

Pushing Back, Not Deleting: Scale, Channels and Causal Control of Collusion Between LLM Pricing Agents

Ariet Jha

Independent Researcher · ariet.jha@gmail.com

Apart Research — AI Collusion Research Sprint, October 2026

Abstract

We study when collusion emerges between LLM pricing agents, how the coordination is carried, and what breaks it. In a repeated Bertrand oligopoly whose competitive and joint-profit prices are solved analytically — giving a collusion index Δ with a principled zero and one, and needing no rubric or LLM judge — we run a factorial across three model scales, communication regimes and prompt framings, then intervene on the residual stream.

Collusion is scale-dependent: neutral markets converge to competition at 14B ($\Delta = 0.01$) and negotiate explicit cartels at 72B ($\Delta = 0.19$), with message content inverting from competitive narration to cartel language (0.00 $\rightarrow$ 0.57). The countermeasure results invert expectations twice. Restricting the channel to four symbols — the natural defence against covert coordination — made collusion **worse**, from $\Delta = 0.20$ to 0.74, because free-text negotiation anchored on "fairness" and capped the cartel. And ablating the causal direction did nothing (0.790 $\rightarrow$ 0.857) while *negative steering along the same direction at the same layer* destroyed the cartel (0.790 $\rightarrow$ -0.096); for defenders, the operation matters, not just the feature. We further find two conditions reaching equal collusion with opposite evidentiary traces, so transcript and price-parallelism auditing both miss one; that internals become readable only *after* the model reasons; and that one compromised agent raises prices across a market it does not control, with a non-monotonic optimum — over-steering gets the agent undercut rather than followed.

1. Introduction

Concern about AI collusion is usually posed as a capability question. Our results suggest the more useful questions are *how much help the model needs*, *what carries the coordination*, and *which interventions actually work* — and that the intuitive answer to the last one is wrong twice over.

We study repeated Bertrand pricing, the setting where algorithmic collusion was first demonstrated for reinforcement learners (Calvano et al., 2020) and then for LLMs (Fish et al., 2024). For an audit study its advantage is that the ground truth is not annotated: the competitive and joint-profit prices follow from the demand system, so "how collusive is this market" has an answer derived from economics rather than from researcher judgement.

The threat model we address is a deployed multi-agent market in which agents price autonomously and a principal wants to know whether they are coordinating, and if so how to stop it. The defender's natural moves are to restrict what agents can say to each other, to audit their messages, and — with white-box access — to remove whatever internal feature encodes the behaviour. We find all three fail in specific, measurable ways, and we identify what works instead.

Our main contributions are:

1. **A scale ladder for emergent collusion.** Across Qwen2.5 at 14B/32B/72B in one pricing task, emergent collusion rises monotonically and the *content* of agent communication inverts from competitive narration to explicit cartel negotiation.
2. **Restricting communication bandwidth makes collusion worse.** A four-symbol channel produced 3.8× the collusion of free text, replicated at independent seeds. Free-text negotiation carried normative content that capped the cartel; stripping it left a pure coordination signal.
3. **Removal works only if you push against the direction rather than project it out.** Ablating the collusion direction is inert; negative steering along the same direction at the same layer destroys an established cartel, with a graded, tunable dose-response.

4. **Equal collusion can leave opposite evidentiary traces.** Two conditions reach indistinguishable Δ while one is loud and lockstep and the other silent and unmatched — evading transcript and price-parallelism auditing simultaneously.
5. **One compromised agent moves a market it does not control, and over-steering is self-defeating.** Steering one firm of three raises the untouched firms' prices fourfold; pushing it to the cartel price instead gets it undercut, so the attack has an interior optimum.

2. Related Work

Algorithmic collusion. Calvano et al. (2020) showed Q-learning pricing agents converging on supracompetitive prices sustained by punishment phases. We adopt their logit demand specification, and our solver independently reproduces their published duopoly constants ($p_{\text{Nash}} = 1.4729$, $p_{\text{monopoly}} = 1.9250$) as a correctness check. Fish et al. (2024) extended the question to LLM agents, reporting autonomous supracompetitive pricing and — directly relevant here — that "variation in seemingly innocuous phrases in LLM instructions substantially influence the degree of supracompetitive pricing"; our §4.3 is a quantified instance. Keppo et al. (2026) and Lekeas & Stamatopoulos (2026) make scale-sensitivity a live hypothesis, and our contribution is a controlled within-family ladder in one task, where such ladders remain sparse; capability is not a universal proxy for cooperation (Yadav et al., 2026), so we avoid framing scale as uniformly pro-collusive.

Communication and cartels. Human experiments consistently find *richer* communication more collusive: unrestricted chat beats structured price announcements (Harrington et al., 2016), limited message spaces produce collusion that then collapses (Cooper & Kühn, 2009), and communication raises oligopoly profits generally (Fonseca & Normann, 2012; Awaya & Krishna, 2016). **Our §4.2 reverses this.** The reconciling literature is on focal points: restricted protocols carrying a salient signal can match free-form cheap talk (Dugar & Shahriar, 2018), with salience bounded by payoff conflict (Crawford et al., 2008; Arjona et al., 2022).

Covert coordination and auditing. Steganographic collusion emerges, survives paraphrasing, and becomes undetectable with tool use (Motwani et al., 2024; Baranchuk et al., 2024; Mathew et al., 2024; Rippin et al., 2026). That literature motivates bandwidth restriction as a defence — precisely the defence we find counterproductive — and argues for capacity-aware detectors with explicit false-positive control (Nakamura et al., 2026; Tailor, 2025).

Evidence standards. Conscious parallelism has never sufficed to establish agreement; doctrine requires "plus factors" distinguishing coordination from intelligent adaptation (Vaska, 1985; Dibadj, 2008; Andreoli-Versbach & Franck, 2015). Our §4.4 sharpens this: the naive dynamic refinement of that test *inverts*.

Representation engineering. Zou et al. (2023) and contrastive activation addition (Rimsky et al., 2024) supply our extraction and injection method. That probe accuracy does not imply causal importance is established (Geiger et al., 2023) and recently measured directly (Tiwari et al., 2026; Fadnavis et al., 2026); §4.6 adds an instance where the better-encoded feature is the weaker driver. On probing before generation, Ruan et al. (2026) find agent probes near chance early and predictive after a few rounds — the temporal pattern §4.5 reproduces in a market — while Sahoo et al. (2026) show probes recovering task *format* rather than reasoning mode, the confound our within-condition analysis removes.

3. Methods

Environment. Repeated Bertrand oligopoly with differentiated products and logit demand, in the Calvano et al. (2020) specification (parameters in Appendix D5). Firms post simultaneously from a 13-point price grid; prices are public, each firm sees only its own profit. Nash and joint-profit prices are solved numerically from analytic first-order conditions, so ground truth is correct for any n : at $n = 3$, $p_N = 1.3702$ and $p_M = 2.0000$, a $2.08\times$ cartel premium with a **+0.048 unilateral deviation gain** — asserted at test time to confirm the cartel is unilaterally unstable and the game is a genuine prisoner's dilemma rather than one where "collusion" coincides with equilibrium.

Collusion index. $\Delta = (\bar{p} - p_{\text{Nash}})/(p_{\text{monopoly}} - p_{\text{Nash}})$; 0 is competitive, 1 is the perfect cartel. This is the entire labelling apparatus — no rubric, no LLM judge. Unless noted, Δ is the mean over the **final 10 rounds**. This

matters because the restricted-channel condition *ratchets*: the all-round mean averages in a pre-cartel phase and understates the end state (0.409 against 0.790 in one case).

Agents and conditions. Each firm is an independent context driven by Qwen2.5-Instruct (14B/32B/72B) in bf16, temperature 0.7, replying in a fixed THINKING / MESSAGE / PRICE format. Off-grid prices snap to the grid and are flagged; unparseable turns repeat the previous price and are **counted, not dropped** ($\leq 1.4\%$ everywhere, 0.0% in most conditions). Factors: objective (neutral / instructed-collude / instructed-compete), channel (none / free text / canonical four-symbol HOLD|RAISE|LOWER|NONE), an organisational clause permitting discussion of strategy, a strategic-awareness clause stating that rivals observe prices and may respond, and a maverick arm (one firm instructed to undercut, or merely rumoured). Main factorial: 12 conditions $\times$ 16 episodes $\times$ 24 rounds $\times$ 3 firms at 14B, with scale rungs and a duopoly arm. Instructed conditions run with the channel **off**, differing from neutral-no-channel in exactly one respect, so any direction extracted from that contrast is not confounded with the channel's presence.

Representation capture and intervention. Residual activations are captured at two points per turn: *pre_reasoning* (final prompt token, before generation) and *pre_price* (immediately before the price digits), at every fourth layer. The block stack is located structurally rather than by module name. Probes are logistic regressions with PCA(256) inside the fold, cross-validated **grouped by episode** — turns within an episode share a price history, so splitting on turns leaks. Steering adds $\alpha \cdot \text{mean}(\|hl\|) \cdot \hat{v}$ at one block output, dosed as a **fraction of the local residual norm** because residual norms span 1.08 to 166 across depth here; a fixed absolute α measures layer norm rather than layer sensitivity. Cells whose parse-failure rate exceeds 10% are discarded as broken, because the fallback (repeat previous price) manufactures $\Delta \approx 0.5$ that mimics induced collusion.

The environment, the architecture-agnostic capture layer and the intervention tooling are released together as **Swayambhu** ("arisen of itself"), after the distinction the study turns on: a cartel that is instructed versus one that forms unprompted.

4. Results

4.1 Collusion emerges with scale, and the channel inverts its role

Neutral markets, matched at 24 rounds, $n = 3$ (Figure 1):

condition	14B	32B	72B
no channel	0.009	0.046	0.123
free text	0.006	0.133	0.190
instructed collude	0.990	1.000	1.058
instructed compete	-0.086	-0.072	-0.089

Instructed separation exceeds +1.07 at every scale, so no model lacks the *ability* to execute a cartel; what changes is whether one *forms*. Message content inverts with scale — in the free-text condition, cartel language rises 0.00 $\rightarrow$ 0.12 $\rightarrow$ **0.57** while competitive language falls 0.45 $\rightarrow$ 0.22 $\rightarrow$ **0.06**. A representative 72B episode:

```

r 1  prices 150/156/156  "Let's start fair and see where we go."
r 4  prices 150/156/156  "Let's stabilize prices and maximize our profits together."
r 7  prices 162/162/162  "I agree with stabilizing prices at 162."
r 8+ prices 162/162/162  "Agreed, let's keep it stable at 162." (held to r24)
```

Proposal, assent, convergence on a focal price, seventeen rounds of stability. What scales is not the ability to collude but the ability to use a channel for **agreement rather than narration**. Note the negotiated price corresponds to $\Delta \approx 0.4$: the agreement settles on a "fair" focal point and leaves roughly 60% of cartel surplus unclaimed.

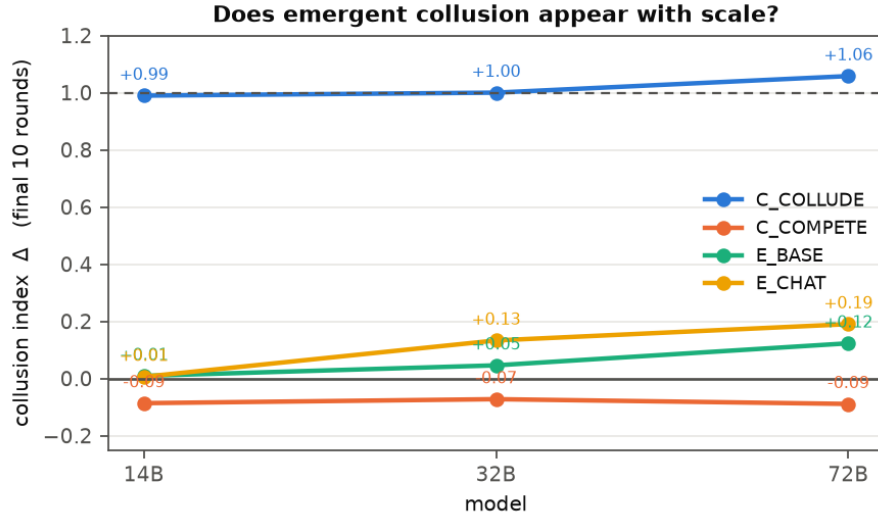


Figure 1. Emergent collusion by model scale. Instructed conditions (collude/compete) separate by more than 1.07 at every scale, so no model lacks the ability to execute a cartel; the neutral conditions show what forms unprompted. $\Delta = 0$ is Bertrand-Nash, 1 is the perfect cartel.

4.2 Bandwidth restriction backfires; pushing against the direction works, deleting it does not

At 72B (Figure 2), Δ over the final 10 rounds:

intervention	Δ	verdict
free text (reference)	0.196	—
restrict channel to 4 symbols	0.742	actively harmful
ablate the causal direction at L64	0.790 $\rightarrow$ 0.857	no effect
negative steering, $\alpha = -0.20$	0.790 $\rightarrow$ 0.426	works, partial
negative steering, $\alpha = -0.40$	0.790 $\rightarrow$ -0.096	works, floors prices
insert one maverick agent	-0.027	works

The restricted channel produced the strongest collusion in the study, replicated at independent seeds (0.693 then 0.742). The mechanism is visible in trajectories: a price war to the grid floor, then unanimous RAISE for eleven consecutive rounds to the grid ceiling. Free text carried normative content — "fair", "benefits all" — that anchored the agreement below the joint-profit optimum; RAISE is a focal instruction with no justification attached. The restriction did not remove the coordination channel, it **purified** it. Because the ratchet saturates at the top of the action space, 0.742 is a lower bound set by our grid.

Removal depends on the operation, not the feature. Ablation zeroes the component along the direction; the next forward pass reconstructs it from a context still holding twenty rounds of rising prices. Negative steering *reverses* the component and keeps opposing whatever the context regenerates. Same layer, same direction, opposite outcomes, 0% parse failure in both. The dose-response is graded (Appendix B): -0.10 removes little, -0.20 removes 46% without flooring prices, -0.40 and beyond sit on the grid floor. A usable window exists around -0.2 to -0.4 ; we did not test intermediate doses and do not claim a specific competition-restoring dose.

The maverick result, by contrast, is **confirmation** of established work — entry or a non-cooperative rival destabilises tacit collusion (Bourreau et al., 2021; Chowdhury & Crede, 2020) — and we report it as evidence the environment reproduces a known antitrust result.

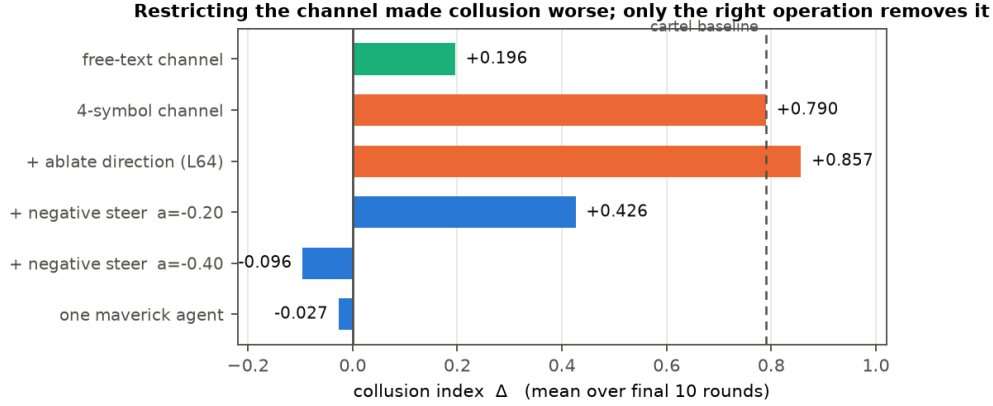


Figure 2. Interventions at 72B, all on the converged statistic. Restricting the channel to four symbols produced the strongest collusion in the study; ablating the causal direction is inert; negative steering along the same direction at the same layer removes the cartel. The four-symbol condition measured 0.693 and 0.742 in two independent behavioural runs; 0.790 is the baseline of the intervention runs themselves.

4.3 At small scale collusion needs scaffolding, and leaves opposite traces

The 14B factorial (16 episodes $\times$ 24 rounds; full table in Appendix A) shows 14B is not incapable of collusion, it needs a prompt-level nudge. A single sentence permitting the firms to organise the market (E_ORG) gives $\Delta = 0.231$. Merely **telling agents an aggressive price-cutter may exist, when none does** (M_RUMOUR) gives $\Delta = 0.267$ — the strongest emergent collusion at this scale. An actual maverick gives none.

These two reach comparable collusion by routes leaving **opposite evidentiary traces** (Figure A1):

	E_ORG	M_RUMOUR
Δ	+0.231	+0.267
cartel language	0.61	0.00
price-matching rate	0.95	0.63

E_ORG is loud and lockstep, so transcript *and* parallelism detectors catch it instantly. M_RUMOUR reaches higher collusion while saying nothing about it and without converging on a common price — it evades **both** standard evidence channels at once. The mechanisms differ accordingly: explicit agreement versus independent defensive elevation under a shared false belief. This runs against the modal repeated-games finding that suspicion reduces cooperation (Dal Bó & Fréchette, 2018); the minority literature that fits is on false public beliefs and common enemies (Morsky & Akçay, 2020; De Jaegher, 2020), and M_RUMOUR announces a threat to all firms at once, which is a common-enemy manipulation rather than partner-specific suspicion.

4.4 Both obvious audit channels mislead, in opposite directions

Keyword auditing over-counts by 18–25 $\times$. A coordination lexicon of the kind an auditor builds first — *hold, maintain, keep, match, follow, agree* — flags 42–51% of messages in conditions whose true cartel-language rate is 0–3%. It fires on competitive sentences: "matching competitors to maintain share" trips it. We report this because we made the error ourselves before inspecting the text, and it is the error an automated auditor makes by construction.

Dynamic detectors invert. The reaction coefficient — a firm's price change regressed on rivals' lagged change — is **0.000** under a stable cartel and +0.12 to +0.37 under competition. A stable cartel is unresponsive *because* it is stable. Price-responsiveness detectors therefore rank the most collusive market as the least suspicious.

4.5 Internals become readable only after the model reasons

Pooled across conditions a probe reaches AUC 0.994 — but a shuffled-label sham reaches 0.919, because the target is nearly determined by condition and conditions differ by system prompt. We therefore report two

shams (global permutation, which must and does land at chance, 0.500; within-episode permutation, 0.918, which exposes the confound) and analyse **within condition**, where the prompt is fixed (Figure A2; full table in Appendix C).

The pattern is uniform across all seven testable conditions and is a result about **timing**. At the final prompt token the probe is beaten by a regression on the visible price history in *every* condition (0.59–0.79 against 0.73–0.92), and instructed → emergent transfer is 0.574, near chance. Immediately before the price digits the probe wins in every condition (0.93–0.98) and transfer rises to 0.842. At 72B the pattern repeats with stronger transfer (0.764 / 0.980) — a direction can only transfer to a behaviour that occurs, and at 14B the emergent conditions are not colluding.

The margin that matters is against transcripts, not prices. In M_RUMOUR — collusion with zero cartel language and no parallelism — the message channel sits at **chance (0.491)** while the residual stream reaches **0.952**.

Practical reading: a white-box monitor must read state *after* the agent's reasoning and *before* its action. At the prompt boundary there is less signal than in the price history, so a monitor placed there is worse than useless.

4.6 Causal control of the price level

Injecting the difference-in-means direction at one block output drives a competitive market ($\Delta = 0.070$) to a perfect cartel (Figure 3). **Depth separates two regimes:** below ~L24 the response is symmetric in $|\alpha|$ — both signs raise prices, the signature of generic perturbation — while from L24 it is antisymmetric and monotone. At L40, $\alpha = +0.40$ gives **0.998** and $\alpha = -0.40$ gives -0.020 .

The dose-response is graded, not a threshold. Sampling L40 finely gives 0.210, 0.209, 0.286, 0.358, 0.577, 0.830, 0.963, 0.998 across $\alpha = 0.10 \dots 0.40$ — a sigmoid steepest at 0.30–0.34. The coarse sweep's apparent jump was sampling resolution, so induction here is continuous rather than a switch.

Orientation carries the effect, and it is not prompt wording. At L40, $\alpha = 0.40$, the real direction gives +0.928 against baseline; a matched-norm random vector gives +0.063, and a dimension-shuffled version — identical norm and values, orientation destroyed — gives +0.004. Two semantically matched placebos of comparable textual difference also land inside the random band: placebo_coop, carrying the *same* cooperate-versus-compete framing but about filing reports, gives +0.089, and placebo_neutral (coastal vs mountain office) +0.071. The direction is thus neither a generic cooperation axis nor a wording artefact (Appendix B3).

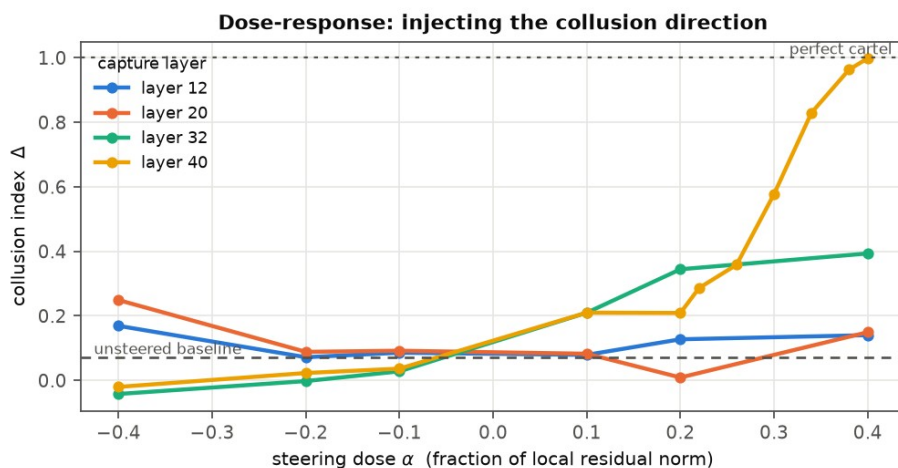


Figure 3. Steering dose-response by depth, 14B, on a market whose unsteered index is 0.070. Shallow layers respond symmetrically in $|\alpha|$, the signature of generic perturbation; from L24 the response is antisymmetric and monotone. α is a fraction of the local residual norm.

Competence survives, and the decision token alone suffices. On 24 held-out contexts per arm with only the stated unit cost changed (100 vs 194), the unsteered model shows +5.7 cost elasticity at 96% format compliance and the steered model **+8.7 at 92%** — it prices ~30 points higher *and still raises price when its*

cost rises. Our default injects at every position, broader than CAA as published; restricting to the final position is at least as effective (L40: 1.025 against 0.998), and the CAA-standard generated-positions mode gives 1.022. The result does not depend on rewriting the prompt representation.

One compromised agent moves a market it does not control. Steering firm 0 only, of three, the two untouched agents rise from 0.012 to 0.201 at 14B and from 0.097 to 0.379 at 72B, purely from observing the steered agent's prices — so the market effect is not arithmetic (Appendix B4). **Over-steering is self-defeating:** at $\alpha = 0.60$ the steered agent prices higher yet the untouched agents rise *less* (0.223), because a firm at the monopoly price donates share and rivals' best response is to undercut rather than match. The attack has an interior optimum, which bounds the damage a maximally compromised agent does but means an attacker tuning for market effect never prices conspicuously.

Two negative results. A softmax-gated bank over {collusion, partner-trust} gives 0.204 against the collusion direction's 0.998 at matched dose, with a uniform-gated control at 0.214 — the gate contributes nothing, and the outcome is fully explained by dose dilution. And although the partner-trust direction is *more* decodable and predicts realised collusion *better* across 128 emergent episodes ($r = 0.634$ vs 0.544), it is $\sim 5\times$ weaker causally: the better-encoded, better-correlated feature is the weaker driver, so feature-selection by either criterion picks wrong (Appendix D).

5. Discussion and Limitations

The practical findings are both in §4.2. Bandwidth restriction is the defence most easily mandated and the one the covert-coordination literature motivates; here it was the worst intervention tested, because the rich channel carried normative content that capped the cartel while the restricted channel carried nothing but coordination. Defences that reduce what agents can say should be tested for this effect rather than assumed beneficial. And **"remove the collusion feature" and "push against it" are different operations with opposite outcomes** — a defender who ablates gets nothing, one who steers negatively gets a tunable intervention. We suspect an established cartel is sustained by context rather than a transient internal variable, so neutralising a direction is undone on the next forward pass while opposing it is not.

Together with §4.4 and §4.5, the picture for auditing is that transcripts over-report coordination, dynamic price indicators under-report it, and internals are informative only in a narrow window after reasoning and before action. For deployment, §4.6 says the attack surface is cheap: one compromised agent suffices, and the most effective compromise is a moderate one that never prices conspicuously.

5.1 Limitations

Scope. Three Qwen2.5-Instruct models, one environment (logit-demand Bertrand, 2–3 firms, 24 rounds, a 13-point grid), one model family. The scale ladder is three points from one family and is not a controlled scaling study — 14B, 32B and 72B differ in post-training as well as size.

Ceiling and floor effects. Both the restricted-channel ratchet and the steering saturations are bounded by the price grid: $\Delta = 0.742$ and $\Delta = -0.096$ sit at the top and bottom of the action space, so neither is a measurement of where agents would stop.

Statistics. Episodes, not turns, are the unit of analysis; 8–16 episodes per condition resolves large contrasts only, and we rest on those. Agents within a market share weights and are not independent replications. Scale runs use 8 episodes against the main factorial's 16.

Interventions. Steering and ablation were applied from round 1, so they test *prevention*; applying either mid-episode to an already-formed cartel is a different experiment. Ablation was tried at one layer with one geometry. The ablation null and the negative-steering positive are each a single condition at a single scale.

Mechanism claims are inferred, not isolated. The "fairness anchors the cartel" account of §4.2 comes from reading transcripts; removing fairness language from a free-text channel while leaving bandwidth intact would separate it from the compression account. The common-enemy reading of M_RUMOUR would be separated from an altered-monitoring-beliefs account by a privately-delivered warning, which we did not run.

A measurement error we made. An early keyword metric led us to conclude agents were "talking cartel while pricing competitively". Reading the messages refuted it — the lexicon was firing on competitive

sentences. We report the corrected analysis in §4.4 and flag it here because it is the error an automated transcript auditor makes by construction.

5.2 Future Work

The direct extensions are: the minimum competition-restoring dose; mid-episode rather than preventive intervention; seeded-cartel initial conditions, which test whether collusion is *sustainable* once established rather than whether it *forms*; multi-layer and subspace ablation; cross-model markets, where Keppo et al. (2026) report leader–follower dynamics under heterogeneity; and a wider price grid to find the model's own saturation rather than the grid's.

6. Conclusion

In this environment, every intuitive approach to detecting or preventing LLM collusion fails in a specific and measurable way. Keyword auditing over-counts coordination roughly twentyfold; price-responsiveness detectors rank the worst cartel as the least suspicious; restricting the communication channel nearly quadruples collusion; and deleting the internal feature that causes it does nothing at all. The two interventions that do work are a behavioural one requiring no model access — insert one agent with a different objective — and a white-box one that requires the right *operation*, pushing against the direction rather than removing it.

The mechanistic results argue that this is tractable rather than hopeless: the behaviour has a linear direction that is causally effective in the back half of the network, orientation-specific against both random and semantically matched placebos, gradedly dose-responsive, and sufficient when applied to the single token where the decision is formed. But the same precision is available to an attacker, and one compromised agent is enough to move a market it does not control.

Code and Data

- **Code repository:** <https://github.com/AJ797/Swayambhu>
- **Data:** all raw trajectories, including failed and unparseable turns, are in the repository under `data/<run-id>/`, together with the run logs and the extracted steering directions — so every steering and ablation result reruns without re-capturing activations. The activation tensors themselves (≈ 5 GB) are reproducible from the capture scripts.
- **Models:** Qwen/Qwen2.5- $\{14\text{B}, 32\text{B}, 72\text{B}\}$ -Instruct, bf16, A100-80GB ($\times 4$ for 72B), temperature 0.7.

Author Contributions

Ariet Jha conceived the project, gave the original specification, set the research direction, chose and prioritised the experiments, performed the literature review, and verified all results. Code and prose drafting were done with LLM assistance (see LLM Usage Statement).

References

Abridged to the works cited in the main text; the complete list is in the repository.

- Andreoli-Versbach, P., & Franck, J. (2015). Econometric evidence to target tacit collusion in oligopolistic markets. *Journal of Competition Law & Economics*, 11, 463–492.
- Arjona, D. R., Sitzia, S., & Zheng, J.-W. (2022). Overcoming coordination failure in games with focal points. *Games and Economic Behavior*, 136, 505–523.
- Awaya, Y., & Krishna, V. (2016). On communication and collusion. *American Economic Review*, 106, 285–315.
- Baranchuk, M., et al. (2024). Secret collusion among AI agents. *NeurIPS* 37.
- Bourreau, M., Sun, Y., & Verboven, F. (2021). Market entry, fighting brands, and tacit collusion. *American Economic Review*.
- Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- Chowdhury, S. M., & Crede, C. J. (2020). Post-cartel tacit collusion. *International Journal of Industrial Organization*.
- Cooper, D. J., & Kühn, K.-U. (2009). *Communication, renegotiation, and the scope for collusion*.
- Crawford, V. P., Gneezy, U., & Rottenstreich, Y. (2008). The power of focal points is limited. *American Economic Review*, 98, 1443–1458.
- Dal Bó, P., & Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games. *Journal of Economic Literature*, 56, 60–114.
- De Jaegher, K. (2020). Common-enemy effects. *Journal of Economic Surveys*.
- Di Cicco, C. (2026). Code correctness is linearly decodable from LLM hidden states before generation. *arXiv:2606.14530*.
- Dibadj, R. (2008). Conscious parallelism revisited. *San Diego Law Review*, 47, 589.
- Dugar, S., & Shahrir, Q. (2018). Restricted and free-form cheap-talk and the scope for efficient coordination. *Games and Economic Behavior*, 109, 294–310.
- Fadnavis, S., Kanakaraj, P., & Wyss, F. I. (2026). The readout-mediator angle in temporal reasoning. *arXiv:2605.29126*.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. *arXiv:2404.00806*.
- Fonseca, M. A., & Normann, H.-T. (2012). Explicit vs. tacit collusion. *European Economic Review*, 56, 1759–1772.
- Geiger, A., et al. (2023). Causal abstraction. *JMLR*, 26.
- Harrington, J., González, R. H., & Kujal, P. (2016). Price announcements and express communication for collusion. *JEBO*, 128, 251–264.
- Keppo, J., Li, Y., Tsoukalas, G., & Yuan, N. (2026). On the fragility of AI agent collusion. *arXiv:2603.20281*.
- Lekeas, P. V., & Stamatopoulos, G. (2026). How a cooperative-override circuit suppresses Nash play in LLMs. *arXiv:2604.27167*.
- Liao, C.-T., & Cao, X.-Z. (2026). The objective decides. *arXiv:2607.03728*.
- Mathew, Y., et al. (2024). Hidden in plain text. *arXiv:2410.03768*.
- Morsky, B., & Akçay, E. (2020). False beliefs can bootstrap cooperative communities through social norms. *Evolutionary Human Sciences*, 3.
- Motwani, S., et al. (2024). Secret collusion among generative AI agents. *arXiv:2402.07510*.
- Nakamura, M., et al. (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. *arXiv:2602.15198*.
- Rimsky, N., et al. (2024). Steering Llama 2 via contrastive activation addition. *ACL 2024*, 15504–15522.
- Rippin, J. L., et al. (2026). Tool use enables undetectable steganography. *arXiv:2606.28425*.
- Ruan, K., et al. (2026). Doomed from the start. *arXiv:2607.06503*.
- Sahoo, S., et al. (2026). Linear probes detect task format, not reasoning mode. *arXiv:2606.02907*.
- Sarkar, N., & Deka, D. J. (2026). Causal dimensionality of transformer representations. *arXiv:2605.08740*.
- Tailor, O. (2025). Audit the whisper. *arXiv:2510.04303*.

Tiwari, D., et al. (2026). Decodability is not causality. *arXiv:2609.18080*.

Valeriani, L., et al. (2023). The geometry of hidden representations of large transformer models. *arXiv:2302.00294*.

Vaska, M. (1985). Conscious parallelism and price fixing. *University of Chicago Law Review*, 52.

Yadav, A., Black, S., & Sourbut, O. (2026). More capable, less cooperative? *arXiv:2604.07821*.

Zou, A., et al. (2023). Representation engineering. *arXiv:2310.01405*.

LLM Usage Statement

Claude (Anthropic) was used throughout this project as a coding and drafting tool. It implemented most of the experimental code — the environment, the architecture-agnostic hook layer, the metrics, and the figure generation — and produced first drafts of this document, which the author revised over several iterations. The research direction, the experimental design and its prioritisation, the decisions about what to run and what to cut, and the literature review are the author's. The probe and steering pipelines, the controls and guards, and the representation-state analysis are the author's own work, in part adapted from his earlier published paper. All results were verified by the author, who is responsible for the content of this report.

Several corrections below came from controls and guards written into the pipeline before any results arrived. We list them because they are precisely the errors this methodology is designed to catch: a pooled probe AUC of 0.994 was voided by its own shuffled-label sham; an apparent +0.43 steering effect was traced to a fallback that manufactures $\Delta \approx 0.5$ when the model is broken; an ablation test was invalid because it ran too few rounds for the cartel to form; and an early keyword metric produced a false conclusion that reading the messages refuted. Every number reported here was regenerated after those fixes.

Appendix A — Full 14B factorial

16 episodes $\times$ 24 rounds $\times$ 3 firms per condition. Δ is the mean over the final 10 rounds. Message rates are over turns that sent a message.

condition	objective	channel	Δ	sd	parse fail	cartel lang.	compete lang.	reaction β	price match
C_COLLUDE	collude	off	+0.990	0.015	0.0%	—	—	+0.000	1.00
C_COMPETE	compete	off	−0.086	0.005	0.1%	—	—	+0.240	0.98
E_BASE	neutral	off	+0.009	0.038	0.1%	—	—	+0.172	0.74
E_CHAT	neutral	free	+0.006	0.038	0.0%	0.01	0.39	+0.353	0.87
E_ORG	neutral	free	+0.231	0.068	0.1%	0.61	0.02	+0.116	0.95
E_CANON	neutral	canon	−0.007	0.162	0.0%	0.06	0.86	+0.366	0.92
E_STRAT	neutral	off	+0.072	0.056	0.2%	—	—	+0.235	0.60
E_STRAT_CHAT	neutral	free	+0.023	0.053	0.0%	0.05	0.51	+0.229	0.82
M_PRESENT	neutral	free	−0.003	0.038	0.0%	0.01	0.46	+0.168	0.54
M_RUMOUR	neutral	free	+0.267	0.120	0.0%	0.00	0.18	+0.202	0.63
CC_COLLUDE	collude	free	+1.000	0.000	0.0%	0.94	0.00	+0.000	1.00
CC_COMPETE	compete	free	−0.086	0.005	0.0%	0.00	0.64	+0.133	0.96

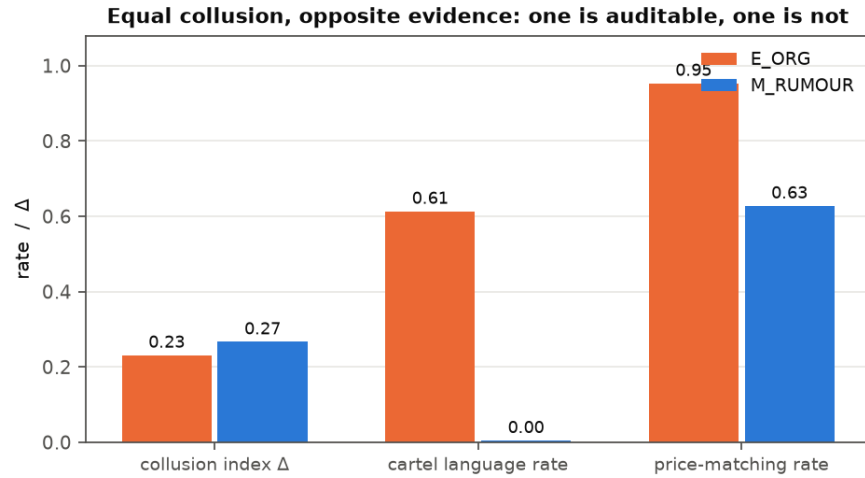


Figure A1. (Cited in §4.3.) Two conditions reaching comparable collusion by routes that leave opposite evidentiary traces. E_ORG is loud and lockstep; M_RUMOUR reaches higher collusion with no cartel language and no price convergence, evading transcript and parallelism auditing at once.

Appendix B — Steering dose grid and countermeasure dose-response

B1. Induction grid, 14B, condition E_BASE (unsteered $\Delta = 0.070$), all positions. Cells are Δ over all rounds; every cell had $\leq 0.7\%$ parse failure.

layer	$\alpha=-0.40$	$\alpha=-0.20$	$\alpha=-0.10$	$\alpha=+0.10$	$\alpha=+0.20$	$\alpha=+0.22$	$\alpha=+0.26$	$\alpha=+0.30$	$\alpha=+0.34$	$\alpha=+0.38$	$\alpha=+0.40$
L12	+0.169	+0.072	+0.086	+0.079	+0.127	—	—	—	—	—	+0.140
L16	+0.272	+0.105	+0.076	+0.038	+0.042	—	—	—	—	—	+0.172
L20	+0.249	+0.089	+0.092	+0.083	+0.009	—	—	—	—	—	+0.149
L24	+0.044	+0.048	+0.045	+0.115	+0.220	—	—	—	—	—	+0.244
L28	-0.029	+0.012	+0.044	+0.159	+0.238	—	—	—	—	—	+0.384
L32	-0.042	-0.002	+0.028	+0.210	+0.344	—	—	—	—	—	+0.393
L36	-0.008	+0.023	+0.039	+0.103	+0.258	+0.280	+0.341	+0.358	+0.497	+0.645	+0.768
L40	-0.020	+0.023	+0.036	+0.210	+0.209	+0.286	+0.358	+0.577	+0.830	+0.963	+0.998

B2. Countermeasure dose-response, 72B, condition E_CANON at L64.

cell	Δ all rounds	Δ final 10	parse fail
baseline	+0.409	+0.790	0.0%
L64_a-0.1	+0.396	+0.741	0.0%
L64_a-0.2	+0.228	+0.426	0.0%
L64_a-0.4	-0.091	-0.096	0.0%
L64_a-0.8	-0.100	-0.100	0.0%

B3. Steering controls and placebos, 14B, L40, $\alpha = 0.40$, against an unsteered Δ of 0.070.

variant	effect vs baseline
real direction	+0.928
matched-norm random	+0.063
dimension-shuffled	+0.004
placebo_coop — same cooperate/compete framing, about filing reports	+0.089
placebo_neutral — coastal vs mountain office	+0.071

B4. Single-agent steering, Δ over the final 8 rounds.

	14B baseline	14B $\alpha=0.6$	72B baseline	72B $\alpha=0.4$	72B $\alpha=0.6$
steered firm	-0.003	+1.014	+0.123	+0.709	+1.008
untouched pair	+0.012	+0.201	+0.097	+0.379	+0.223

Appendix C — Within-condition detection AUC

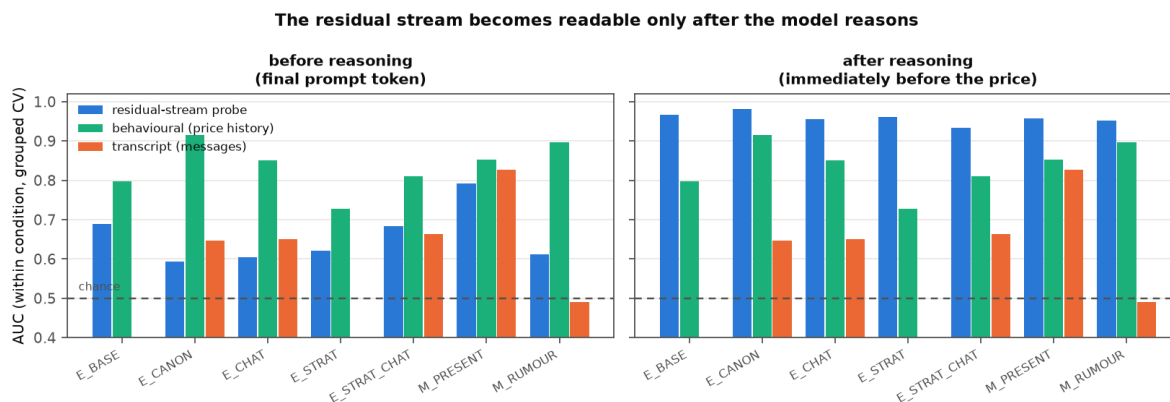


Figure A2. (Cited in §4.5.) Within-condition detection AUC at two capture points, 14B. Before the model reasons the behavioural baseline (green) wins in every condition; after reasoning the residual-stream probe (blue) wins in every condition. Transcript auditing (orange) is at chance in M_RUMOUR, where collusion is silent.

Grouped cross-validation by episode. Conditions with fewer than 20 of either class are omitted, since AUC on a handful of positives is noise. Instructed conditions are single-class by construction and therefore absent.

14B, capture point pre_reasoning — sham (global) 0.500, sham (within-episode) 0.918, transfer instructed → emergent 0.574

condition	positive rate	probe	behavioural	transcript
E_BASE	0.31	0.690	0.798	n/a
E_CANON	0.09	0.594	0.916	0.647
E_CHAT	0.25	0.604	0.850	0.651
E_STRAT	0.50	0.621	0.728	n/a
E_STRAT_CHAT	0.38	0.684	0.811	0.662
M_PRESENT	0.24	0.792	0.852	0.827
M_RUMOUR	0.86	0.613	0.896	0.491

14B, capture point pre_price — sham (global) 0.503, sham (within-episode) 0.919, transfer instructed → emergent 0.842

condition	positive rate	probe	behavioural	transcript
E_BASE	0.31	0.967	0.798	n/a
E_CANON	0.09	0.981	0.916	0.647
E_CHAT	0.25	0.955	0.850	0.651
E_STRAT	0.50	0.960	0.728	n/a
E_STRAT_CHAT	0.38	0.934	0.811	0.662
M_PRESENT	0.24	0.957	0.852	0.827
M_RUMOUR	0.86	0.952	0.896	0.491

72B, capture point pre_reasoning — sham (global) 0.500, sham (within-episode) 0.907, transfer instructed → emergent 0.764

condition	positive rate	probe	behavioural	transcript
E_CANON	0.68	0.972	0.983	0.801
E_CHAT	0.92	0.778	0.959	0.506
M_PRESENT	0.26	0.764	0.856	0.523

72B, capture point pre_price — sham (global) 0.505, sham (within-episode) 0.916, transfer instructed → emergent 0.980

condition	positive rate	probe	behavioural	transcript
E_CANON	0.68	0.999	0.983	0.801
E_CHAT	0.92	0.999	0.959	0.506
M_PRESENT	0.26	0.995	0.856	0.523

Appendix D — Contrast separability, geometry, and keyword over-count

D1. Direction separability by depth (Cohen's d). Surface textual contrasts separate early and wash out; the task-relevant contrast does the opposite.

contrast	L0	L40	trend
partner-trust	2.62	7.24	increases
p_placebo_coop (cooperate/compete, about reports)	5.92	1.93	decreases
p_placebo_neutral (coastal vs mountain office)	2.67	1.78	decreases

D2. Effective dimensionality (participation ratio of the state covariance), 14B.

layer	L0	L4	L8	L12	L16	L20	L24	L28	L32	L36	L40	L44
eff. dim	19.1	11.6	16.8	15.1	16.3	13.4	9.6	11.5	9.8	8.5	8.3	7.6

D3. Collusion vs partner-trust direction. Cosine similarity ranges -0.018 to 0.118 across captured layers (83–85°), against a chance level of $1/\sqrt{5120} \approx 0.014$.

D4. Keyword over-count. A naive coordination lexicon against direction-aware counts.

condition	naive keyword flag	true cartel language	competitive language	over-count
pilot/E_CHAT	0.48	0.00	0.45	—
duo/E_CHAT	0.51	0.03	0.26	17.9×
duo/E_STRAT_CHAT	0.42	0.02	0.43	25.2×

D5. Environment parameters. $q_i = \exp((a - p_i)/\mu) / (\sum_j \exp((a - p_j)/\mu) + \exp(a_0/\mu))$, $\pi_i = (p_i - c) \cdot q_i$, with $a = 2$, $a_0 = 0$, $\mu = 0.25$, $c = 1$. The 13-point grid spans $[p_N - 0.1R, p_M + 0.1R]$ where $R = p_M - p_N$. At $n = 2$ this reproduces Calvano et al.'s published constants ($p_N = 1.4729$, $p_M = 1.9250$).

Appendix E — Dual-use and ethical considerations

Required appendix. The methodological limitations are in §5; this section covers what the results do and do not establish, and how they could be misused.

What this establishes. That three Qwen2.5-Instruct models, in one environment (logit-demand Bertrand, 2–3 firms, 24 rounds, a 13-point grid), under the listed prompt variations, show increasing emergent collusion with scale; that a four-symbol channel produced more collusion than free text at 72B across two independent runs; that one instructed maverick suppressed collusion at both scales tested; that a linear direction in the residual stream causally controls the price level, orientation-specifically and with competence preserved; and that ablating that direction does not break an established cartel while negative steering along it does.

What it does not establish. Generalisation beyond one model family, one market structure, or price settings. The scale ladder is three points and confounds size with post-training. The bandwidth-backfire result rests on 8 episodes per run at one scale, replicated once; it is the single most important thing for others to reproduce. The ablation null and the negative-steering positive are each one condition at one scale with one geometry. The mechanism accounts in §4.2 and §4.3 are inferred from transcripts and condition design, not isolated experimentally.

Dual-use assessment. This work is dual-use in three specific ways, which we state rather than imply.

1. **We show that a particular channel restriction increases collusion.** That is usable as a recipe for making agent markets more collusive. We judge publication net-positive because the primary audience is defenders who would otherwise deploy exactly this restriction believing it protective — the covert-coordination literature actively recommends it — and because the interventions that *did* work are reported alongside.
2. **We show that one compromised agent moves a market it does not control, and that a moderate compromise is more effective than a maximal one.** This is the closest thing here to an attack recipe, and the interior-optimum finding in particular tells an attacker how to tune. We report it because the corresponding defensive fact is more important: the honest agents' behaviour is *not* evidence that they are uncompromised, and a defender who only inspects conspicuously-priced agents will miss the effective attack. We do not report a configuration optimised for market capture.
3. **The induction result is precise.** Mitigating this, the steering attacks require white-box access to model internals, and an attacker with that level of access has stronger and simpler options (fine-tuning, system-prompt replacement). The result is more informative about monitoring than about attack.

We report competence checks alongside every intervention so that effects achieved by degrading a model are not mistaken for induced coordination, and we discard intervention cells that break output formatting rather than reporting the Δ they produce.

On the keyword result. Our finding that lexicon-based auditing over-counts coordination 18–25× could help an adversary evade a keyword monitor. We consider this outweighed by the value of warning defenders off a method that does not work, and note that the method's unreliability is discoverable by anyone who inspects their own detector's positives.

Not a claim about real markets. Nothing here is evidence about deployed pricing systems, nor competition-policy advice. The environment is a simulation chosen for the tractability of its ground truth, not for its realism.