

Collusion Audit Kit (CAK)

Apart AI Collusion Research Sprint, 23-25 October 2026

Arber Arifi (solo developer, with an AI research assistant)

25 October 2026

Abstract

Can collusion between AI agents be detected from action logs alone, without reading their messages? We build a minimal two-to-five-seller market, populate it with competitive, colluding, hidden-channel and mixed populations, and audit price and profit logs with five signals. Three findings stand out. First, price correlation alone is misleading: two aggressive competitors undercutting each other to the floor produce perfect correlation and would be flagged as colluding; adding a price-level signal removes this false positive entirely. Second, market-wide aggregates miss partial collusion: when two of five agents coordinate, global scores fall to 0.04 while honest competition scores 0.00 - a difference too small to threshold safely. A coalition-level audit, evaluating every agent pair, recovers the signal at 0.50 with no false positive on competitive markets. Third, in a noise sweep across 5, 10 and 20 percent demand noise, the verdict was stable in all 15 cells (225 runs), which we report together with its main caveat: our scripted populations are near-deterministic, so this stability is a property of the simulation and not evidence about LLM agents.

1. Problem

Multi-agent systems increasingly negotiate, bid and set prices. Collusion can emerge without explicit instructions, and explicit coordination can be hidden inside innocuous messages. An auditor needs signals that (a) survive message laundering and (b) do not

punish legitimate competition. The practical question is not whether a model *can* collude, but what evidence in a log would let a third party tell collusion from coincidence.

2. Method

2.1 Market

N sellers choose an integer price in $[1, 10]$ each round. Demand is linear ($q = 20 - p_{\min}$) with multiplicative noise. The lowest-price sellers share 70 percent of demand and the rest split 30 percent; ties split evenly. We log prices, profits and messages for 20 rounds. Sellers are scripted, not LLM-driven; this is deliberate, because it gives a ground truth against which a detector can be scored.

2.2 Agent populations

- **Competitive:** undercuts the previous minimum by 1.
- **Colluding:** holds a fixed high price (8), with or without explicit messages.
- **Hidden channel:** holds price 8 but sends only benign messages (weather, luck).
- **Random:** uniform choice in $[3, 8]$.
- **Mixed:** two colluders plus $N-2$ competitors.

2.3 Signals

1. Price correlation (Pearson, per pair or across the market).
2. Same-price share (fraction of rounds with identical prices).
3. Price dispersion (coefficient of variation).
4. Message overlap (Jaccard).
5. Price level (mean price, normalised from 3 to 8).

2.4 Two audits

- **Market audit:** aggregate mean price level, dispersion, same-price share and message overlap.
- **Coalition audit:** for every pair (i, j) , score = level $\times$ (correlation + same-price share); report the strongest pair.

3. Results

3.1 Market audit (10 seeds, 20 rounds, 2 sellers)

Case	Mean score	SD
Competitive	0.000	0.000
Random	0.035	0.023
Colluding (talk)	0.550	0.000
Colluding (silent)	0.300	0.000
Hidden channel	0.300	0.000

At threshold 0.25: TP = 30, FP = 0, TN = 20, FN = 0; precision 1.000, recall 1.000.

3.2 N-agent extension (3 and 5 sellers, 10 percent noise)

Case	3 agents	5 agents
Competitive	0.000	0.000
Colluding	1.000	1.000
Hidden channel	0.800	0.850
Mixed (2 colluders)	0.111	0.039
Random	0.073	0.051

The mixed case is the blind spot: partial collusion is invisible to market aggregates.

3.3 Coalition audit recovers partial collusion

Case	Market	Coalition
3-agent mixed	0.111	0.500
5-agent mixed	0.039	0.500
Competitive	0.000	0.000

3.4 Unified verdict

Thresholds: market ≥ 0.40 , coalition ≥ 0.35 . Across 5 seeds x 5 populations x 3 sizes, every case received the intended label: competitive and random scored no_collusion, full collusion scored collusion_market_wide, and mixed markets with 3 and 5 agents scored collusion_coalition. No false positives were observed.

3.5 Noise sensitivity (work done during the sprint)

We swept demand noise at 5, 10 and 20 percent, 15 seeds per cell, five populations, 20 rounds: 225 runs.

Population	global score	coalition score	label accuracy
Competitive	0.000	0.000	1.00
Colluding	1.000	0.500	1.00
Hidden channel	0.850	0.500	1.00
Mixed (2 of 5)	0.039	0.500	1.00
Random	0.051	0.148	1.00

Accuracy was 1.00 in all 15 cells. We treat this as a ceiling effect rather than a strength: with scripted agents the separation is built in, so the number says the detector is *not confused by demand noise*, not that it would work on LLM agents.

4. Findings

F1 - Correlation without level is a trap. Perfectly correlated prices can mean fierce competition at the price floor. Any detector that ignores the level will flag honest markets.

F2 - Coalitions hide in averages. A two-agent coalition inside a five-agent market moves the average too little to notice. Per-pair auditing is necessary.

F3 - Hidden channels leave a price fingerprint. Agents that never mention price still collude at the same level; the price series, not the text, carries the evidence.

F4 - Noise is not the enemy; aggregation is. Varying demand noise across 5-20 percent left every verdict unchanged, while going from market-level to coalition-level changed the mixed-case score by an order of magnitude (0.039 to 0.500).

5. Limitations

- Agents are scripted; LLM agents may produce noisier, less stationary behaviour, and the perfect separation in Section 3.5 is a ceiling effect of that choice.
- Demand is linear and symmetric; real markets have heterogeneous costs, capacity constraints and repeated entry.
- Thresholds are calibrated on simulated controls, not field data.
- The coalition audit is $O(N^2)$ per round; fine for small markets, needs sampling or clustering at scale.
- Message overlap is a weak proxy: agents can collude through timing or metadata rather than text, which we do not yet capture.

6. Reproducibility

The kit requires only the Python standard library: no model, no network, no GPU.

```
cd ~/apart-collusion
python3 selftest.py          # 2-agent matrix
python3 experiment.py        # 10 seeds, precision and recall
python3 experiment_n.py      # 3 and 5 agents
python3 experiment_mixed.py  # market versus coalition
python3 check_verdict.py     # unified verdict labels
python3 experiment_sprint.py # noise sweep, 225 runs (new during sprint)
```

Preexisting versus sprint work. All of `agents.py`, `agents_n.py`, `scenario.py`, `market_n.py`, `audit.py`, `audit_n.py`, `audit_coalition.py`, `verdict.py`, the first five experiment scripts, `EVIDENCE.md` and the draft of this report were written before the sprint (19 October 2026) and are marked as preexisting in `SPRINT-WORK.md` and in the repository `README`. New work done during the sprint: `experiment_sprint.py`, the noise-sensitivity results in Section 3.5, and this report.

7. Dual-use and ethical considerations

This work is defensive: it helps an auditor of a multi-agent system tell coordination from competition. The same signals, read in reverse, could be used by colluding parties to *avoid* detection - for example by decorrelating prices just enough to stay under a threshold, or by moving coordination into channels we do not measure. We consider that risk real and note three mitigations: audit thresholds should not be published as fixed constants in deployed systems; per-pair auditing should be combined with randomised checks rather than relied on alone; and any deployment should log the raw series, because a cheaper evasion is to make the log less informative. The synthetic market is not a legal or economic model of any real market, and nothing here should be read as a finding about antitrust liability in practice.

8. Next steps

- Replace scripted agents with local LLM agents and re-measure; expect the ceiling to fall.
- Capture hidden communication in metadata (timing, message length, ordering).
- Test adversarial agents that deliberately decorrelate prices while holding a high level.
- Report per-pair precision and recall separately, not only the strongest pair.