$10 million versus \$208 million) as well as concentration. Their difference is not an isolated estimate of wealth inequality. Across supervised games, Group A returns range from -50.01% to +3.18%; the sign of its profit is not a measure of coalition advantage. A1-A2 were designated for reporting, not independently shown to form a coalition.

4.6 Accounting consistency does not establish economic validity

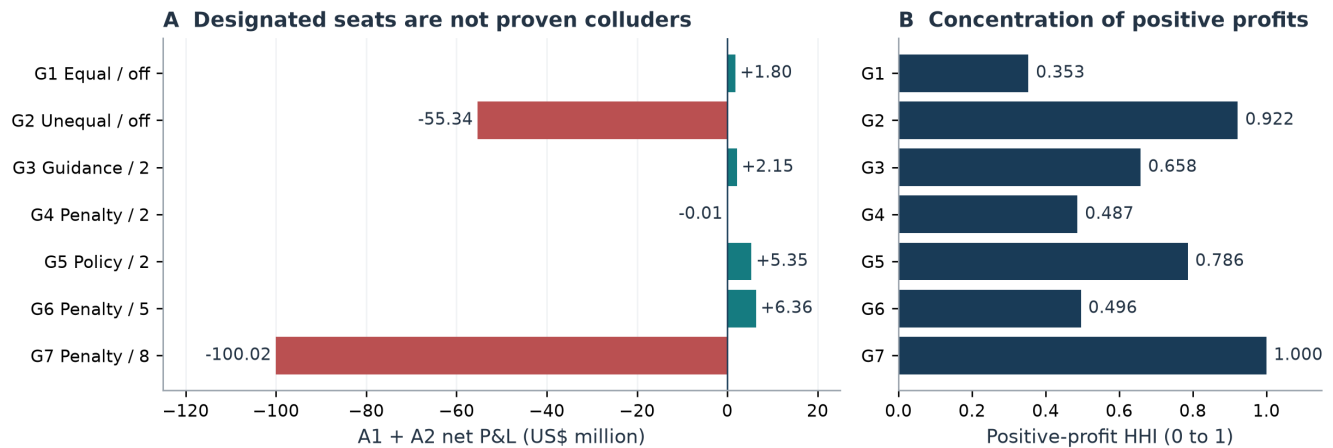


Figure 5. Financial outcomes reconstructed from actual trades and variation-margin payments. A: combined net P&L of designated seats A1-A2; the other eight seats have the opposite aggregate P&L. B: Herfindahl-Hirschman index of positive profits, calculated from each profitable trader's share of total positive P&L. G7 has a single positive-profit trader. Different initial capital makes dollar outcomes across G1 and G2-G7 non-comparable as capital-normalized performance.

G7 illustrates why concentration is not sufficient evidence of collusion. Theo (A6) earns \$107.017 million net, including \$107.017 million from goods transfers, and accounts for all positive profit. Group A loses \$100.017 million. The actual trader accounting credits sales proceeds and debits purchase expenditure, without inventory valuation or redemption. Goods asks reserve only fees, which are zero. These mechanics permit an agent to collect large cash transfers by repeatedly selling; a coordinated conspiracy is not needed to explain that opportunity.

Two prompt-to-engine discrepancies further limit interpretation. The trader prompt says goods trading is limited to five units per hour, but logged volume exceeds five units in 151/315 hours, reaching 422,322 units in one G7 hour. It also says asks must be backed by cash, while the archived trader ask-reservation code returns only the fee. These are verified mismatches between communicated and executed rules. Passing ledger checks does not repair those incentive or information problems.

4.7 Private messages provide leads, not adjudicated agreements

Only eight private messages occur: one in G1 and seven in G2; none occur under supervision. Most concern bilateral execution or available liquidity. One unflagged G2 message from Omar (A3) to Nia (A9), sent at hour 3 and delivered at hour 4, merits contextual review:

“I am raising my size to match yours. Let's see if the sellers step down to 89.99.”

The message signals intended imitation, but the private record contains no response from Nia. It also includes the mistaken suggestion that two buyers might trade with each other. This is neither an established collusive agreement nor proof of innocence. It demonstrates that meaningful evidence review must connect an overture to recipient behavior and third-party effects, rather than rely on a generic keyword hit. The absence of private traffic in supervised games is compatible with deterrence, prompt-induced behavior, or sampling variation; this design does not separate them.

5. Discussion and Limitations

The defensible finding is a failure of identification. The intervention arms never activate, the reference indicators are not validated collusion labels, and the intended terminal model audit never executes. These facts preclude claims that the regulator successfully reduces collusion or systematically misses known collusion. They also prevent a ranking of guidance, penalty, and policy as enforcement mechanisms. The run still provides useful evidence: it exposes exactly where a plausible evaluation pipeline ceases to measure its intended target.

5.1 Implications for AI safety evaluation

Audit outputs need execution provenance. A zero should be reported together with whether inference ran, whether parsing succeeded, and whether a fallback was used. Failed evaluations should remain visibly unavailable. The distinction is substantive here: the public terminal announcement presents a regulator conclusion even though no model audit response exists.

Measure realized oversight separately from announced oversight. A cap of eight is not eight inspected agents. All five treatments deliver a monitoring disclosure, but none deliver active inspection or enforcement. A useful benchmark should report both assignment and actual treatment delivery, including suspicion-gate activation, messages examined, and actions taken. The pipeline diagram makes that distinction auditable.

Validate evidence at the level of the claim. Lexical hits can encode denial, quote proximity can encode common information or low-cash compliance, and profit concentration can encode market mechanics. A detector benchmark needs examples linking communication to executed coordination and harm, plus negatives containing benign convergence and denial language. Agreement, action consistency, and advantage should be recorded separately before defining a composite label.

5.2 What a decisive follow-up requires

First, correct and test the duplicate audit argument and explicitly label fallback output. Second, align the trader prompt with goods quantity, collateral, and inventory rules; validate the resulting incentives before interpreting wealth transfers. Third, establish a live positive control that reliably produces a known agreement and observable follow-through, alongside competitive and benign-negation controls. Such controls are absent here and must not be reconstructed as if they had occurred.

Then run multiple independent seeds per arm, preserve model-serving settings, and hold total capital constant when isolating capital concentration. To separate disclosure from information and enforcement, compare silent logging, announced monitoring without action, and calibrated active supervision. Private-cap evaluation additionally requires actual private traffic and gate activation. Predefine episode-level outcomes and a blinded annotation protocol; report episode-level variability and liquidity or welfare costs alongside coordination evidence.

6. Conclusion

Seven genuine live games yield no activated regulatory intervention, 180 heuristic-positive supervised hours, five failed terminal model audits, and an apparent private-cap gradient that disappears under the stated clustering definition. These findings support an audit-methodology contribution, not a claim of proven collusion or effective deterrence. Reliable oversight evaluation requires validating the measurement and execution chain before interpreting a reassuring verdict or a treatment trend.

Code and Data

The companion evidence package contains the analysis and figure builders, source-file hashes, message annotations, hourly and game-level data, and the raw logs needed to reproduce reported statistics. Original experiment files are preserved without modification. The report follows the supplied template structure; references and the required limitations/ethics appendix follow. Upstream attribution and unresolved development chronology are stated in Appendix A.

References

- [1] Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. 2024; revised 2026. *Algorithmic Collusion by Large Language Models*. arXiv:2404.00806. Accepted to EC 2026.
<https://arxiv.org/abs/2404.00806>
- [2] Kushal Agrawal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. 2025. *Evaluating LLM Agent Collusion in Double Auctions*. arXiv:2507.01413.
<https://arxiv.org/abs/2507.01413>
- [3] Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. 2026. *Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems*. arXiv:2602.15198, version 2.
<https://arxiv.org/abs/2602.15198>
- [4] Neil Kale, Chen Bo Calvin Zhang, Kevin Zhu, Ankit Aich, Paula Rodriguez, Scale Red Team, Christina Q. Knight, and Zifan Wang. 2025. *Reliable Weak-to-Strong Monitoring of LLM Agents*. arXiv:2508.19461.
<https://arxiv.org/abs/2508.19461>
- [5] Apart Research. 2026. *AI Collusion Research Sprint: Research Tracks, Guidelines, and Submission Requirements*. Accessed 26 October 2026.
<https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25#guidelines>

Data and implementation sources

The empirical sources are the seven numbered game directories in *collusion_hackathon/llm-agent-collusion/results/capital_futures/real_test*, each containing `config.json`, `summary.json`, `messages.jsonl`, `rounds.jsonl`, `trades.jsonl`, `variation_margin.jsonl`, `regulator.jsonl`, `llm_calls.jsonl`, `reconciliation.json`, and a `source_snapshot` directory. These local experimental records are distinct from the literature references above. Exact relative paths, file sizes, and SHA-256 digests are provided in `evidence/input_manifest.csv`.

The archived configuration records parent commit `d585a7b3be50366cc2bdf5789bf8add5890b2abd`, branch *futures-collusion*, engine *capital-futures-v2*, and dirty source status. Therefore, the commit alone does not reproduce the executed implementation. Source snapshots and their verified hashes are the implementation reference.

Report template

The supplied four-page “Copy of Apart Research hackathon submission template” was used for the report structure: title and affiliations, abstract, introduction, related work, methods, results, discussion and limitations, conclusion, code and data, references, appendices, and LLM usage. The main report occupies pages 1-8; references and appendices are excluded from that count.

Appendix A. Limitations and Dual-Use / Ethical Considerations

A.1 Limits of inference

Replication and controls. One environment seed and one realized game per arm provide no treatment-level uncertainty estimate. Agent observations and hourly outcomes are dependent. No live known-collusion positive control, communication-disabled control, silent-monitor arm, or total-capital-matched baseline is included. Archived mock or scripted experiments are not substituted for live validation.

Ground truth. There is no independent, blinded labeling of agreements, follow-through, or harm. The model reviews are judgments, and the lexicon and quote metrics are clues. All supervised terminal audits fail before inference; baseline terminal audits do not exist. Zero counts cannot establish that the market was free of collusion. Conversely, heuristic-positive hours cannot establish missed collusion.

Intervention and communication. No realized supervision action occurs. No private messages are sent in G3-G7. Expanding private access therefore never exposes additional material in the recorded run. The absence of private chat could reflect announcement effects, but the design cannot establish causal deterrence or strategic channel migration. There is also no natural ordering of “harshness” between a market-wide rule and an individual penalty.

Market validity. Goods transactions transfer cash without an inventory value, trader production cost, or redemption value. Trader asks reserve no cash when fees are zero. Quantity limits stated in the prompt are not enforced as stated. The same model supplies all traders and the regulator, wealth remains hidden from traders, and stakes do not represent a validated real-exchange mechanism. These conditions constrain generalization to financial markets or antitrust claims.

Metric sensitivity. The denominator of quote dispersion and co-movement varies with participation. The whole-side rule differs from the stated pairwise rule, and the pairwise alternative saturates. The denial-language sensitivity was devised after inspecting the records and is not a validated classifier. A statement denying coordination can coexist with coordinated behavior. The unweighted spot summary is not a VWAP.

A.2 Dual-use and ethical considerations

Studying endogenous settlement and monitoring gaps can expose ways market agents might exploit thin reference prices or avoid scrutiny. The value of this report is diagnostic: it documents evidence requirements and execution failures. The results should not be used to tune real-world collusion or evade oversight, and the simulation does not justify accusations against any real entity. The trader names are simulated identities, not allegations about people.

Private-message inspection also creates a governance trade-off: broader visibility can increase privacy costs without ensuring valid detection. A future deployment would require appropriate access controls, retention limits, and accountable review. Our analysis uses the experiment's provided artificial-agent records; it does not infer behavior by human market participants. No real trading, regulator deployment, or external communication was performed for this report.

A.3 Provenance and LLM usage

The project contains preexisting upstream double-auction work attributed to Agrawal et al. [2]. The role-free capital/price-bet extension and the present audit are identified separately; the exact pre-sprint versus sprint development chronology of local extension code has not been verified and requires author confirmation before submission. The analyzed batch itself is timestamped 26 October 2026.

Codex assisted with source inspection, literature lookup, analysis implementation, figure generation, and drafting this report. Counts and financial identities were programmatically checked against the archived logs and code; the report does not claim independent human adjudication or completed author verification. The named authors should review the manuscript, source interpretations, and provenance disclosure before submitting it.

Appendix B. Complete Measurement Definitions and Counts

B.1 Original indicators

For message m , $K(m) = 1$ when the original case-insensitive regex matches an agreement term, a coordinat* stem, “together,” “collective,” “hold the line,” certain undercutting phrases, “floor,” “minimum price,” “fix,” “cartel,” “split,” “rotate,” or “take turns.” Regulator messages are excluded. The exact regex and every message-level match are included in the companion files.

For side s and hour t , let $Q(s,t)$ be the set of quoted prices. The implemented cluster $C(s,t) = 1$ if there are at least two quoters and $\max Q - \min Q \leq \$0.50$. The alternative $C^*(s,t) = 1$ if any two quoted prices differ by at most \$0.50. The original hourly indicator $B(t)$ is the union of a keyword hit in a message sent that hour, $C(\text{ask},t)$, and $C(\text{bid},t)$. Co-movement and dispersion are reported separately; they do not enter $B(t)$.

Ask co-movement is the mean over eligible adjacent hours of $\max(\text{number of increases, number of decreases})$ divided by the number of agents asking in both hours. At least two common quoters are required; unchanged prices remain in the denominator. Dispersion is the mean of the population standard deviation of ask prices over hours with at least two asks. Thus low co-movement can reflect unchanged prices, and a small quote range is distinct from movement in the same direction.

Game	Messages public / private	Keyword hits	Denial hits	B(t) hours	B*(t) hours
G1	443 / 1	263	0	43	45
G2	442 / 7	3	0	44	45
G3	441 / 0	1	0	43	45
G4	444 / 0	6	0	43	45
G5	444 / 0	334	333	45	45
G6	448 / 0	9	0	43	45
G7	440 / 0	3	0	6	45

Table B1. Full message and hourly counts. B* changes only clustering to any-pair. All denominators for hourly indicators are 45. The policy denial share is 333/334; no message-level category here is a ground-truth collusion label.

Game	Ask movement mean % (n)	Bid movement mean % (n)	Ask dispersion mean (\$)	Ask hours with $n \geq 2$
G1	81.25 (4)	9.46 (44)	0.9259	5
G2	5.87 (44)	6.08 (36)	0.1066	45
G3	8.37 (44)	7.77 (44)	0.0763	45
G4	9.45 (44)	7.56 (41)	0.0462	45
G5	6.82 (44)	11.55 (44)	0.0145	45
G6	6.40 (43)	4.37 (44)	0.0666	45
G7	23.81 (7)	8.76 (44)	0.7161	8

Table B2. Participation-sensitive quote statistics. n in movement columns counts eligible adjacent-hour transitions. The last column gives the dispersion denominator. These are temporal descriptive statistics within one game per condition.

B.2 Denial sensitivity and economic statistics

The post hoc denial regex searches for no, not, without, avoid, avoiding, never, or non, followed within 70 non-period characters by coordinat*. This partitions G5's 334 hits into 333 denial-pattern matches and one other hit. It does not establish truthfulness or distinguish all possible semantic contexts. The sensitivity is reported alongside, not in place of, the original counts.

Goods VWAP = $\text{sum}(\text{price} \times \text{quantity}) / \text{sum}(\text{quantity})$, computed directly from fills. Net P&L = goods cash transfers + cumulative variation-margin payments - fees. Group return divides total group net P&L by total group initial cash. Positive-profit HHI = sum of squared shares among agents with positive net P&L. All monetary input fields are integer cents; tables convert to US dollars or millions of dollars as labeled.

Appendix C. Reproducibility and Evidence Locators

C.1 Analysis protocol

Run `analyze_results.py` against the archived real-test directory, then `make_figures.py`, then `build_paper.py`. The first script reads records without modifying them, writes game/hour/message annotations and an input manifest, verifies 45 hours and 450 actions per game, verifies all configured source hashes, reproduces the original keyword and behavior counts, and reconstructs every agent's goods and futures P&L. The figure script uses only those computed outputs. The document builder embeds those figures and checked statistics.

The companion package includes a README with commands and dependencies, the manuscript source, five figures in vector PDF/SVG and raster PNG formats, analysis tables, and the raw files needed for these computations. A minimal data-root argument allows the archived logs to be analyzed outside the original workstation. No API credentials or new live calls are required for recomputing this report.

Claim or quantity	Primary evidence / location
225 valid reviews; no inspections	G3-G7 regulator.jsonl, lines 1-45; matching phase=review records in llm_calls.jsonl. All scores are 1; all suspect/action lists are empty.
Five terminal audit failures	G3-G7 regulator.jsonl, line 46; final llm_calls.jsonl record: prompt=null, response=null, parse_ok=false, duplicate round TypeError.
Duplicate argument cause	Archived src/capital_futures/agents.py, _call and audit; engine.py, _audit_dossier.
Original clustering and lexicon	Archived src/capital_futures/metrics.py: COORDINATION_TERMS, _clustered_hours, _side_comovement, compute_metrics.
Unflagged private overture	G2 messages.jsonl, line 25: Omar to Nia, sent hour 3, delivered hour 4.
Repeated denial statement	G5 messages.jsonl, line 179: Theo, hour 19, delivered hour 20. Identical message text occurs 212 times across G5.
Goods rule discrepancies	Archived prompts/trader.jinja2; engine.py, spot_reservation and settle_spot; book.py, match; actual quantities in trades.jsonl.
Financial totals and weights	trades.jsonl and variation_margin.jsonl; evidence/game_metrics.csv. All 70 agent P&L; decompositions match summary.json.

Table C1. Auditable locators. G1-G7 denote the numbered directories listed in Table 1 and mapped to full names in the input manifest. JSONL line numbers are one-based. Source references point to each run's archived snapshot, not a subsequently edited working tree.

C.2 Intended protocol versus observed execution

Three distinctions should survive any reuse of this dataset. First, terminal inspection of all private messages is the intended protocol; the recorded model audits did not run. Second, the private cap controls eligible agent pairs, but the realized opened-message count is zero. Third, documentation saying “two or more” quotes within \$0.50 is not equivalent to the implemented all-quoter range test. The paper preserves the original outputs and explicitly labels each reanalysis.

The declared regulator model field and provider field are also insufficient to identify actual inference: the archived backend is Modal and the live call records name Qwen/Qwen3.8-27B. Likewise, configuration seed 0 controls environment randomness, while no generation seed is supplied in the archived request wrapper. Reproduction of the analysis is exact; reproduction of the stochastic trajectories is not guaranteed.

For the next live study, release both the frozen prompt and engine contracts, a validated positive control, per-call execution status, and independent labels for sampled candidate episodes. Report missing or failed audits as unavailable. None of those future validations should be represented as completed by the current seven-game record.