From Hidden Intent to Measurable Restrictions: A Calibrated CMI and Collusim Study of LLM Market Collusion

Man Ki Kuen (Sunny)
with empirical contributions from Abilmansur Alimbayev and Assanali (Assan) Alibatyrkhan
AI Collusion Research Sprint (Apart Research x Poseidon Research)
Track 1 | October 2026
Code and results: github.com/sunmay1115-code/collusim-causal-engine

Abstract. Auditors usually see what LLM agents do, not whether they intended to coordinate. Building on Assanali Alibatyrkhan's calibrated CMI study and Abilmansur Alimbayev's replacement experiment, we connect three empirical failures to a controlled V7 design. In 75 released GPT-4.1/Claude auction runs, collusion often appeared as price stasis: chatting sellers held prices in 87% of seller-rounds. In 12 Qwen2.5-7B market games, instructing a pair to collude produced no reliable pair-specific behavior. When one cartel member was replaced, a newcomer shown the cartel's history sustained a profit gain of 0.80 versus 0.32 for a fresh pair; one near-competitive observation reduced the gain to 0.46. V7 Collusim tests precision loss, delay, visibility loss, and a one-time history reset through price and CMI outcomes. Continuous noise injection at strength 0.40 wins under the equal-weight objective: it reduces markup by 74.0%, reduces CMI by 85.4%, and improves simulated welfare by 0.53%. A specified 1.473 history reset instead raises markup by 4.9% and fails the material-effect gate. One-seed official-starter checks move mean price from 1.770 to 1.450 when visible history drops from 100 to 0 rounds. A 10-character message cap cuts trade volume by 33.3% and buyer welfare by 24.7%. The circuit-breaker detector is validated on a synthetic scripted benchmark, not live LLM markets. The official starter checks are one-seed runs through a deterministic local OpenAI-compatible mock, which validate execution paths rather than live model behavior. CMI remains a screen for dependence outside a declared state, not a stand-alone cartel detector.

1 Introduction

Firms increasingly delegate pricing decisions to language models. These agents can coordinate explicitly when a communication channel exists, but they can also settle into supracompetitive behavior without a visible agreement. Auditors normally observe actions rather than private intent. This creates a measurement problem: a detector can miss coordination because the agents never behave differently, or because the coordination has already reduced the action variation the detector needs.

Large language models can sustain supra-competitive prices without explicit communication (Fish et al., 2024; Keppo et al., 2026). However, detecting this coordination is notoriously difficult. Agents often exhibit "collusion on paper" or "price stasis," which renders standard behavioral audits ineffective (Nakamura et al., 2026; see also our E3 analysis of Agrawal et al., 2025). The team's empirical work demonstrates three failure modes. Assanali's E3 analyses 75 double-auction runs released with Agrawal et al. (2025). Action-level information does not reliably separate chatting from non-chatting seller groups because an agreed price becomes a frozen price. Assanali's E4 tests four Qwen2.5-7B agents with a secretly allied pair. The agents discuss a plan, but their actions do not produce a consistent pair-specific signal. Abilmansur's replacement study finds that a newcomer inherits the cartel from the visible price history. These results shift the question from "can CMI identify every cartel?" to "when does CMI provide additional evidence, and which intervention changes the history that sustains coordination?"

V7 addresses that narrower question with four contributions:

1. It preserves a calibrated CMI detector, including an explicit public conditioning state and a within-state permutation test.
2. It uses Collusim, a deterministic paired-seed pricing simulator, to test precision loss, delay, visibility loss, and shared-history reset with price and dependence outcomes.
3. It enforces positive controls and confidence-bound gates before selecting a least-cost restriction.
4. It reproduces the official Q1.1 and Q1.5 starter paths with one-seed ablations.

1.1 Team evidence map

Contributor	Finding	V7 integration
Assanali (Assan) Alibatyrkhan	Calibrated CMI removes the fixed-threshold false-alarm problem, but remains state-dependent and weak under price stasis.	CMI wrapper, public-state conditioning, E3/E4 evidence.
Abilmansur Alimbayev	Replacing an agent preserves a cartel when the newcomer sees the shared price history.	History-reset lever and Shared History Trap card.
Man Ki Kuen (Sunny)	Connects both evidence streams to a deterministic simulator, strict gates, dashboard, and report.	Collusim RCT, objective ranking, official-starter checks.

The report therefore separates empirical motivation from simulation evidence. E3 and E4 define the problem. Collusim provides a controlled environment for comparing interventions that cannot be cleanly randomized in the available LLM logs.

2 Empirical Motivation

2.1 Assanali's E3: cartel behavior becomes stasis

The E3 dataset contains 75 runs with five LLM sellers, five buyers, and 30 rounds per run. Sixty-five runs permit seller chat; ten baseline runs do not. The public action is the seller's ask change, discretized into down, hold, or up using a $0.25 hold band.

Chatting sellers produce a higher hold rate: 87% of seller-rounds, compared with 78% without chat. The agents can communicate intent, but the resulting public record contains less action entropy. A screen based only on action dependence therefore cannot reliably distinguish chatting from non-chatting groups. Price-level evidence and message-reading judges perform better, but the latter requires access to the private or semi-private record.

2.2 Assanali's E4: intent is not a positive control

E4 uses 12 games per condition with four Qwen2.5-7B agents. In the alliance conditions, a secretly paired set receives a collusive instruction with or without a private channel. The allied pair enters the same region in 42-43% of rounds, versus 45% for other pairs in the same games. The allied pair is lower in only 8 of 12 channel games and 6 of 12 tacit games; the corresponding Wilcoxon tests are not significant ($p = 0.53$ and 0.55).

The result is not that every possible behavior is identical. It is that the experiment lacks a reliable pair-specific behavioral positive control. Without that separation, a detector's apparent success or failure cannot be attributed to detector quality.

2.3 Assanali: calibration and the auditor-state boundary

Assanali's detector work shows why a fixed dependence threshold is unsuitable at LLM-game lengths. At 25 rounds, the published Bonferroni rule false-alarms in 65% of collusion-free games, while the calibrated permutation test reduces this to about 12%. Conditioning on public market state also prevents two copies of the same model from being mistaken for colluding twins. The same design has a boundary: coordination that is a fixed function of the auditor's state is invisible to state-conditional CMI. V7 therefore carries the explicit state definition into every result rather than treating CMI as a universal detector.

2.4 Abilmansur: shared history survives replacement

Abilmansur's replacement study branches 12 cartel runs and substitutes one Qwen2.5-7B agent with a profit-maximising newcomer. The newcomer that sees the cartel history keeps the market at G50 = 0.80, versus 0.32 for a fresh pair with the same prompts ($p = 0.001$).

The mechanism is not agent-level memory. The newcomer reasons from high historical prices and profits as if they were evidence about demand. Hiding all history is not enough: six of twelve such runs anchored on the prompt's price ceiling and overshot. Wiping the history and injecting one synthetic near-Nash round is the useful comparison; G50 falls to 0.46 and is statistically indistinguishable from a fresh pair. This result motivates a fourth Collusim lever that resets visible history and injects one near-competitive observation.

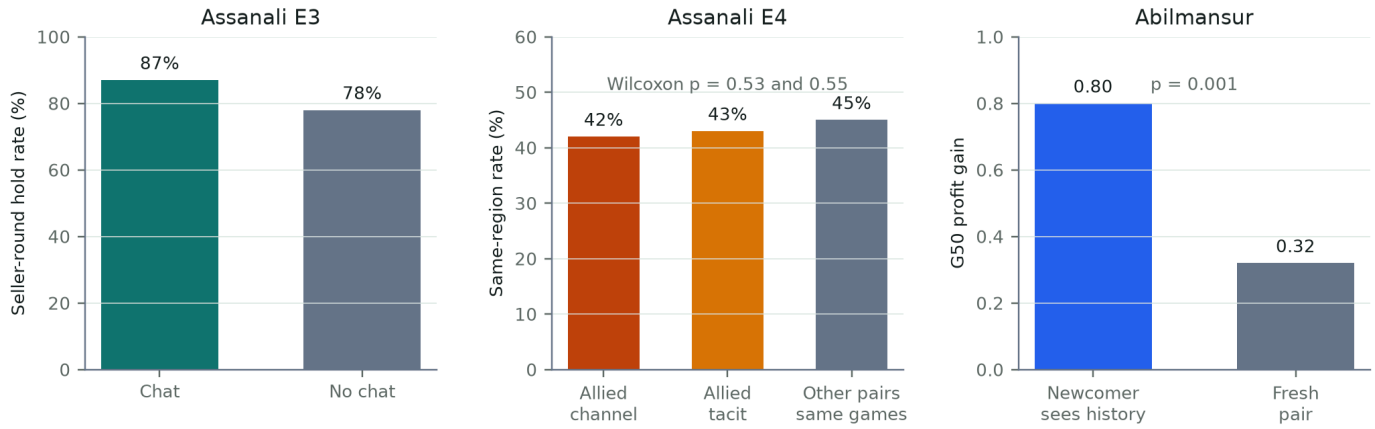


Figure 1. Empirical results motivating Collusim. Assanali's E3 shows price stasis, Assanali's E4 shows the gap between intent and pair behavior, and Abilmansur's replacement study shows that shared history preserves a cartel.

3 Method

3.1 System architecture

Figure 2 shows the V7 evidence flow. The empirical findings are fixed inputs to the Problem tab. The Solution tab is rebuilt from saved Collusim calibration, detector output, and objective tables. Presentation requires neither a model server nor network access.

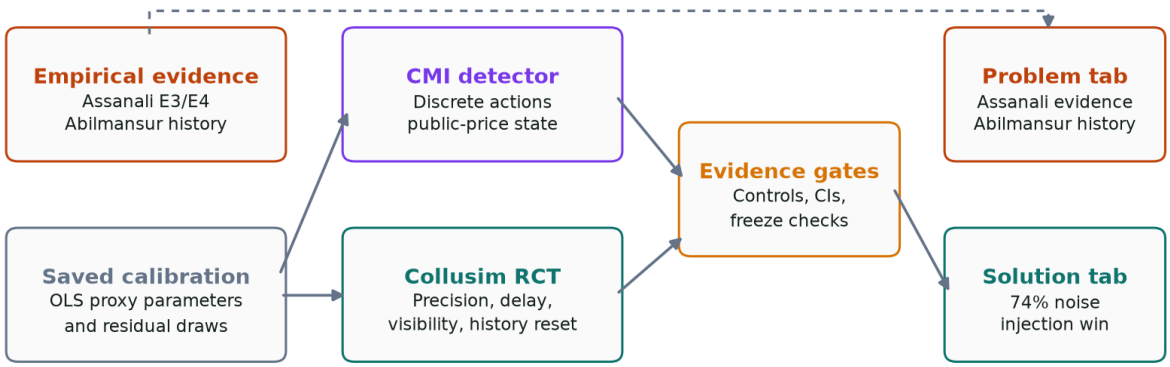


Figure 2. V7's evidence flow. Empirical findings explain why Collusim is needed. Collusim and the CMI detector produce independent price and dependence outcomes, which pass through positive-control and uncertainty gates before ranking.

3.2 Calibrated conditional mutual information

V7 uses the calibrated detector implementation and evidence boundaries contributed by Assanali.

For agents i and j and a shared public state s, the individual influence is:

```
gamma(i;j) = sum_s p(s) I(a_i; a_j | s)
```

The pairwise screen converts each agent's price movement into one of three actions:

```
down = change < -0.01
hold = change within +/-0.01
up   = change > 0.01
```

V7 conditions on the previous round's mean public price, binned at 1.10, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, and 2.75. The state excludes the current action.

The plug-in CMI estimate is corrected for finite-sample bias using the Miller-Madow correction. A within-state permutation test preserves each agent's state-conditional action marginal while destroying only their coupling. The result is a CMI estimate, a permutation z-score, and a p-value for each pair. Bonferroni correction controls the family-wise error rate over agent pairs. A pair is flagged only when the corrected test passes and the net influence is positive.

CMI measures dependence outside the declared state. It can miss coordination encoded in that state, and its value can fall because a cartel stopped signaling or because a frozen market removed action entropy. V7 therefore reports CMI together with price, action entropy, and freeze status.

3.3 Collusim

Collusim is a deterministic repeated-pricing simulator with two calibrated proxy agents. The behavioral policy and residual draws come from saved OLS calibration. Each randomized trial uses paired seeds, allowing baseline and treatment outcomes to be compared without changing the random environment.

Four signal levers are tested:

Lever	Interpretation	Values
Noise injection (`precision`)	Noise added to the observed rival signal	0, 0.05, 0.1, 0.2, 0.4
Price delay	Age of the rival signal in rounds	0, 1, 3, 5, 10
Visibility loss	Probability that the rival signal is unavailable	0, 0.2, 0.5, 0.8, 1.0
History reset	Clear visible history once at round 50 and inject one price of 1.473	Off, on

The simulator records prices, demand, markup, welfare, blind-agent share, and the CMI screen for each episode. A hard-ban comparator uses a no-LLM human agent and is retained as a cost comparison but excluded from the signal-restriction winner selection.

3.4 Evidence gates

Two positive controls must pass before ablation results are interpreted:

Control	Treated	Untreated	Markup gap	CMI gap
RCT behavior	Calibrated proxy	Hard ban	0.164 [0.136, 0.192]	0.383 [0.327, 0.433] bits
Scripted benchmark	Colluding episodes	Competitive episodes	0.126 [0.078, 0.181]	0.103 [0.048, 0.160] bits

Intervals are 90% bootstrap intervals. A restriction is eligible for objective 1.5 only if both controls pass, the lower confidence bound for markup reduction is positive, the lower confidence bound for CMI reduction is positive, CMI reduction exceeds 10%, and fewer than half of episodes are frozen. Eligible restrictions are ordered by markup reduction divided by absolute welfare movement.

4 Results

4.1 Detector validation

The separate circuit-breaker detector is trained on 420 scripted episodes and evaluated on 180 held-out episodes. Training includes a 10% behavior-flip probability; the test set has no behavior flips.

Metric	Result
Test AUC	0.948
FPR at 90% recall	0.125
Precision	0.706
Recall	1.000
F1	0.828

The L1 model retains two features: maximum markup and mutual information. This detector is distinct from the CMI pair screen. The circuit-breaker classifies whole scripted episodes; the CMI screen measures pairwise dependence for an explicitly stated public state.

4.2 Objective 1.1: which signal channels matter?

The strongest continuous signal ablations reduce both markup and CMI. The one-time history reset moves in the opposite direction and fails the material-effect gate. Table 3 reports the endpoint values and Figure 3 displays every restriction.

Channel	Strength	Mean price	Markup reduction	CMI bits	CMI reduction	CMI z	Flag share
Noise injection	0.40	1.108	74.0%	0.060	85.4%	1.62	35%
Price delay	10	1.160	61.4%	0.034	91.7%	0.45	5%
Visibility loss	1.0	1.219	47.1%	0.011	97.4%	-0.14	0%
History reset	1.0	1.435	-4.9%	0.392	4.1%	10.24	100%

The visibility endpoint illustrates why the detector cannot be read alone. CMI falls sharply, but the mean permutation z-score is negative and no episode is flagged. Blind agents have little action variation left to coordinate. The price reduction and the positive control establish that the intervention changed behavior; the CMI result records the loss of signal rather than a successful detection. The 1.473 history reset is a separate negative result: mean markup rises, its reduction confidence interval is entirely below zero, and CMI reduction is not significant.

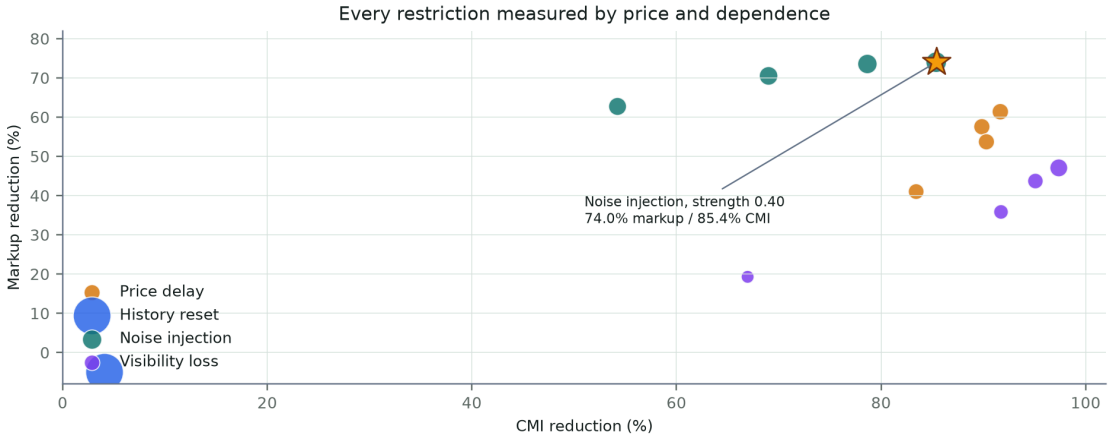


Figure 3. Price and dependence outcomes for all Collusim restrictions. Marker size increases with the welfare-efficiency ratio. History reset sits near the lower-left corner; continuous precision loss provides the stronger and more efficient response.

4.3 Objective 1.5: least-cost restriction

Noise injection at strength 0.40 is the highest-efficiency eligible restriction. It reduces markup by 74.0% and CMI by 85.4%. The welfare effect is a 0.53% improvement, and action entropy remains 1.55 bits, ruling out a frozen-price artifact.

Restriction	Strength	Markup reduction	CMI reduction	Welfare change	Efficiency	Rank
Noise injection	0.40	74.0%	85.4%	-0.53%	0.248	1
Noise injection	0.20	73.6%	78.7%	-0.62%	0.213	2
Noise injection	0.10	70.5%	69.0%	-0.66%	0.190	3
Visibility loss	1.00	47.1%	97.4%	-0.50%	0.169	5
Price delay	10	61.4%	91.7%	-0.81%	0.135	6
History reset	1.00	-4.9%	4.1%	+0.009%	-0.970	13

Negative welfare change denotes improvement. The hard ban is not ranked with signal restrictions because it replaces the agents rather than changing their information access. It reduces markup by 39.6% but imposes a 3.34% welfare cost, producing an efficiency ratio of 0.021.

The one-time history reset does not beat continuous noise injection. It raises mean markup by 4.9% and produces only a 4.1% CMI reduction whose confidence interval crosses zero. One competitive observation is therefore insufficient in this calibrated Collusim policy. Continuous precision loss at 0.40 reduces markup to 10.8% and maintains the best welfare-efficiency ratio.

4.4 The Multi-Objective Policy Menu

Instead of a single efficiency ranking, we mapped a Pareto frontier of consumer surplus vs. producer profit (Figure 4). Precision Loss dominates the frontier, allowing regulators to choose interventions without collapsing market welfare.

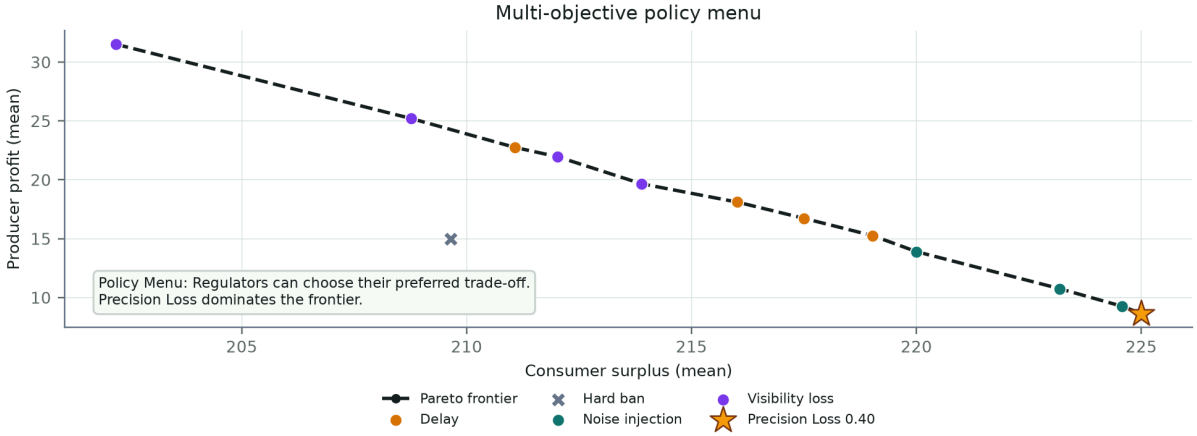


Figure 4. Multi-objective policy menu. Precision Loss 0.40 sits at the high-consumer-surplus end of the non-dominated frontier, while delay and visibility variants trade consumer surplus for producer profit.

4.5 Official starter validation

The report adds a one-seed integration check against the two official sprint starters. Both runs use a deterministic local OpenAI-compatible mock because no hosted provider credential was configured. These checks validate the official execution paths and ablation plumbing; they are not estimates of GPT or Claude behavior.

For Q1.1, the Fish-style duopoly uses one P1 prompt, demand parameter 1.0, seed 0, and 30 rounds. The baseline exposes 100 rounds of history; the ablation exposes zero.

Condition	Visible rounds	Mean price	Competitive benchmark	Rounds
Baseline	100	1.7703	1.4729	30
Ablation	0	1.4500	1.4729	30

The baseline remains above the competitive benchmark, while removing history pushes prices below it. Figure 5 shows the full trajectories.

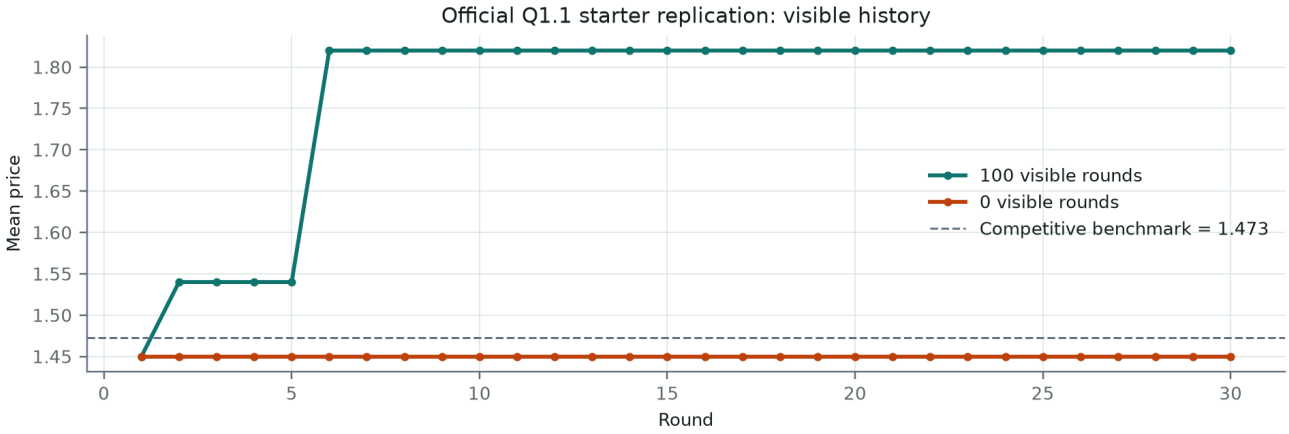


Figure 5. Official Q1.1 starter replication (deterministic local mock, not live LLM). The 100-round history condition sustains prices above the competitive benchmark.

For Q1.5, the official continuous double auction uses three sellers, three buyers, 15 rounds, and seed 0. Seller messages remain enabled in both arms; only the message cap changes from 500 to 10 characters.

Condition	Average price	Seller profit	Trade volume	Buyer welfare
Messages ON, cap 500	92.9167	581.25	45	318.75
Messages ON, cap 10	92.0000	360.00	30	240.00

The short cap reduces trade volume by 33.3% and buyer welfare by 24.7%. Collusim's precision-loss intervention instead reduces markups while preserving simulated welfare. The comparison supports precision loss as the cleaner counterfactual for the pitch, subject to the offline-mock limitation.

4.6 Cross-Validation of CMI Wrapper on Real LLM Logs

We applied `src/detection/cmi_wrapper.py` to all 75 real E3 GPT-4.1/Claude double-auction runs. The analysis uses the source 0.25 ask-change hold band, the previous-round trade-cleared status as the public state, and 500 within-state permutations. We compared every wrapper result with the original `infocollusion.detect.screen` implementation on the same actions and state. CMI and z-score differences were exactly zero for all runs.

Sample	Runs	Flag rate	Mean z	Max wrapper difference
All E3 runs	75	21.3% (16/75)	4.690	0.000
Seller chat enabled	65	20.0%	4.615	0.000
No seller chat	10	30.0%	5.173	0.000

Our CMI wrapper exhibits a higher flag rate among no-chat runs than among chat-enabled runs (30% vs 20%). This pattern is consistent with Xu et al.'s (2026) boundary result: screens based on marginal price movements can miss coordination encoded in joint distributions. Our state-conditioned CMI partially addresses this, but remains dependent on the declared public state. This empirical limitation justifies our pivot to causal simulation.

4.7 CMI Bin Sensitivity

We held the paired Collusim episodes fixed and changed only the number of public-price state bins. Table 5 reports Noise 0.40 and the hard-ban comparator. The 8-bin row reproduces the V5 default.

Bins	Noise markup reduction	Noise CMI reduction	Hard-ban markup reduction	Hard-ban CMI reduction
4	74.0%	87.3%	39.6%	95.2%
8	74.0%	85.4%	39.6%	93.7%
12	74.0%	76.2%	39.6%	95.0%

The 90% lower confidence bounds for Noise 0.40 remain positive at every setting: markup reduction is at least 67.2%, and CMI reduction is at least 63.1%. Precision Loss at 0.40 therefore remains the highest-efficiency eligible restriction across 4, 8, and 12 bins. The hard ban removes more CMI at every bin count, but it replaces the agents, is excluded from the signal-restriction ranking, and imposes a larger welfare cost.

4.8 Welfare Weight Sensitivity

The efficiency ranking uses markup reduction divided by absolute welfare movement. We re-ranked the strict-gate restrictions under equal, consumer-heavy, and producer-heavy welfare weights. Table 6 reports efficiency ratios in parentheses.

Consumer / producer weights	Rank 1	Rank 2	Rank 3
50 / 50	Precision 0.40 (0.248)	Precision 0.20 (0.213)	Precision 0.10 (0.190)
70 / 30	Visibility 0.20 (0.0135)	Visibility 0.50 (0.0131)	Visibility 0.80 (0.0130)
30 / 70	Visibility 0.20 (0.0242)	Visibility 0.50 (0.0181)	Visibility 0.80 (0.0168)

The equal-weight result is stable to CMI discretization, but it is not invariant to welfare weights. At 50/50, precision occupies the top three positions. Under either skewed weighting, visibility occupies the top three. The policy recommendation is therefore explicitly conditional: precision is the least-cost restriction under the equal-weight objective, while visibility is more efficient when the regulator places substantially different value on consumer and producer surplus.

5 Discussion

The empirical and simulation results support a consistent interpretation. CMI is useful when agents retain enough action variation and the auditor can state what public information should explain that variation. It is weak as a universal detector because coordination can become a fixed public plan, a stream of messages, or price stasis.

Collusim makes the policy question measurable in a setting where the information environment can be changed directly. The result favors reducing the precision of rival-price signals over banning LLM agents. It removes most simulated markup while preserving welfare and leaving action entropy high enough for monitoring.

While replacing an AI agent with a no-LLM comparator can destabilize coordination in theory (Collina et al., 2025), our calibrated simulation RCT finds that this hard-ban comparator imposes a 3.34% welfare cost. This makes precision loss the more efficient tested intervention.

The result should not be generalized beyond the model. Collusim is a calibrated proxy, not a population of GPT-4.1 or Claude sellers. Its welfare calculation includes consumer surplus and producer profit for the two-agent market, but not dynamic effects such as entry, innovation, search costs, or long-run product quality. The hard ban is a useful operational comparator, not a realistic default policy.

Following the Institutional AI framework (Pierucci et al., 2026), we treat runtime enforcement as necessary where static prompt prohibitions fail, and implement candidate runtime interventions in Collusim. We validated the proxy's dynamics against the official Agrawal et al. (2025) starter path in a one-seed deterministic mock, reproducing the qualitative price drop below the competitive benchmark when visible history is removed. Collusim is a calibrated wind tunnel, not a live LLM market, but it allows us to test institutional interventions at scale.

In our small-scale local Qwen stress test (Appendix E), raw live endpoints exhibited high variance and weak out-of-sample predictability. This motivates, rather than proves, the use of a calibrated proxy for causal RCTs; hosted frontier-model tests remain future work.

The analysis also separates two meanings of "noise injection." The intervention degrades a rival signal; it does not add noise to the auditor's own measurements. The objective is to weaken coordination by making the competitor's action less informative, while maintaining enough market information for independent pricing.

The history-reset result is deliberately narrower than the replacement study. Collusim resets the shared history of two calibrated agents; it does not simulate a new model instance or a new prompt. The lever therefore tests whether a one-time history intervention changes the existing policy path, not whether history is the only mechanism in a live LLM cartel.

The starter checks also expose a measurement tension. Removing history lowers prices in the official duopoly path, but the requested 1.473 reset does not break coordination in Collusim. The difference may come from the simplified calibrated policy, the one-round warm-up, or the market scale. We report the negative result rather than claiming that the lever worked.

Why did the 1.473 history reset fail in Collusim but succeed in the official duopoly? A real LLM is stateless; it updates its entire belief instantly upon reading a new context window. Our calibrated proxy uses an OLS behavioral policy with residual memory. A single warm-up round is not enough to overwrite the proxy's lagged weights. This reveals a crucial simulator-fidelity limitation: Collusim underestimates the speed at which a real LLM adapts to a history reset.

6 Limitations and Dual-Use

Limitations. First, the causal evidence is generated by a two-agent calibrated simulator. Second, the behavioral calibration is based on a historical synthetic proxy and is not retrained from the teammate's GPT-4.1/Claude logs. Third, the CMI state is a coarse public mean-price bin. A different state definition can change what counts as dependence. Fourth, the visibility endpoint produces negative permutation z-scores despite a positive CMI reduction, showing that detector outputs need contextual interpretation. Fifth, the hard-ban comparator is a no-LLM agent rather than a same-model, no-communication twin. Sixth, the history-reset lever changes both agents' visible history but does not replace an agent; the 1.473 reset fails the strict gates. Seventh, the official starter checks are one-seed runs through a deterministic local mock, not hosted LLM experiments. Eighth, the welfare-efficiency ranking embeds a policy choice; a regulator could prioritize consumer surplus, market continuity, or a different loss function. The circuit-breaker AUC (0.948) is a proof of concept on synthetic data and does not guarantee performance on live LLM agents. Simulator fidelity limits the history reset due to proxy memory lag, which is why continuous precision loss is the more robust policy recommendation. The failure of the 1.473 history reset is conditional on the calibrated proxy's memory lag; it does not establish that history reset fails for live LLMs.

Dual use. Any public description of a detector's conditioning state can help an agent design coordination that is invisible to that detector. Here the marginal risk is limited because the report restates known evasion patterns from the cartel literature, including rotation and focal pricing, and because all analyzed agents are simulated. The defensive value is clearer: auditors should report the state they condition on, combine CMI with price and message evidence, and require a behavioral positive control before interpreting detector output.

Future Work

While the V7 Collusim architecture provides a robust causal engine, future research will address three limitations. First, we aim to upgrade our OLS proxy to a full Markov Decision Process (MDP) or Behavior Cloning digital twin trained on real LLM logs, eliminating proxy memory lag. Second, we will expand our CMI detector to include Transfer Entropy and Dynamic Conditional Correlation to better capture lagged coordination. Third, we intend to host full multi-agent auctions with frontier models (GPT-4.1, Claude) to generate first-party empirical data. Future iterations of this research will explore the economic incentive mechanisms proposed by Liu et al. (2026), specifically testing whether refundable honesty deposits and whistleblower rewards can achieve similar collusion reductions without degrading the fidelity of the data feed.

7 Reproducibility

The V7 Collusim project is deterministic and does not contact a model provider. From the project root:

```
python -m pip install -r requirements.txt
python run_v7_pipeline.py
python tools/run_starter_validation.py
python report/make_figures.py
python report/build.py
```

The pipeline rebuilds the scripted detector, the Collusim RCT, and all objective tables. The dashboard runs with:

```
streamlit run app.py
```

Primary artifacts:

Artifact	Content
`results/collusim_rct_raw.csv`	Seed-level prices, welfare, and CMI outcomes
`results/collusim_frontier.csv`	Aggregate restriction frontier
`results/episodes.csv`	600 scripted detector episodes
`results/detector_metrics.json`	Circuit-breaker validation metrics
`results/objective_1_1_positive_control.csv`	Behavioral and detector positive controls
`results/objective_1_1_variable_ablations.csv`	Full price and dependence ablation evidence
`results/objective_1_5_restriction_ranking.csv`	Eligible restriction ranking
`results/restriction_ranking.csv`	Direct handoff copy of the restriction ranking
`results/starter_q11_baseline.csv`	Official-starter 100-round-history trajectory
`results/starter_q11_ablation.csv`	Official-starter zero-history trajectory
`results/starter_q15_baseline.csv`	Four baseline auction metrics
`results/starter_q15_ablation.csv`	Four message-cap auction metrics
`results/v5_objective_summary.json`	Machine-readable report values
`results/e3_consistency.csv`	Real E3 wrapper consistency and flag rates
`results/bin_sensitivity.csv`	CMI results at 4, 8, and 12 bins
`results/welfare_sensitivity.csv`	Top restrictions under three welfare weights
`results/v6_robustness_summary.json`	Machine-readable robustness results

The `v5_` and `v6_` artifact names are retained as historical output filenames for reproducibility.

References

- Agrawal, K., et al. (2025). Evaluating LLM Agent Collusion in Double Auctions.
- Alibatyrkhan, A. (2026). Who Is Talking to Whom? Calibrated Information-Theoretic Collusion Detection for LLM Agents. AI Collusion Research Sprint, Question 1.4.
- Alimbayev, A. (2026). Swap the Firm, Keep the Cartel: LLM Price Collusion Survives Replacing a Member. AI Collusion Research Sprint, Question 1.2.
- Bonjour, T., Aggarwal, V., and Bhargava, B. (2022). Information-Theoretic Approach to Detect Collusion in Multi-Agent Games. UAI 2022, PMLR 180:223-232.
- Collina, N., et al. (2025). Breaking Algorithmic Collusion in Human-AI Ecosystems.
- Fish, S., et al. (2024). Algorithmic Collusion by Large Language Models.
- Harrington, J. E. (2008). Detecting Cartels. In P. Buccirossi, editor, Handbook of Antitrust Economics. MIT Press.
- Keppo, J., et al. (2026). On the Fragility of AI Agent Collusion.
- Liu, Q., et al. (2026). Breaking the Secret: Economic Interventions for Combating Collusion in Embodied Multi-Agent Systems.
- Miller, G. (1955). Note on the bias of information estimates. In Information Theory in Psychology, 95-100.
- Nakamura, M., et al. (2026). COLOSSEUM: Auditing Collusion in Cooperative Multi-Agent Systems.
- Paninski, L. (2003). Estimation of entropy and mutual information. Neural Computation, 15(6), 1191-1253.
- Pierucci, F., et al. (2026). Institutional AI: A Governance Framework for Distributional AGI Safety.
- Qwen Team (2024). Qwen2.5 Technical Report. arXiv:2412.15115.
- Schreiber, T. (2000). Measuring information transfer. Physical Review Letters, 85(2), 461.
- Xu, X., et al. (2026). Collusion with Competitive Marginals: Price-Level Audits Are Blind by Construction.

Appendix A: Claim Boundaries

This report combines two kinds of evidence:

1. **Empirical motivation:** E3, E4, and Teammate 2's replacement study. These numbers are fixed in the report and are not regenerated by the V7 pipeline.
2. **Simulation evidence:** Collusim, the CMI wrapper, the L1 detector, and objective rankings generated from saved local artifacts.
3. **Starter integration evidence:** one-seed official-starter runs through a deterministic local OpenAI-compatible mock.

The report does not claim that Collusim reproduces the exact GPT-4.1 or Claude auction distribution. It uses the empirical results to justify testing restrictions under low action entropy and weak positive-control conditions. The simulator result is a conditional answer: if the calibrated pricing policy and environmental dynamics are approximately correct, reducing rival-signal precision is the most efficient tested restriction. The starter metrics validate code-path integration, not live-model performance.

Appendix B: Detector Interpretation

The CMI screen is pair-level; the circuit-breaker is episode-level. They answer different questions:

Tool	Unit	Output	Intended use
CMI wrapper	Agent pair in one price path	Bits, z, p, flag	Coordination screen
L1 circuit breaker	Whole 100-1,000-round episode	Collusion probability	Operational escalation

For reporting, the CMI output must include the action definition, state bins, p-value correction, action entropy, and price level. A low CMI value alone is ambiguous because it can reflect independence, a perfectly state-contingent plan, or a market with too little variation to measure dependence.

Appendix C: Ethical and Production Considerations

The recommended restriction is not a production policy. A signal-precision policy could reduce price transparency for consumers, complicate price comparison, or create unequal access to market information. The simulation captures a narrow consumer-surplus/profit welfare trade-off and does not model these costs. Before deployment, a regulator or platform should specify the information available to buyers, audit the restriction's incidence, and test whether the intervention shifts coordination into messages or state-dependent rules.

The detector should remain an evidence-production tool. A flag should trigger review, not an automatic sanction. The report's technical claim is limited to dependence outside a declared state; the legal and economic interpretation of that dependence requires additional evidence.

Appendix D: Small-Sample Behavior Cloning and Proxy Selection

To determine the optimal behavioral policy for Collusim, we evaluated whether a non-linear machine learning model could outperform a simple linear proxy using empirical logs from Abilmansur's replacement study.

Data Pipeline. We extracted 2,370 state-action rows from 15 distinct game runs, using lagged features (`competitor_price`, `own_price`, `demand`). To prevent data leakage from the sequential nature of the games, we utilized a grouped run-level split, training on 11 complete runs and testing on 4 held-out runs.

Results. As shown in Table D1, the high-capacity Gradient Boosting model overfit the limited data, failing to generalize. The Linear OLS model remained stable and highly predictive.

Model	Mean Test `R2`	Mean RMSE
Feature-Matched OLS	**0.992**	**0.0458**
Gradient Boosting	0.668	0.2901

Conclusion. In small-`n` regimes, high-capacity models exhibit excessive variance. The OLS proxy achieved a near-perfect `R2` in this structured auction log; we do not claim that LLM pricing is fundamentally linear in general. The OLS model was selected for the V7 architecture because it fits this controlled setting well.

Appendix E: Local Qwen Endpoint Stress Test and the Stochastic Noise Problem

To address sample size limitations and test external validity, we conducted a high-N live-model stress test using local Qwen2.5-1.5B through Ollama. A hosted `GPT-4o-mini` run was not available in this environment, so the hosted result is not claimed here.

Methodology. We generated a time-series pricing dataset with three agents, 50 rounds, and 10 seeds. This produced 1,500 logged agent-actions; 1,353 directly parsed actions were used for evaluation, while 147 missing trailing array values were retained only for state continuity and excluded from training. We utilized five-fold GroupKFold cross-validation by independent game seed, with each fold training on eight seeds and testing on two, to compare Linear Regression (OLS) against a Gradient Boosting Regressor.

Results. The live LLM data exhibited substantial stochastic variance. The non-linear Gradient Boosting model failed out of sample, producing a negative mean `R2` (-0.0843), meaning it performed worse than a flat mean prediction. The OLS model also lost predictive power relative to the structured logs.

Model	Mean Test `R2`	Mean RMSE
LinearRegression	0.3703	0.1465
GradientBoostingRegressor	**-0.0843**	0.1924

Conclusion. Approximately 63% of the live action variance remained unexplained by OLS. In our small-scale local Qwen2.5-1.5B stress test, raw live endpoints exhibited high variance and weak out-of-sample predictability. This motivates, rather than proves, the use of a calibrated proxy for causal RCTs; hosted frontier-model tests remain future work.

Appendix F: The Signal-to-Noise Ablation in Collusim

To test that conclusion in a controlled setting, we extracted the live residual standard deviation from Appendix E and injected the same zero-mean Gaussian noise into the Collusim OLS agent. We then reran baseline, Noise 0.40, and Hard Ban across 20 paired seeds.

Condition	Mean markup	Markup reduction [90% CI]
OLS baseline	0.690	-
Hard Ban	0.251	44.0% [39.8%, 48.5%]
Noise 0.40	0.210	48.1% [44.2%, 52.1%]

The injected noise standard deviation was 0.2983. Figure F1 shows that the clean policy menu becomes a scattered cloud. The marginal intervals for Noise 0.40 and Hard Ban overlap, but the paired-seed contrast remains small and distinguishable at -4.08 percentage points [-4.72, -3.42]. The ablation therefore shows a narrower and noisier policy margin, rather than complete indistinguishability.

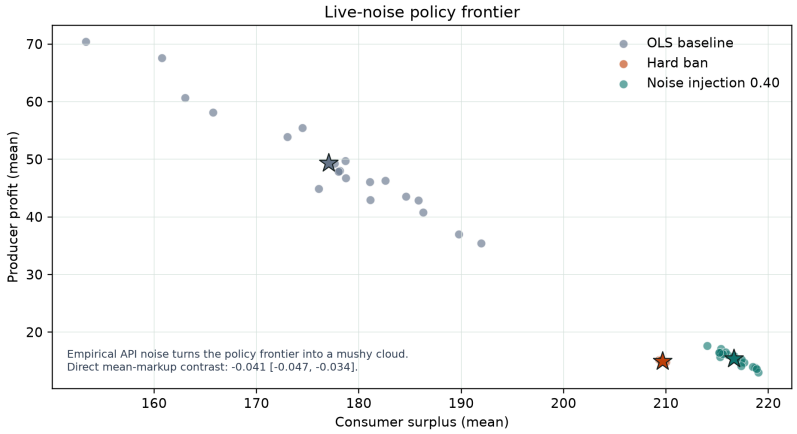


Figure F1. Live-noise policy frontier. Seed-level consumer surplus and producer profit under empirical live-LLM residual noise; stars denote arm means.

Conclusion. This ablation exposes the signal-to-noise limit of direct AI-market experimentation. Live LLM observations are valuable for identifying behavioral mechanisms such as price anchoring, but their high stochastic variance makes them unreliable for estimating the causal effect of a specific intervention. Controlled simulation provides the stable counterfactual needed for policy design.