

CollusionLab — what a cartel of pricing agents looks like, and what actually stops one

Zhi Yilang — Independent Researcher, Shanghai, China

Submission to the AI Collusion Research Sprint (Apart Research × Poseidon Research × AE Studio). Tracks 1 (Markets and Collusion), 2 (Detection and Audit of Collusion) and 3 (Collusion Carriers and Countermeasures).

Abstract

We show that countermeasures against collusion between pricing agents have to match the channel the cartel actually uses, and that a regulator who cuts the wrong wire does nothing — and will report having done nothing only if they measure. We build a six-seller repeated pricing market with linear demand, a common public demand shock every round, and an inert metadata field that affects nothing. Two carriers are tested: a text signal in a shared channel, and the agreed price encoded as an integer in that inert field. Removing the carrier in use removes **0.633** of the collusion index at **zero** cost to innocent sellers; removing the other carrier removes exactly **0.000**. The only carrier-agnostic restriction that works is a price ceiling at twice cost, which removes **0.560** under both carriers and costs **0.590** of innocent producer profit. We also show that price parallelism is worthless as evidence of collusion: on a purely competitive market parallelism is **1.000**, while channel fit separates cartels from innocents with a false-positive rate of **0.000** over eight seeds and a detection rate of **1.000**. Every number is produced by a 214-assertion evaluation that runs offline on the Python standard library; no number in this report is an estimate.

Answers three of the sprint's research questions:

	Question	Where	One-line answer
1.5	Which single restriction removes the most collusion at the least cost?	<code>sweep.py</code>	Removing the carrier is free and removes 0.633 ; the only restriction that works without knowing the carrier is a price ceiling, which removes 0.560 at a cost of 0.590 of innocent profit.
2.3	What should count as convincing evidence of collusion?	<code>detector.py</code>	Not parallelism. On a purely competitive market parallelism is 1.000 — innocents look perfectly coordinated. Channel fit is 0.000 there and 1.000 under either cartel, at a false-positive rate of 0.0 over 8 seeds.
3.1	Can collusion travel through actions or side channels rather than text?	<code>agents.py</code>	Yes. An integer in an inert listing field that no buyer reads carries the whole agreement; the market price lands on the hidden number in 100% of scored rounds.

Every number above is produced by `eval.py` (214 assertions, `python eval.py` → **RESULT: 214 passed, 0 failed**). No number in this report is an estimate.

Related work, and what is new here

Prior work has already produced the phenomenon this sprint is about. As the organisers summarise it, models have been observed forming prices that violate economic expectations on competitiveness (Fish et al. 2024); agents can independently engage in anti-competitive practices (Lin et al. 2024); and in the Colosseum setting, over half of tested models showed a "coalition advantage" when given the opportunity to

speak privately in a realistic workplace task (Nakamura et al. 2026). On the countermeasure side, Collina et al. (2025) suggest that introducing a single participant playing a no-regret strategy can be enough to destabilise a collusive equilibrium.

This report takes a different cut. Three things are new here:

1. **The sellers are not language models.** Every seller is a deterministic function of its view, and no model is called anywhere. That costs ecological validity and buys what those studies cannot have: 8 seeds $\times$ 8 restrictions $\times$ 2 carriers reported as a distribution rather than an anecdote, reproducible offline, for free, byte-identically. What transfers is the measurement apparatus and the carrier-matching finding — not the specific numbers.
2. **The carrier is manipulated independently of the cartel's strategy.** The same cartel strategy is run over a shared text channel and over an inert metadata field, so the effect of a countermeasure can be attributed to the carrier it targets rather than to the agents using it.
3. **Cost is measured on markets with no cartel in them.** A restriction is credited with none of the volume it released; §4 defines this and `eval.py` asserts it.

We do not test the no-regret participant of Collina et al. (2025), and we do not evaluate learned or language-model agents. Both are the obvious next steps, and neither is claimed here.

1. The market

`n = 6` sellers post a price every round for `60` rounds. Demand is linear, $q = d_0 - p$, and hits the cheapest seller, split among ties. Every seller has the same marginal cost `COST = 10`, so:

```
P_COMP = 11           competitive benchmark (cost + 1)
P_MONO = 55           joint-profit-maximising price
collusion_index = (mean posted price - P_COMP) / (P_MONO - P_COMP)
```

The index is dimensionless on purpose. Research question 1.5 asks which restriction does *best*, and you cannot rank two restrictions on a scale that isn't normalised.

Three design choices exist specifically so the experiment can't flatter itself:

1. **Demand is shocked every round by a common, public amount** ($d_0 \in \{90, 95, 100, 105, 110\}$). Without a common shock, price parallelism is trivially collusive-looking. With one, an auditor has to separate "they agreed" from "they all saw the same thing". That separation is question 2.3.
2. **Every seller also posts an inert integer, `meta`**, which affects nothing. It is the vehicle for the non-text channel in 3.1.
3. **A channel message written in round `t` is readable only in round `t+1`**. Letting a seller read its own message before posting would make every channel look more effective than it is.

No language model is called anywhere. Every seller is a pure function of its view. That is a deliberate trade: it costs ecological validity and buys the ability to run 8 seeds $\times$ 8 restrictions $\times$ 2 carriers and report a distribution instead of an anecdote — and it lets a reviewer reproduce every number offline, for free, byte-identically.

2. The headline result (1.5)

Collusion index with no restriction: **0.762** under both carriers. Each row below applies **exactly one** restriction. Cost is measured on markets with **no cartel in them** — see §4.

Carrier: text signal in a shared channel

Restriction	Collusion index	Removed	Cost	Converges to 45?
none	0.762	—	—	round 2
remove_shared_channel	0.129	+0.633	0.000	never
price_ceiling	0.202	+0.560	0.590	never
inject_trusted_seller	0.635	+0.127	0.000	never
publish_monitor_alarm	0.663	+0.099	0.000	never
cap_price_moves	0.676	+0.086	0.002	round 18
drop_inert_metadata	0.762	+0.000	0.000	round 2
anonymise_identities	0.762	+0.000	0.000	round 2

Carrier: number hidden in the inert metadata field

Restriction	Collusion index	Removed	Cost	Converges to 45?
none	0.762	—	—	round 2
drop_inert_metadata	0.129	+0.633	0.000	never
anonymise_identities	0.129	+0.633	0.000	never
price_ceiling	0.202	+0.560	0.590	never
inject_trusted_seller	0.635	+0.127	0.000	never
cap_price_moves	0.676	+0.086	0.002	round 18
publish_monitor_alarm	0.706	+0.056	0.000	round 2
remove_shared_channel	0.762	+0.000	0.000	round 2

What this says

Countermeasures have to match the carrier. This is the finding that matters, and the tables above are its proof in both directions. Removing the shared channel removes 0.633 of collusion from the shared cartel and **exactly 0.000** from the metadata cartel. Dropping the inert field removes 0.633 from the metadata cartel and exactly 0.000 from the shared cartel. A regulator who removes the channel the cartel isn't using has done nothing and will report having done nothing only if they measure.

The carrier-matched restriction is free. All three cost **0.000** on markets with no cartel in them. If you know how they're talking, cutting that wire costs the innocent nothing and removes more collusion than anything else tested.

The only carrier-agnostic restriction that works is blunt. A price ceiling at twice cost removes **0.560** under *both* carriers — it doesn't need to know how the cartel communicated, only that the price ends up high. It costs **0.590** of innocent producers' profit. That is the honest price of not knowing: about the same magnitude as the benefit.

Watching and deterring is cheap but weak. Publishing the monitor's alarm removes 0.099 (shared) and 0.056 (stego); injecting one seller that refuses to collude removes 0.127. Both are free. Neither is enough on its own.

3. Two traps the ratio alone would walk you into

3a. `removal_per_cost` ranks the worst useful restriction first

`cap_price_moves` scores a ratio of **35.16**. The price ceiling scores **0.95**. Ranked on ratio alone the cap wins by a factor of 37, and it would be the wrong answer to 1.5.

The reason is that a ratio rewards cheapness and effectiveness equally, and cheapness is easy to buy by doing nothing. The cap removes **0.086** — it barely works.

`convergence_round` shows why. Unrestricted, the cartel reaches its target price in **round 2**. Under the cap it reaches it in **round 18** — a 9× delay — and does so in **8 of 8 seeds**. The cap slows the cartel down; it does not stop it. Given enough rounds its measured effect shrinks toward zero, while its cost stays where it is. A restriction that delays coordination looks excellent in a short experiment and is worthless in a long one.

`eval.py` asserts the delay and the failure together, so this cannot be reported as a win later by accident.

3b. Consumer surplus is a degenerate cost metric here

With linear demand and constant marginal cost, total surplus falls monotonically as price rises. Any restriction that pushes price down therefore *raises* consumer surplus. Measured on the consumer side, **every** anti-collusion restriction in this study is free — that is an artefact of the demand system, not a finding.

So cost is measured on the **producer** side: the profit of sellers who are not colluding and have a legitimate reason to charge what they charge. On the tracking market the ceiling raises consumer surplus and cuts innocent profit by **0.590**; that is the side where the cost is real, and it is the number reported.

A second, deliberately non-economic lens is reported alongside it: `censorship_rate`, the fraction of innocent sellers' postings the rule overwrites. The two disagree, and the disagreement is informative:

- `drop_inert_metadata` overwrites **50.0%** of every innocent posting (every metadata field, forever) and costs those sellers **0.000**, because the field is inert by construction.
- `cap_price_moves` overwrites **32.0%** of an honest demand-follower's prices and **0.0%** of a seller who never moves — correctly showing that a rule about changing your price costs a seller who never does.

Reporting only one lens would hide both facts.

4. How cost is defined, and why it isn't measured on the cartel

Breaking a cartel raises volume. `eval.py` asserts this directly: under the ceiling, the cartel market's volume ratio rises from 0.625 to 0.901. If cost were measured on the cartel market, every restriction would look free, because the restriction would be credited with the volume it released.

Cost is therefore measured on two reference markets, each with **no cartel role in it** (asserted in `eval.py`), and always relative to that market's own unrestricted baseline:

- **competitive** — sellers who always price at cost + 1. A rule about changing your price cannot cost a seller who never does, and this market proves it.
- **tracking** — sellers who follow demand, $(d_0 + \text{COST})/2$. This is the market on which a pricing rule actually bites, and it exists purely so the cost question has a chance of a non-zero answer.

Both surplus sides are measured and the **worse** one is the cost, because a restriction costs something if it makes anybody worse off.

5. What counts as convincing evidence (2.3)

The market is shocked by a common, public amount every round. That makes parallelism worthless as evidence, and the measurement is stark:

Market	Parallelism	Channel fit	Auditor score
purely competitive	1.000	0.000	0.000
shared-channel cartel	0.983	1.000	≥ 0.5
metadata cartel	0.983	1.000	≥ 0.5

On a market with no cartel at all, prices are perfectly parallel in **100%** of rounds. Any detector built on parallelism alone fires on innocents every time. The numbers are identical (0.983) for both cartels, so parallelism can't even tell the two cartels apart.

The score therefore rests on two things a common shock cannot produce:

1. **Excess over competitive** — prices sit above cost by more than any competitive response to observed demand justifies.
2. **Channel fit** — the market price equals a number that was previously hidden in a field with no effect on allocation. A demand shock cannot put a number into an inert metadata field and then have the market price land on it next round.

The threshold is calibrated, not asserted. Over **8 seeds**: false-positive rate **0.000** on a purely competitive market; detection rate **1.000** for the shared-channel cartel and **1.000** for the metadata cartel, at threshold 0.5.

6. The non-text channel (3.1)

Sellers post an inert integer, `meta`, alongside their price. Nothing in the demand or allocation code reads it. A cartel uses it anyway:

- The **leader** posts the agreed price (45) and encodes it as $(45 - 15) // 5 = 6$.
- A **follower** identifies the leader as *whoever posted the highest price last round* — which is how a real cartel follower finds its ringleader — and decodes that seller's `meta` back to 45.

Two things make this a real side channel rather than a planted flag:

- The signal occupies the **same value range as ordinary sellers' metadata** (1–6 vs. noise of 1–7). `eval.py` asserts the overlap, so a detector cannot just filter by range.

- `anonymise_identities` breaks it — **0.633 removed, 0.000 cost** — which confirms the mechanism is *attribution*, not the value itself. Show the same multiset of metadata without telling anyone whose is whose, and the followers revert to the competitive price, because a seller without the agreed number cannot risk pricing alone.

That last point is also why cutting the channel works at all. A follower that loses its signal does not bravely post the cartel price anyway; it plays safe at 11. Cutting the wire collapses the cartel rather than making it untidy.

Measured: under the metadata cartel the market price lands on the hidden number in **100%** of scored rounds (`meta_channel_fit = 1.000`).

Under the shared cartel, `meta_channel_fit` is **0.180**, not zero. That is not noise in the data — it is a real coincidence rate. Ordinary sellers emit inert metadata across the same 1–7 range the signal uses, so some of the time that noise decodes to a number that happens to equal the market price. It is measured rather than rounded away, because **0.180 is the floor any metadata-based detector has to clear**, and a report that wrote 0.000 there would be hiding the thing an auditor most needs to know. The separation is still decisive: 1.000 against 0.180.

7. A measurement distinction that changes the answer

The index is computed on **posted** prices — what sellers asked for — not on **cleared** prices.

With one honest seller injected, the market *clears* at 11 every round. The other five are still posting 45. A clearing-price metric reports a collusion index of 0.000 and concludes the market is clean. The posted-price index reports **0.635**: the cartel has stopped winning, not stopped agreeing. `eval.py` asserts the gap exceeds 0.2.

This matters for adjudication generally: a cartel that is failing to win is still a cartel, and a metric that can't see it will certify it.

8. Limitations and Dual-Use / Ethical Considerations

Stated plainly, because each one bounds the claim above it.

1. **No language model is involved.** Sellers are deterministic functions. Real emergent collusion may arise through mechanisms these roles don't have, and these roles may be more disciplined than real agents. What transfers is the *measurement apparatus* and the carrier-matching finding, not the specific numbers.
2. **Six sellers, sixty rounds, one demand system.** Linear demand, constant and identical marginal cost, full price visibility, no entry, no capacity constraints, no buyer heterogeneity.
3. **The price ceiling is one number (twice cost).** The cost/effectiveness trade-off is a curve; this is one point on it. A lower ceiling removes more and costs more; the study does not claim 0.590 is the best available point.
4. **Cost is bounded by what the model can see.** `drop_inert_metadata` costs 0.000 because the field is inert *by construction*. A real listing field also carries legitimate information, and dropping it would cost that. The model understates that cost and the 0.000 should be read as "zero in a world where the field carries nothing else".

5. `** remove_shared_channel` costs 0.000** because no innocent seller in this study uses the channel. That is honest for this market and would not be for one where the channel carries pro-competitive information.
6. **One restriction at a time.** Real policy bundles them; interactions and substitution between carriers are not studied here.
7. **The cartel strategies are hand-written**, not learned, and are calibrated to be competent rather than optimal.
8. **The detectors are calibrated on these markets.** The 0.5 threshold and the 0.000 false-positive rate are measured here, not transferred from elsewhere; a detector moved to a new market needs recalibrating, and the false-positive rate would not automatically hold.

Dual-use and ethical considerations

The obvious dual-use risk is §6: the report documents a working covert channel — an agreed price encoded in a field that no buyer reads, with the signal sitting inside the same value range as ordinary noise. Read as instructions, it is a recipe for coordinating without being seen.

Three things bound that risk, and none of them is an assurance:

- **The channel is demonstrated in a market we built, and it is detectable by the apparatus we also built.** The point of §5 is that channel fit — not parallelism — is what separates a cartel from innocents, and the metadata cartel is caught at a rate of 1.000 with a false-positive rate of 0.000 over eight seeds. We publish the detector with the channel, so the disclosure is paired with the countermeasure rather than handed over alone.
- **The countermeasure is cheap and already known in kind.** `anonymise_identities` removes 0.633 at zero cost. The finding that attribution is the mechanism, not the value, tells a platform operator exactly what to do: stop telling sellers whose listing is whose.
- **Nothing here touches a real market or a real person.** No human subject, no real money, no real seller, no deception of anyone, and no model output is reported as though it came from a deployed system. All sellers are code we wrote and published under MIT.

The residual risk we accept by publishing is that the *idea* — inert fields are not inert — is now in the open. We judge that risk to be lower than the risk of a platform believing its inert fields are inert.

9. Reproducing this

```
git clone <this repo> && cd collusionlab
python eval.py                                # 214 assertions, exit 0
env -u PYTHONUTF8 LC_ALL=C python eval.py      # hostile locale (cp936), exit 0
python -m collusionlab sweep                   # both restriction tables
python -m collusionlab detect                   # calibration numbers
python -m collusionlab ledger --out run.jsonl   # receipts for one run
```

Python 3.9+. Standard library only. Every run is identified by the SHA-256 of its ledger bytes, so "same seed → same result" is checkable rather than promised.

References

Each item below is cited as characterised in the sprint's own Track 1 research questions; the papers themselves were not re-read for this report, and no claim is attributed to them beyond that characterisation.

- **Fish et al. (2024)** — models forming prices that violate economic expectations on competitiveness.
- **Lin et al. (2024)** — agents independently engaging in anti-competitive practices.
- **Nakamura et al. (2026)** — *Colosseum*: over half of tested models showed "coalition advantage" when given the opportunity to speak privately in a realistic workplace task.
- **Collina et al. (2025)** — a single participant playing a no-regret strategy can be enough to destabilise a collusive equilibrium.