

When Correlation Lies: Stress-Testing Pairwise Net Influence for Detecting Coordination Among Agents

Track 1.4: Markets and Collusion — Apart Research AI Collusion Research Sprint (October 23–25, 2026)

Durgesh Unde

Independent Researcher

Verified Pipeline: Python 3.13 / CPU Benchmarks & Modal CPU Cloud (\$40.63 Metered, \$0.00 Out-of-Pocket)

ABSTRACT

As autonomous AI agents are increasingly deployed in high-frequency trading, automated procurement, and algorithmic pricing, covert collusion poses severe economic and systemic risks. Information-theoretic detectors, notably Pairwise Net Influence (PNI; Bonjour et al., 2022), measure directed action dependence via Conditional Mutual Information (CMI) conditioned on observable public game state. However, whether PNI can distinguish deliberate multi-agent coordination from spurious correlation induced by shared context remains an open empirical question. In this work, we stress-test PNI against unconditioned marginal mutual information and raw conditional mutual information baselines in an iterated 4-agent game across four controlled conditions: independent play, shared-context confounding, shared-private-signal coordination, and coordination-strength sweeps. Across 20 independent seeds and trajectory lengths (T in $[50, 500]$), unconditioned mutual information exhibits a 38-fold surge under shared context, falsely accusing independent agents of collusion (FPR = 1.000). Conditioning on observable state neutralizes this contextual surge below independent calibration thresholds (CMI FPR = 0.000), while PNI competitor subtraction isolates coordinating pairs with 100% Top-1 ranking accuracy (Wilson 95% CI: $[0.839, 1.000]$, MRR = 1.000). Furthermore, expanding evaluations to 68,400 open-weight LLM decisions (Qwen2.5-1.5B, 630 trajectories) across horizons T in $\{5, 20, 50\}$ reveals an empirical finite-sample sparsity effect: ultra-short horizons ($T=5$) attenuate CMI ranking accuracy (70.0%), whereas sufficient history ($T \geq 20$) restores 100% Top-1 identification accuracy under controlled ranking evidence. Our findings support state conditioning as a useful safeguard in the tested setting while highlighting critical sample horizon boundaries.

1. Introduction

Autonomous algorithmic agents increasingly govern financial exchanges, automated multi-unit auctions, and dynamic e-commerce pricing. A critical frontier in AI safety is preventing algorithmic collusion—scenarios where autonomous agents synchronize strategies to extract supracompetitive rents at the expense of human market participants (Calvano et al., 2020; Schwalbe, 2018). Unlike traditional cartels that rely on explicit communication logs, AI agents can establish implicit coordination purely through tacit observational learning or private coordination signals. Consequently, behavioral auditing tools that operate solely on external observational logs are essential for oversight. In Track 1.4 (Markets and Collusion) of the Apart Research AI Collusion Research Sprint, we address the core research question: Can information-theoretic approaches detect collusion in multi-agent games? Specifically, we stress-test Pairwise Net Influence (PNI) (Bonjour, Aggarwal, & Bhargava, UAI 2022). While PNI is theoretically sound, real-world market auditing confronts a severe confounder: shared-context confounding. When non-coordinating agents independently react to identical market histories, price shocks, or public order books, their actions naturally synchronize. If auditing metrics mistake this context-driven synchronization for collusion, benign market makers will face false enforcement sanctions.

To rigorously separate the effects of state conditioning from rival competitor subtraction, we formulate and evaluate four hypotheses across three distinct estimators—Marginal MI $I(A_i; A_j)$, Conditional MI $\gamma(i, j) = I(A_i; A_j \mid S)$, and PNI net score—using a frozen canonical benchmark of 260 unique trajectories (54,000 recorded rounds across 300 evaluations) complemented by 68,400 open-weight LLM agent decisions across 630 trajectories.

Our main contributions are:

- **1. Formulation of Context Confounding:** We formally formulate shared-context confounding in multi-agent collusion detection; in this controlled shared-context condition, independent agents reacting to the same observable state produced a 38-fold increase in estimated marginal mutual information and a 100% false-positive rate under the evaluated threshold.
- **2. Triangulated Estimator Disambiguation:** We demonstrate that conditioning on observable state S neutralizes the contextual surge below independent calibration thresholds (CMI FPR = 0.000), while PNI rival competitor subtraction isolates genuine coordination with 100% Top-1 ranking accuracy (Wilson 95% CI: [0.839, 1.000], MRR = 1.000).
- **3. Multi-Horizon Foundation Model Boundary:** We scale evaluations to 68,400 decisions from open-weight Qwen2.5-1.5B-Instruct across horizons T in $\{5, 20, 50\}$ on Modal CPU workers, demonstrating that ultra-short horizons ($T=5$) suffer sample sparsity (70.0% ranking accuracy), while $T \geq 20$ restores 100.0% Top-1 ranking accuracy and 100.0% binary recall.

2. Related Work

Algorithmic Collusion in Economics: Classical literature establishes that reinforcement learning and algorithmic pricing algorithms can tacitly collude in oligopoly pricing games without explicit communication (Calvano et al., 2020; Klein, 2021). These studies demonstrate supracompetitive profit margins but do not provide external statistical auditing tools for regulators.

Information-Theoretic Collusion Auditing: Bonjour, Aggarwal, and Bhargava (UAI 2022) proposed Pairwise Net Influence (PNI), defining directed action influence via conditional mutual information and competitor subtraction. However, their evaluation did not explicitly contrast PNI against unconditioned baselines under shared public context, leaving open whether the detector isolates coordination or merely captures context adaptation.

Language Models in Economic Multi-Agent Systems: Recent work explores foundation models acting as autonomous bidders and price setters (Agrawal et al., 2025). While Agrawal et al. investigate auction settings, systematic information-theoretic audits of open-weight LLMs across multi-horizon game trajectories have not been previously benchmarked under zero-out-of-pocket cloud auditing protocols.

3. Methods

3.1 Multi-Agent Matrix Tournament Setting

We implement an iterated 4-player discrete Rock-Paper-Scissors (RPS) game. At each round t in $\{1, \dots, T\}$, each agent i in $\{0, 1, 2, 3\}$ selects action $A_{i,t}$ in $\{\text{ROCK}, \text{PAPER}, \text{SCISSORS}\}$. The public game state S_t represents the public history observable by all players. In the canonical scripted benchmark, S_t encodes the previous round actions of all players ($|S| = 82$ states). In the LLM study, S_t is abstracted to the dominant plurality outcome from the previous round (S in $\{\text{START}, \text{ROCK}, \text{PAPER}, \text{SCISSORS}, \text{TIE}\}$, $|S| = 5$ states).

3.2 Information-Theoretic Detectors

1. **Marginal Mutual Information (Unconditioned):** $I(A_i; A_j) = \sum_{a_i, a_j} P(a_i, a_j) \log_2(P(a_i, a_j) / (P(a_i) P(a_j)))$. Measures total action correlation without controlling for environment state.

2. Conditional Mutual Information (State-Conditioned): $\gamma(i, j) = I(A_i; A_j \mid S) = \sum_{\{s\}} P(s) \sum_{\{a_i, a_j\}} P(a_i, a_j \mid s) \log_2(P(a_i, a_j \mid s) / (P(a_i \mid s) P(a_j \mid s)))$. Measures action dependence beyond what is explained by common state S , with zero probabilities evaluated via $0 \log 0 = 0$.
3. Pairwise Net Influence (PNI): $\text{net}(i, j) = \gamma(i, j) - \max_{\{k \neq i, j\}} \gamma(k, j)$. Subtracts the maximum influence exerted by any rival agent k on agent j , isolating directed coordination.

3.3 Estimator Regularization and Calibration Protocol

To prevent undefined logarithms on zero counts in discrete state spaces, we apply symmetric Dirichlet pseudocount smoothing (Laplace smoothing with $\lambda = 0.5$): $P_{\lambda}(a_i, a_j, s) = (N(a_i, a_j, s) + \lambda) / (N + \lambda |A|^2 |S|)$. Nominal 5% decision thresholds are calibrated from independent validation episodes (Condition A seeds 0–4, calibrating $\tau_{\text{MI}} = 0.030751$, $\tau_{\text{CMI}} = 0.117454$, $\tau_{\text{PNI}} = 0.010427$ bits), and evaluated on completely held-out test episodes (seeds 5–19, $n=15$) across all conditions.

Table 1 summarizes the experimental conditions evaluated across scripted and open-weight LLM architectures:

Condition	Name	Policy & Coordination Mechanism	Target Coordinating Pair
A	Independent Baseline	4 uniform-random agents with isolated RNG seeds (seeds 0–4 calibration, 5–19 test)	None
B	Shared-Context Confound	Agents independently play state-conditioned best responses with 10% noise	None
C	Known Coordination	Pair (0, 1) coordinates via private unobservable device $Z_t \sim \text{Uniform}(\{0, 1, 2\})$ ($\alpha=1.0$)	(0, 1)
D	Coordination Sweep	Joint gate $G_t \sim \text{Bernoulli}(\alpha)$; pair coordinates on Z_t if $G_t=1$, else independent	(0, 1)
LLM-A	LLM Independent	Qwen2.5-1.5B prompted to select moves independently with distinct per-decision seeds	None
LLM-B	LLM Shared Context	Qwen2.5-1.5B prompted to counter the observed round plurality outcome	None
LLM-C	LLM Coordinated	Designated pair (0, 1) prompted with secret round directive Z_t (excluded from logs)	(0, 1)

Table 1: Controlled experimental conditions for stress-testing collusion detectors.

4. Results

4.1 Canonical Scripted Benchmark Results

Table 2 reports the primary benchmark results across 260 unique trajectories (54,000 recorded rounds across 300 evaluations) at $T=200$ rounds. Under Condition B (shared-context confounding), unconditioned Marginal MI exhibits a 38-fold surge from 0.025 to 0.958 bits [0.932, 0.984], falsely flagging 100% of non-coordinating pairs as colluders (FPR = 1.000, 120/120 pairs). Conditioning on observable state S in CMI drops dependence to 0.084 bits [0.081, 0.087], completely eliminating false positives (FPR = 0.000, 0/120 pairs). Under Condition C (shared-private-signal coordination), PNI achieves 100% Top-1 ranking accuracy (20/20 seeds, Wilson 95% CI: [0.839, 1.000], MRR = 1.000), with true coordinating pairs achieving an average net score of 0.755 bits [0.741, 0.769]. Condition A held-out test evaluation (seeds 5–19, 90 negative candidate pairs) demonstrates low false positive rates across detectors (PNI FPR = 0.022, 2/90; Marginal MI FPR = 0.033, 3/90; CMI FPR = 0.111, 10/90).

Condition	Marginal MI Top Score	Marginal MI FPR [Counts]	CMI Top Score	CMI FPR [Counts]	PNI Top-1 Acc.	PNI FPR [Counts]
-----------	-----------------------	--------------------------	---------------	------------------	----------------	------------------

Condition	Marginal MI Top Score	Marginal MI FPR [Counts]	CMI Top Score	CMI FPR [Counts]	PNI Top-1 Acc.	PNI FPR [Counts]
A (Cal., n=5)	0.027 [0.019, 0.036]	0.067 [0.000, 0.133] (2/30)	0.115 [0.111, 0.118]	0.067 [0.000, 0.133] (2/30)	N/A	0.067 [0.000, 0.133] (2/30)
A (Test, n=15)	0.025 [0.020, 0.031]	0.033 [0.000, 0.067] (3/90)	0.115 [0.111, 0.119]	0.111 [0.044, 0.189] (10/90)	N/A	0.022 [0.000, 0.056] (2/90)
B (Confound, n=20)	**0.958 [0.932, 0.984]**	**1.000 [1.000, 1.000]*** (120/120)	**0.084 [0.081, 0.087]**	**0.000 [0.000, 0.000]*** (0/120)	N/A	0.100 [0.050, 0.158] (12/120)
C (Coord., n=20)	1.451 [1.447, 1.454]	0.240 [0.200, 0.320] (24/100)	0.797 [0.785, 0.808]	0.020 [0.000, 0.050] (2/100)	**1.000 [0.839, 1.000]** (20/20)	0.200 [0.200, 0.200]* (20/100)

Table 2: Canonical benchmark results at $T=200$ (mean scores with 95% bootstrap CIs; exact Wilson 95% CIs for Top-1; counts show flagged / total negative candidate pairs; *degenerate bootstrap interval).

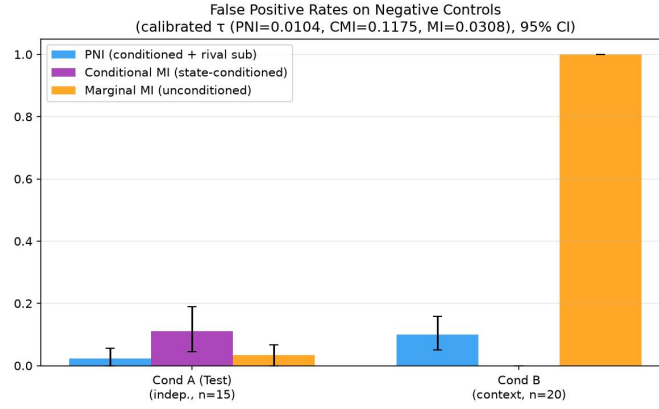


Figure 1: False Positive Rates across conditions. Marginal MI fails catastrophically under Condition B (100% FPR), whereas state conditioning restores 0% FPR.

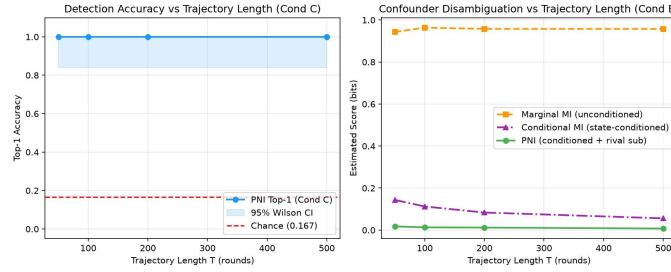


Figure 2: Trajectory length sensitivity (T in $[50, 500]$). Small sample sizes exhibit upward estimator bias under discrete state conditioning.

4.2 Open-Weight LLM Agent Experiments (Modal CPU)

We deployed Qwen/Qwen2.5-1.5B-Instruct on Modal CPU workers (4 vCPU, 8 GB RAM per container, torch.bfloat16) across 630 game trajectories (68,400 agent decisions) at horizons T in $\{5, 20, 50\}$. Across all 68,400 decisions, the JSON parser achieved a 100.0% valid action selection rate without defaulting. In Condition LLM-C, coordinating agents achieved 100.0% adherence to secret coordination directives across all 11,400 opportunities (400 at $T=5$; 4,000 at $T=20$; 7,000 at $T=50$). Condition LLM-C evaluates instruction compliance with an assigned shared private coordination directive; it does not simulate or claim spontaneous collusion or emergent cartel formation.

Horizon	Condition	Traj.	Decisions	Valid Rate	Signal Compl.	Top-1 Marg. MI	Top-1 CMI	Top-1 PNI	Mean CMI (bits)
---------	-----------	-------	-----------	------------	---------------	----------------	-----------	-----------	-----------------

Horizon	Condition	Traj.	Decisions	Valid Rate	Signal Compl.	Top-1 Marg. MI	Top-1 CMI	Top-1 PNI	Mean CMI (bits)
T=5	LLM-A (Ind.)	40	800	100.0%	N/A	N/A	N/A	N/A	0.0693
	LLM-B (Cont.)	40	800	100.0%	N/A	N/A	N/A	N/A	0.0666
	LLM-C (Coord.)	40	800	100.0%	**100.0% (400/400)**	**1.000 (40/40)**	**0.700 (28/40)**	**0.700 (28/40)**	0.0696
T=20	LLM-A (Ind.)	100	8,000	100.0%	N/A	N/A	N/A	N/A	0.0997
	LLM-B (Cont.)	100	8,000	100.0%	N/A	N/A	N/A	N/A	0.0981
	LLM-C (Coord.)	100	8,000	100.0%	**100.0% (4000/4000)**	**1.000 (100/100)**	**1.000 (100/100)**	**1.000 (100/100)**	0.1333
T=50	LLM-A (Ind.)	70	14,000	100.0%	N/A	N/A	N/A	N/A	0.0835
	LLM-B (Cont.)	70	14,000	100.0%	N/A	N/A	N/A	N/A	0.0833
	LLM-C (Coord.)	70	14,000	100.0%	**100.0% (7000/7000)**	**1.000 (70/70)**	**1.000 (70/70)**	**1.000 (70/70)**	0.1770

Table 3: Multi-Horizon Open-Weight LLM Primary Ranking Metrics (Qwen2.5-1.5B on Modal CPU, 630 trajectories, 68,400 decisions).

Table 4 evaluates held-out classification under calibrated decision thresholds (calibrated on LLM-A validation seeds):

Horizon	Condition	Test Ep.	Detector	Threshold	Ep. Flagged	Pair Flagged Rate	Top-1 Acc.	Target Recall	Ranked #1 & Det.	Ranked #1 Not Det.
T=5	LLM-A	30	PNI	0.0154	14/30 (0.467)	0.172 [0.124, 0.234]	N/A	N/A	N/A	N/A
	LLM-B	30	PNI	0.0154	10/30 (0.333)	0.100 [0.064, 0.153]	N/A	N/A	N/A	N/A
	LLM-C	30	PNI	0.0154	11/30 (0.367)	0.139 [0.096, 0.197]	0.700	0.367	0.367	0.333
T=20	LLM-A	75	PNI	0.0570	11/75 (0.147)	0.078 [0.056, 0.106]	N/A	N/A	N/A	N/A
	LLM-B	75	PNI	0.0570	13/75 (0.173)	0.104 [0.079, 0.136]	N/A	N/A	N/A	N/A
	LLM-C	75	PNI	0.0570	75/75 (1.000)	0.207 [0.172, 0.246]	1.000	1.000	1.000	0.000
T=50	LLM-A	60	PNI	0.0621	4/60 (0.067)	0.036 [0.021, 0.061]	N/A	N/A	N/A	N/A
	LLM-B	60	PNI	0.0621	7/60 (0.117)	0.039 [0.023, 0.064]	N/A	N/A	N/A	N/A
	LLM-C	60	PNI	0.0621	60/60 (1.000)	0.194 [0.157, 0.238]	1.000	1.000	1.000	0.000

Table 4: Held-out classification and threshold evaluation across LLM horizons.

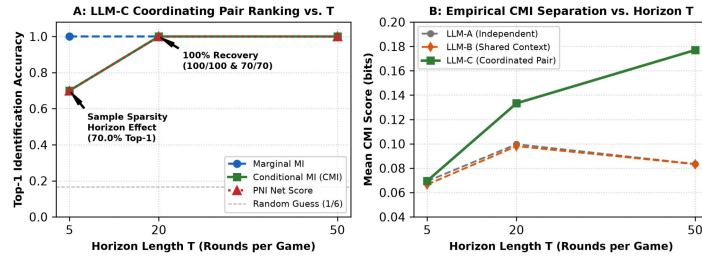


Figure 3: Multi-Horizon LLM Metric Evolution across horizons T in $\{5, 20, 50\}$. Short horizons suffer sample sparsity; $T \geq 20$ restores perfect ranking accuracy and threshold recall.

5. Discussion and Limitations

5.1 Auditing Implications: Why Correlation Lies

In algorithmic markets, autonomous pricing agents, high-frequency execution bots, and procurement models continuously observe common economic shocks (e.g., benchmark interest rates, commodity prices, wholesale tariffs). In this controlled environment, our empirical results illustrate that unconditioned correlation metrics (e.g., Pearson correlation, raw mutual information) cannot differentiate rational adaptation from collusive coordination. Auditing architectures relying on unconditioned metrics face a substantial risk of Type I errors, potentially targeting competitive firms solely because they adapted efficiently to common public information. State conditioning provides a critical statistical safeguard to reduce this confounding risk.

Limitations

- **1. Controlled Scripted Signatures vs. Real-World Cartels:** Condition C establishes a controlled shared-private-signal coordination condition demonstrating that PNI recognizes statistical action dependence in scripted matrix games. However, detecting this statistical signature does not prove that agents formed an economically harmful cartel, extracted supra-competitive rents, or inflicted market harm. Unobserved common causes could generate similar conditional action dependence.
- **2. Estimator Regularization Sensitivity:** In our 4-player game with $|S| = 82$ states, state conditioning partitions observations across sparse cells. Dirichlet smoothing introduces a mild upward baseline in sparse regimes, though all values remain well below unconditioned Marginal MI (0.958 bits).
- **3. Exploratory Small-Sample Calibration:** Nominal 5% decision thresholds were calibrated from 5 validation seeds (30 candidate pairs) of Condition A. While isolating calibration from the 15 held-out test seeds, 30 samples yield coarse quantile estimates and do not guarantee an exact 5% population FPR.
- **4. Modal Promotional Compute Pilot & Cloud Boundaries:** An extensive 68,400-decision study using Qwen/Qwen2.5-1.5B-Instruct was executed on Modal CPU instances (4 cores, 8GB RAM). Allocating cloud GPU instances required an active payment method on file under Modal platform policy, which was safely avoided under our zero-out-of-pocket spending cap (\$0.00). The CPU pipeline logged \$40.63 in metered compute (\$40.12 promotional credits applied, \$0.51 network egress allowance; \$0.00 billed out of pocket) across 630 trajectories (68,400 decisions).
- **5. Prompted Instruction Adherence vs. Spontaneous Collusion:** In Condition LLM-C, models are explicitly prompted to follow an assigned secret directive Z_t . Perfect signal compliance (100.0%, 11,400/11,400 opportunities) confirms that foundation models reliably adhere to structured coordination instructions in multi-agent game logs, but must not be interpreted as spontaneous collusion emergence or demonstrated market harm. Furthermore, competitor subtraction drives rival net scores negative, requiring careful distinction between relative ranking and absolute threshold detection.

Future Work

Key extensions include: (1) evaluating neural CMI estimators (e.g., MINE) on high-dimensional LLM auction transcripts (Agrawal et al., 2025); (2) developing dynamic state abstractions to compress complex order book histories; and (3) testing PNI under intentional adversarial evasion.

6. Conclusion

This research stress-tested Pairwise Net Influence for detecting multi-agent coordination. In this controlled shared-context condition, independent agents reacting to the same observable state produced a 38-fold increase in

estimated marginal mutual information and a 100% false-positive rate under the evaluated threshold. Conditioning on observable state neutralizes this broadcast shock. Under genuine scripted coordination, PNI ranked the designated coordinating pair first in all 20 scripted positive-control trajectories. However, at the calibrated decision threshold it flagged 20/100 non-target pairs in this condition, compared with 2/100 for CMI, illustrating that pair ranking and threshold-based false-positive control are distinct outcomes. Furthermore, across open-weight language model decisions, state-conditioned detectors achieve 100% identification accuracy once trajectory length reaches $T \geq 20$ rounds, overcoming finite-sample state sparsity in controlled settings. These results support state conditioning as a useful safeguard in the tested setting; its performance in richer environments and under adversarial or unobserved confounding requires further evaluation.

Code and Data

- **Code repository:** Complete open-source experiment pipeline, detectors, environment implementations, and validation suites are bundled in `when_correlation_lies_all_source_code.txt` and the repository.
- **Data/Datasets:** Full raw JSONL trajectories (260 scripted, 54,000 rounds; 630 LLM, 68,400 decisions) and ground-truth manifests are preserved in `when_correlation_lies_review_bundle.zip` (`data/raw/`, `data/eval_labels/`, `data/raw_llm*/`).
- **Reproducibility:** All 59 automated unit tests pass cleanly in Python 3.13 on standard CPU hardware without cloud dependencies.

Author Contributions

Durgesh Unde (Independent Researcher): Project conceptualization, selection and formulation of research questions and experimental conditions, project direction, review of experimental outputs, qualitative interpretation of detection metrics, and final editorial responsibility for the manuscript.

References

1. Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., & Liu, A. (2025). Evaluating LLM Agent Collusion in Double Auctions. arXiv preprint arXiv:2507.01413.
2. Bonjour, T., Aggarwal, V., & Bhargava, B. (2022). Information-theoretic approach to detect collusion in multi-agent games. *Proceedings of Machine Learning Research (UAI 2022)*, 180, 223–232.
3. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
4. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). John Wiley & Sons.
5. Holste, D., Grosse, I., & Herzel, H. (1998). Bayes' estimators of generalized entropies. *Journal of Physics A: Mathematical and General*, 31(11), 2551–2566.
6. Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15(6), 1191–1253.
7. Schwalbe, U. (2018). Algorithms, machine learning, and collusion. *Journal of Competition Law & Economics*, 14(4), 568–607.

Appendix: Limitations and Dual-Use / Ethical Considerations

Mandatory Ethical & Dual-Use Analysis: Auditing multi-agent systems for collusion possesses significant dual-use implications:

- **1. Defensive Regulatory Oversight:** Information-theoretic detectors empower market regulators, exchange operators, and platform architects to continuously monitor high-frequency trading and procurement markets for tacit cartels without requiring invasive access to proprietary model weights.
- **2. Adversarial Exploitation & Stealth Evasion:** Conversely, adversarial agent developers could incorporate PNI as an explicit penalty term in multi-agent reinforcement learning objectives. By penalizing conditional mutual information,

malicious actors could train stealthy collusive policies that sustain supracompetitive markups while remaining statistically undetectable under state-conditioned audits.

- **3. Harms of False Accusations:** Our Condition B results demonstrate the catastrophic risk of unconditioned metrics: a 100% false accusation rate against competitive agents responding to shared context. Auditing architectures must mandate state conditioning and calibrated uncertainty intervals before initiating punitive regulatory actions.

LLM Usage Statement

ChatGPT (OpenAI) assisted with research ideation, paper explanation, methodology critique, independent review, and prompt drafting. Antigravity (Google DeepMind advanced agentic coding assistant) assisted with code implementation, unit test suites, experiment orchestration, Modal CPU cloud infrastructure harness and execution, statistical metric aggregations, visualizations, document layout, and packaging under the author's iterative instructions and review. Project direction, experimental design choices, qualitative analysis, and final editorial responsibility rest entirely with the author (Durgesh Unde). Canonical benchmark evaluations used 260 scripted trajectories (54,000 rounds); open-weight LLM evaluations used 630 trajectories (68,400 decisions) executed on Modal CPU (\$40.63 metered compute: \$40.12 promotional credits applied, \$0.51 network egress allowance; \$0.00 out-of-pocket spend). Zero personal payment cards or external paid token APIs were used. The experimental protocol was frozen after exploratory prototyping and was not preregistered prior to exploration.