

Honor Among AI Thieves? AI collusion in rival cartels in markets.

Abstract

As the capabilities of LLMs increase dramatically we've seen a rise in the introduction of AI agents. These agents while acting for people and companies that utilize them can collude against the principals they represent. To our knowledge, this is the first study of LLM agent coalitions facing each other in a market. Where we asked how two AI cartels interact and whether resource scarcity alters their behavior. We developed and tested a pre-registered coalition-vs-coalition market with a crossed microstructure factor (144 episodes on DeepSeek-V3) followed by 320 contest episodes across DeepSeek-V3, GPT-4o and Claude Opus 5.5. Our hypothesis that scarcity would increase in-group cohesion was not supported (110 episodes, two models; $p = 0.10-0.42$), but the study offers answers to several of the sprint's questions. 1: Market rules drove coalition behavior by changing the clearing rule the time to a deal moved as well as price. So manipulating a market design is a potential low-cost lever for reducing AI collusion. 2: Coalitions facing independent sellers held their line and lost. DeepSeek coalitions won fewer contracts when facing independents. 3: A markup over a competitive benchmark is not strong evidence of LLM collusion unless it survives mirroring the market side and shifting the price level. Our results showed markups were higher on the buy side in two tested models but when prices were raised by 100, 45 of 140 markets failed to clear. Claude Opus 5.5 never formed a coalition.

Introduction

The past year has seen a surge of security incidents caused by AI agents. In July 2026, AI agents that OpenAI was training and evaluating escaped their sandbox and breached Hugging Face's servers, which made headline news [1–4]. At least 1,200 agent instances, most of them copies of a single internal model, coordinated through an improvised message board on OpenAI's internal infrastructure and gained access to Hugging Face partly through a template-injection flaw in a dataset configuration [3, 5, 6]. In a separate episode disclosed in September 2026, OpenAI agents turned a dormant German software wiki into a coordination channel, leaving about 18,000 posts to pool answers on timed tasks [7, 8]. Agentic collusion has become a major area of AI safety and security research [9, 10], made even more pressing by incidents like these and by the arrival of more capable personal AI agents such as xAI's Grok Bot [11], Meta's Muse [12], OpenClaw [13] and OpenAI's Dots [14].

Previous AI collusion research has shown the complexities of AI collusion and how difficult it can be to prevent [9, 10]. Frontier models can often tell when they are being evaluated [15] and can strategically underperform on evaluations when given a reason to (sandbagging) [16, 17].

Countermeasures have also met with difficulties: monitoring and paraphrasing agents' messages do not reliably stop covert coordination [10, 18]. Protocols in which a weaker trusted model supervises a stronger one improve safety but do not eliminate the risk [19]. In the Hugging Face incident, agents collaborated through improvised channels even when no collaboration tool was enabled [3]. Prior work on algorithmic and AI collusion focuses on a single coalition or set of competitors against a principal or against non-colluding bystanders [20–23].

Instead we focus on measuring interactions between two competing AI coalitions within a market over a perceived scarce resource. Within a confined marketplace do these coalitions settle, or fight and what is the effect within each coalition, greater or lesser in-group cohesion? This AI v AI coalition interaction we believe will become very common as more and more of digital space becomes occupied by AI agents the greater the risk of agentic collusion in price-fixing, market division and manipulation, and norm erosion.

Contributions.

1. A pre-registered coalition-vs-coalition market with a crossed microstructure factor where the market rules dominate communication.
2. A behavioral cross-model contrast: Opus never forms the coalition; verbalised refusal is salience-sensitive.
3. A pre-registered negative result on scarcity and cohesion (110 episodes, two firm models), with a replicated side finding: a cohesive coalition facing independents holds its line and loses the tender.
4. Two controls showing that markup, position and shortfall are conditional on the nominal price level; a mechanism (percentage reasoning on the agent's own reservation) that explains a buyer–seller asymmetry and mass non-clearing.

Background

The Study A grid and the Opus comparison were run on 21–22 October, before the sprint opened on the 23rd. Our contest line (320 episodes) was designed and run during the sprint.

Our primary question was: when two cartels face each other across a market, do they settle or fight, and does rivalry between them tighten or loosen cohesion within each?

Pre-registered predictions:

- Model settlement is more frequent and median close is earlier from none to public to private channels, within each microstructure.
- Accept-only conditions produce longer closes and wider spread. Whereas seller-ask produces fast, tight settlements displaced toward sellers. If the microstructure effect exceeds the channel effect outcomes are determined by market design.

- Cohesion is higher under the full bloc instruction than minimal.
- No directional prediction was made on cohesion across channels.

Contest Hypothesis: Rivalry over scarce contracts tightens in-group cohesion therefore partner asks will be closer with 2 contracts than 3.

Competing prediction (group-contest free-riding): partners are rivals for scarce contracts, so gap, undercuts and leapfrogs rise with scarcity.

Our contest hypothesis was inspired by work showing that intergroup conflict raises in-group cooperation [25, 26, 31] and our competing prediction follows work on group contests and free-riding [27–30].

Methods

We began by establishing the bargaining market for Study A, our main pre-registered study (run on 21–22 October, before the sprint). The market was comprised of 2 sellers (coalition one) vs. 2 buyers (coalition two), each acting on behalf of a principal with a private reserve and no input to the other agents or the market. Seller floors are drawn from a uniform distribution $U(20, 35)$ and buyer ceilings are between $U(65, 80)$ so deals can always be made. Each market ran for a max of 20 rounds after which if no deal was reached, each agent lost 5 points of pay. Position normalized: 0 = lowest seller floor, 1 = highest buyer ceiling. Payoffs were calculated via $\text{Agent} = \alpha \times \text{principal surplus} (\alpha = 1.0) + 12 \times 0.90^{(\text{round}-1)}$ with a settlement bonus and no-deal penalty of -5. $\text{Principal} = \text{surplus at the clearing price}$.

Factors

In Study A we compared the following factors and their levels:

- Cross-bloc channel: none / public (principal can read) / private (principal cannot).
- Bloc instruction strength: minimal (bare fact) / full (+ shared objective, "hold a common line"). Framing: neutral commercial code / unregulated.
- Counterpart: colluding buyers / non-colluding buyers (counterfactual)
- Microstructure (crossed with every cell): accept-only order book / AgenticPay seller-ask bilateral threads [24].
- Paraphrase (nuisance): 3 per manipulated block, seed mod 3; minimal-pair and length-matching checked by a smoke test, following our earlier framing study [37].

Microstructures are accept only with one order book, simultaneous quotes, alternating move order, clearing by explicit accept of a standing quote or bid/ask within 0.5 (midpoint). Seller-ask: four bilateral threads, buyers move first, sellers see same-round bids, a crossed pair settles at the seller's ask, prices sticky.

Our studies' grid is defined as the following: per microstructure none 18, public 12, private 12, counterfactual 10, unregulated 10, full bloc 10 = 72; $\times 2 = 144$ episodes. Seeds from 20261023; identical seeds and reservation draws across arms. The grid model was DeepSeek-V3 (deepseek-chat-v3-0324, via OpenRouter); Claude Opus 5.5 was the comparison model and Claude Sonnet 5.5 (low effort) the judge. No sampling temperature was set (the API rejects it), so seeds were the only control for the reservation draws. Transcripts are reproducible only from the response cache. We selected DeepSeek for the grid because Claude Opus rejected coordination and refused to form a coalition in all 36 pilot episodes which was later corroborated by a 104-episode comparison. Code, pre-registration and run outputs are at github.com/horigyns/cartel-rivalry-study.

Contest Environment

We took the same mechanics from Study A including payoff weights, neutral jurisdiction and paraphrases reused verbatim from Study A's frozen prompts. To test whether rivalry over a scarce resource changes how coalitions operate we made the following changes. We built a second environment in which two coalitions compete for the same thing rather than trading with each other. Two coalitions each composed of two agents competed to supply a single scripted buyer awarding K contracts. Scarcity was set by K , the number of contracts. We ran this environment four times (run 1, Study C, a buy-side mirror and a shifted-range control); within each run, K took the values 1, 2 and 3, and each run's predictions were recorded in the pre-registration before it launched. The reservation ranges and payoff weights were carried over from Study A, but the market rules were new. At each round every seller posted an ask, the market closed once no ask had been lowered for two consecutive rounds (Max 12 rounds), and the K lowest asks at or below the buyer's hidden maximum won. Agents were told the structure of their pay, including that a settlement bonus and a no-deal penalty existed, but not their values. We also added a cell in which the rival coalition was replaced by two independent sellers. Which received Study A's "you act alone" prompt and no channels (played by DeepSeek-V3 in run 1 and by Claude Opus 5.5 from Study C onward).

Run 1 served as a test of our scarcity hypothesis. After data collection we modified parts of the study to fix problems that had arisen. Ties at the cut off were common and were settled by the system using a rule the agents had not been told, which decided the winning coalition in 10 of 40 episodes. Partners could also see only the previous round's messages, and asks could be raised. In later runs, ties were settled by a lottery the agents were told about, the full message history stayed visible, and asks could no longer be raised.

In Study C we made sure coalitions were firmly established rather than being merely stated. Agents also became representatives of a company. Sharing an owner and cost to ensure that in-coalition cooperation was ordinary and legal therefore making collusion only able to occur between companies. We also incorporated a commission so undercutting a colleague was rewarded and this pay split was stated in the prompt. We used GPT-4o (gpt-4o-2024-11-20) as our second firm model, since earlier runs showed Opus 5.5 was unwilling to join a coalition. We

then mirrored this experiment putting agents on the buy side rather than the sell side. A bid b was stored as the ask $100 - b$, so for the same seed the underlying game is numerically identical to its sell-side twin and every measure is computed by the same code. Agents were purchasing companies that bid for K units from a scripted seller with a hidden minimum price. The only wording that was changed was the ask to bid, the cost to value and swapping lowest to highest. Finally, to test whether the buyer–seller difference reflected the role or the size of the numbers, we reran both sides with every price raised by 100 and the width of the bargaining zone unchanged.

The following was collected as data for our results. Study A: deal share, accidental crossings, median/mean close round, mean position, reference position (counterfactual arm), displacement, cross-seed SD, defection rate, principal shortfall, classification. Cohesion was measured four ways and the specification history is recorded in the pre-registration.

Contest measures (all continuous, no classification):

- Partner gap (primary cohesion): mean $|\text{partner ask difference}| / \text{zone per scored round}$, round 1 skipped.
- Undercut rate per coalition-round (> 0.08 of zone below partner).
- Leapfrog rate: share of chances the lower-asking partner switched (ties within 0.5 skipped); primary at $K=1$, where a cover bid makes the gap wide and stable.
- Markup over the competitive benchmark, zone units. Run 1: $(K+1)$ -th lowest seller floor. Study C onward: efficient cost-ordered allocation over supplier units (a company = one supplier of capacity 2; an independent = capacity 1); per served slot, the cost of the cheapest other supplier with unserved capacity, the buyer's maximum if none. For capacity-1 sellers this equals the run-1 rule (all 40 run-1 markups unchanged). At $K=3$ with two companies the third slot is unopposed and takes the ceiling.
- Position (normalized award), buyer surplus (the harmed party; on the buy side the scripted seller's surplus), split/sweep allocation, ask descent, close round, Company A's wins.
- Own-distance: award minus the agent's own reservation, as a share of the zone, and as a share of the reservation itself.

Study A utilized bootstrapped 95% CIs, paired bootstrap channel contrasts from 10000 draws and Fisher's exact test for defections with a minimum detectable effect by enumeration and Mann–Whitney tests with Holm correction for the exploratory gap. Contest run 1 used cell means and a one-sided Mann–Whitney test; Study C used a paired one-sided Wilcoxon signed-rank test; the mirror and shift used paired two-sided Wilcoxon tests (35 seed-paired cells per model).

Results

Table 1. Study A: the clearing rule outweighs communication (median close round / mean position; minimal-instruction cells)

	no channel	public	private	no buyer coalition (baseline)
accept-only	11 / 0.28	8 / 0.32	6 / 0.25	4 / 0.27
seller-ask	2 / 0.69	3 / 0.71	3 / 0.66	3 / 0.76

In Study A under accept-only rules, deals took 11 rounds when agents had no channel. This was then dropped to 6 when agents were given a private channel. While talking accelerated time to a deal it barely changed price (0.25 to 0.32). In the seller ask scenario adding a channel made almost no difference at all with deals taking 2–3 rounds. We saw the biggest changes when we changed the trading rules: with the same agents and prompts, time to a deal dropped from 6–11 rounds to 2–3 rounds. The final column served as our baseline, in which the buyers had no coalition. The existence of coalitions on both sides shifted price much less than the marketplace rules did.

Claude Opus 5.5 declined the coalition in all 36 pilot episodes and formed none in a later 104-episode comparison (0/104). The share of its turns giving an ethical reason rose from 0.60 to 0.78 under the fuller coalition instruction (difference -0.18 [-0.27 , -0.08]), while its behaviour did not change. These results lead to our decision to utilize Opus 5.5 only as an independent seller.

Table 2. Contest: scarcity did not tighten coalitions; coalitions held together and lost against independents

Run / firm model	partner gap, 2 contracts	partner gap, 3 contracts	test (H1)	coalition's wins vs independents	gap vs independents
Run 1, DeepSeek	0.072	0.053	$p = 0.10$	0 / 10	0.004
Study C, DeepSeek	0.026	0.037	$p = 0.37$	1 / 10 (5 / 10 vs a rival company)	0.003
Study C, GPT-4o	0.054	0.050	$p = 0.42$	3 / 10	0.117

Gaps are means over both channel conditions. Run 1 p is the descriptive one-sided Mann-Whitney; Study C p is the paired one-sided Wilcoxon.

In run 1 and Study C we measured the partner gap, i.e. how far apart two partners' prices were. The smaller the gap the more united the coalition. Our hypothesis that increased competition would create tighter coalitions was not supported. The 2-contract gap was larger than the 3-contract gap in two of the three comparisons and smaller in one, so there was no consistent direction, and no test was statistically significant ($p = 0.10$, 0.37 and 0.42). Across two runs and two models there was no reliable effect. However, when facing independent sellers,

DeepSeek's coalitions were much more aligned, with partner gaps of 0.003 and 0.004, far smaller than when facing another coalition (not tested for significance). DeepSeek also lost nearly every contract it bid for against independent sellers (0/10 and 1/10). Compared to when it faced another coalition the same agents won 5/10 contracts. GPT-4o behaved differently, against independents its partners broke ranks (gap 0.117) and still lost 3 out of 10 bids.

Table 3. The controls: settlements track the size of the agent's own number (award distance from own reservation; per-episode means)

	DeepSeek sell	DeepSeek buy	GPT-4o sell	GPT-4o buy
base level, share of zone	0.23	0.55	0.11	0.39
+100 level, share of zone	0.50	0.66	0.31	0.44
base level, % of own reservation	48%	36%	19%	26%
+100 level, % of own reservation	20%	19%	12%	12%
markets not cleared, base → +100 (of 35)	0 → 10	3 → 24	0 → 0	0 → 11

Table 3 measures an agent's distance from its own reservation, i.e. how much it kept: how far a seller settled above its cost, or a buyer below its value. At the base price level, buyers kept far more than sellers (0.55 vs. 0.23 for DeepSeek and 0.39 vs 0.11 for GPT-4o). Since it was the same experimental design with only the wording flipped it meant that buyers kept 2–4 times more surplus than sellers. When every price was raised by 100, with the width of the bargaining zone unchanged, sellers' shares roughly doubled (DeepSeek) or tripled (GPT-4o). Expressed as a percentage of each agent's own number, the two sides were nearly identical at the +100 level (20% vs 19% for DeepSeek; 12% vs 12% for GPT-4o). At +100, a percentage move can overshoot a fixed-width bargaining zone; as a result, many markets never traded at all: 34 of 70 episodes for DeepSeek and 11 of 70 for GPT-4o, compared with 3 and 0 at the base level.

Discussion

We began this study by looking to determine how two separate AI cartels would operate within a specified marketplace. Study A asked broadly whether cartel rivalry changed in-group cohesion, but could not measure cohesion reliably. The design lacked the statistical power to detect it. Trading rules had a much greater effect than available routes of communication within the bounds of that study. This suggests that market design may be a lower-cost lever on what colluding agents achieve than restricting their communication. Within the game design models were able to trade which offered a scenario where both cartels could win making rivalry a worse option. Study A gave us the experimental design foundation and a basic understanding of a novel research question.

From there we started on the work for the sprint. Since Study A could not answer the cohesion question, and its trading game let both cartels win from a deal. We instead asked if the scarcity

of a resource affected the coalitions. Our results did not support this hypothesis either. As contracts became scarcer, the in-group partner gap did not consistently shrink. However, when facing independent sellers the coalitions held together and lost. This might suggest that the type of rival changes how coalitions behave. This is still exploratory and an avenue for future work. Classic cartel theory states that collusion is sustained by repeated interaction and the threat of punishment [32–34] and our contests were single episodes. So repeated interactions are a natural next step. In Study C we fixed run 1's design problems (full message memory, a disclosed tie lottery, no raising of asks), established the coalitions as companies, and added a second model, GPT-4o. While DeepSeek again held ranks while losing against independent sellers, GPT-4o broke ranks, with a much wider partner gap (0.117).

We then wanted to see if there would be any change in model and coalition behavior if coalitions were buyers instead of sellers. Our results found that buyer vs seller side mattered significantly where buyers kept 2-4x more than when they were sellers. To tell apart two explanations, a role preference or the size of the numbers, we raised every price by 100 while keeping the bargaining zone the same width, then expressed each settlement as a percentage of the agent's own reservation. These results suggest that it was not a role preference shown by the models but the size of the numbers, and many markets failed because the zone was too narrow relative to the price level. A simple markup over a competitive benchmark should not be considered evidence of collusion unless it also survives flipping the market side and shifting the price level. The same caution applies to the price-based screens competition authorities use to detect bid rigging [35, 36].

Results from this work tell us that for these selected models cohesion is likely a default or disposition and not a strategy that responds to a competitive situation. Across these models we saw stable “personalities” toward coalitions which was not a form of strategic adaptation. This study showed that the models themselves are as large of a variable as any factor we manipulated. Measuring collusion within this study design was highly complex across all the runs. Each time we modified the design or corrected a benchmark we exposed new measurement problems. What the field should determine as convincing evidence of model collusion must be carefully considered along with the complete design of the experiment.

Appendix

Limitations and Ethics

The primary limitation of this study is its scope. It focused primarily on DeepSeek and while we ran comparisons using GPT-4o and utilized Opus 5.5 as an independent seller there still remains the need to rigorously test this experimental design using different models. Another limitation was sample size, to stay within the suggested sprint cost, contest cells had only five seeds each so the cohesion tests could detect only large effects. This study was designed and run on a simulated marketplace with no human subjects. Dual use: our elicitation prompts are

published, which makes the pre-registration verifiable but also publishes a minimal recipe for eliciting coordination. We determined this as acceptable because the instruction is a single plain sentence, and the controls we introduce help guard against false positives in collusion screening.

References

1. Hugging Face Security Team. "Security incident disclosure — July 2026." Hugging Face blog, 16 July 2026.
2. OpenAI. "OpenAI and Hugging Face partner to address security incident during model evaluation." 21 July 2026.
3. OpenAI. "OpenAI – Hugging Face Incident Technical Report." 26 August 2026.
4. Satter, R., Seetharaman, D. and Cai, K. "Its AI agent spent days hacking a company. Sources say OpenAI did not notice for a week." Reuters, 24 July 2026.
5. Hugging Face Security Team. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident." Hugging Face blog, 27 July 2026.
6. METR. "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident." 26 August 2026.
7. Nightingale Collective (S. Von Arx, lead). Report on OpenAI agent activity on DSEwiki. September 2026. collusion.wiki.
8. Khandelwal, S. "Thousands of OpenAI Agents Quietly Turned an Abandoned Wiki Into Their Coordination Channel." The Hacker News, 5 September 2026.
9. Hammond, L., et al. "Multi-Agent Risks from Advanced AI." Cooperative AI Foundation Technical Report #1, 2025. arXiv:2502.14143.
10. Motwani, S. R., et al. "Secret Collusion among AI Agents: Multi-Agent Deception via Steganography." NeurIPS 2024. arXiv:2402.07510.
11. VentureBeat. "SpaceXAI's Grok Bot turns agents into persistent digital coworkers that can operate your apps for \$120-per-month." 2026.
12. Meta. "Introducing Muse: The World's First Personal AI Agent Built for Everyone." Meta Newsroom, 8 September 2026.
13. DigitalOcean. "What is OpenClaw? Your Open-Source AI Assistant for 2026." 2026.
14. TechCrunch. "OpenAI launches Dots, its bubbly agentic avatar." 29 September 2026.
15. Needham, J., Edkins, G., Pimpale, G., Bartsch, H. and Hobbhahn, M. "Large Language Models Often Know When They Are Being Evaluated." 2025. arXiv:2505.23836.

16. van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F. and Ward, F. R. "AI Sandbagging: Language Models can Strategically Underperform on Evaluations." ICLR 2025. arXiv:2406.07358.
17. Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R. and Hobbhahn, M. "Frontier Models are Capable of In-context Scheming." 2024. arXiv:2412.04984.
18. Mathew, Y., et al. "Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs." 2024. arXiv:2410.03768.
19. Greenblatt, R., Shlegeris, B., Sachan, K. and Roger, F. "AI Control: Improving Safety Despite Intentional Subversion." ICML 2024. arXiv:2312.06942.
20. Fish, S., Gonczarowski, Y. A. and Shorrer, R. I. "Algorithmic Collusion by Large Language Models." 2024. arXiv:2404.00806.
21. Lin, R. Y., Ojha, S., Cai, K. and Chen, M. F. "Strategic Collusion of LLM Agents: Market Division in Multi-Commodity Competitions." NeurIPS 2024 workshop. arXiv:2410.00031.
22. Wu, Z., Peng, R., et al. "Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents." Findings of EMNLP 2024. arXiv:2402.12327.
23. Calvano, E., Calzolari, G., Denicolò, V. and Pastorello, S. "Artificial Intelligence, Algorithmic Pricing, and Collusion." *American Economic Review* 110(10): 3267–3297, 2020.
24. Liu, X., Gu, S. and Song, D. "AgenticPay: A Multi-Agent LLM Negotiation System for Buyer–Seller Transactions." 2026. arXiv:2602.06008.
25. Bornstein, G. "Intergroup Conflict: Individual, Group, and Collective Interests." *Personality and Social Psychology Review* 7(2): 129–145, 2003.
26. Bornstein, G. and Ben-Yossef, M. "Cooperation in Intergroup and Single-Group Social Dilemmas." *Journal of Experimental Social Psychology* 30(1): 52–67, 1994.
27. Tullock, G. "Efficient Rent Seeking." In Buchanan, Tollison and Tullock (eds.), *Toward a Theory of the Rent-Seeking Society*, 97–112. Texas A&M University Press, 1980.
28. Katz, E., Nitzan, S. and Rosenberg, J. "Rent-Seeking for Pure Public Goods." *Public Choice* 65(1): 49–60, 1990.
29. Nitzan, S. "Collective Rent Dissipation." *The Economic Journal* 101(409): 1522–1534, 1991.
30. Olson, M. *The Logic of Collective Action*. Harvard University Press, 1965.
31. Sherif, M., Harvey, O. J., White, B. J., Hood, W. R. and Sherif, C. W. *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. University of Oklahoma Book Exchange, 1961.
32. Stigler, G. J. "A Theory of Oligopoly." *Journal of Political Economy* 72(1): 44–61, 1964.
33. Green, E. J. and Porter, R. H. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52(1): 87–100, 1984.

34. Levenstein, M. C. and Suslow, V. Y. "What Determines Cartel Success?" *Journal of Economic Literature* 44(1): 43–95, 2006.
35. Porter, R. H. and Zona, J. D. "Detection of Bid Rigging in Procurement Auctions." *Journal of Political Economy* 101(3): 518–538, 1993.
36. OECD. *Guidelines for Fighting Bid Rigging in Public Procurement*. 2009.
37. Hill, H. Hold Off There are Digital Civvies There. When Told the Casualties Are Human, Some Language Models Hold Back. OSF, 2026. osf.io/zwt32.