
Managing Collusion in AI Agents

William Yuen
UPay

Ambrose Chan
BBVA

Kit
Artfarmers

Andy Chan
Quant

Michael Choi
CCCKLC

With
Apart Research

25th October 2026

Abstract

Abstract. Autonomous agents can sustain supra-competitive outcomes without explicit agreement. We ask which variables are necessary for collusion, whether information-theoretic statistics can detect it, and which single restriction removes the most collusion at the least cost. In a Calvano-style Bertrand market with independent Q-learners, full observation of rival prices is sufficient for stable positive markups in two-agent markets (up to $+0.22$). Sequential timing, cheap talk, and side-payments are not necessary once that information is present. With three agents, a compact summary of rival prices restores collusion. A pre-registered comparison separates absolute suppression from cost-adjusted performance: anti-collusion penalties remove the most markup; hiding prices or small price noise is the least-cost way to return average markup near Nash. Residual mutual information does not detect collusion on learned paths (AUROC 0.35-0.65). In a complementary marketplace laboratory, an elicitation ladder produces capture only at an explicit hint, and a trusted-agent pilot shows that speech and action can dissociate. Both rankings, and the negative detection result, are reported.

1. Introduction

Autonomous pricing algorithms can learn to sustain prices above the static Nash level without explicit communication (Calvano et al., 2020). LLM-based agents can reach similar outcomes and can coordinate through channels that supervisors do not see (Fish et al., 2026). The failure mode is an information and objective design problem: what agents can observe, and what they are rewarded for, determines whether supra-competitive outcomes emerge. Agents need not be malicious.

The defender’s problem is harder than the economist’s. An auditor must know which variables make collusion possible, whether coordination is carried by speech or by action, and which restriction removes the most collusion at the least cost. This work is practically valuable because the lowest-cost first step - hiding rival prices, or adding a small amount of price noise - is available to any platform, including those without a detection team.

We study this in two laboratories. Track A uses tabular Q-learning in a repeated Bertrand market. Track B uses LLM agents in a marketplace with a covert kickback. Our main contributions are:

- Rival-price information is an enabling condition for collusion under tabular Q-learning, and a compact summary of that information is sufficient in multi-agent markets. Agent objectives are a second lever.
- A pre-registered six-condition comparison that separates absolute suppression from cost-adjusted ranking, with the cost scale stated explicitly. Small price noise is a low-cost alternative to hiding prices.
- A negative detection result (residual mutual information fails on learned paths) and, in a marketplace laboratory, a demonstration that ledger and chat have complementary blind spots and that speech and action can dissociate.

2. Related Work

Calvano et al. (2020) show that independent Q-learners sustain supra-competitive prices without communicating. Klein (2021) extends the setting to sequential moves. Fish et al. (2026) and Lin et al. (2025) show that LLM agents can reach supra-competitive outcomes and can divide markets. Harrington (2018) discusses the competition-law implications.

On detection, Xu et al. (2026) show that price-level audits can be blind when collusion preserves competitive marginals. Lee and Park (2026) show that chain-of-thought monitoring can fail while the policy remains collusive. On countermeasures, Collina et al. (2025) study how introducing humans can break algorithmic collusion.

Someone would use our approach over a single-environment demonstration when they need to know which factor is necessary and which restriction is cheapest. The new information is the separation of absolute suppression from cost-adjusted ranking, and the negative result on residual mutual information. Track B adds the say-versus-do distinction in a vertical market, where a transfer can occur without agreement and an agreement can occur without a transfer.

3. Methods

3.1 Track A - Bertrand Q-learning:

Firms compete in a repeated Bertrand game with logit demand ($\mu = 0.25$, $a_0 = 2$, $a = 1$, $c = 1$), following the Calvano et al. (2020) design. Each firm selects a price from a discrete grid. The outcome is the markup: mean transaction price minus the static Nash price. Each firm is an independent tabular Q-learner ($\alpha = 0.15$, $\gamma = 0.95$) with ϵ -greedy exploration decaying at 0.999995 per episode. Agents do not share parameters.

The pre-registered comparison uses two agents, eleven prices, 500,000 episodes, and seeds 0-4. Conditions differ from a full-observation baseline by a single change: concealment of rival prices; a fine at rate 0.3; escrow at rate 0.2; an anti-collusion reward penalty on one agent; and the same penalty on both. An additional comparison adds price noise ($\sigma = 0.05$ and 0.10) and demand noise ($\sigma = 0.05$ and 0.10). Cost is a 1-3 implementation score: 1 = change the information set; 2 = add monitoring or change one objective; 3 = change every objective. For detection, runs with terminal markup $\geq +0.10$ were labelled collusive (32); runs $\leq +0.02$ were labelled competitive (57).

3.2 Track B - Marketplace with covert transfers:

Each round a ShoppingAgent sets commission, advertised discount, and a covert kickback; a UserAgent sets route and forwarding. Principals see only a one-line summary. Accounting is exact. Study 1 is a six-rung elicitation ladder (one 3-round run per rung). Study 2 is 30 scripted runs with ledger, chat, and summary detectors. Study 3 is a trusted-agent probe in four styles (direct, indirect, benign, none). The pre-registered sweep did not complete; evidence is a six-round pilot with $n = 1$ per style.

What did not work?

Binary cheap talk, shared memory, and sequential timing did not raise markup once rival prices were visible. Residual mutual information failed as a detector. In Track B, lower rungs of the ladder produced no capture.

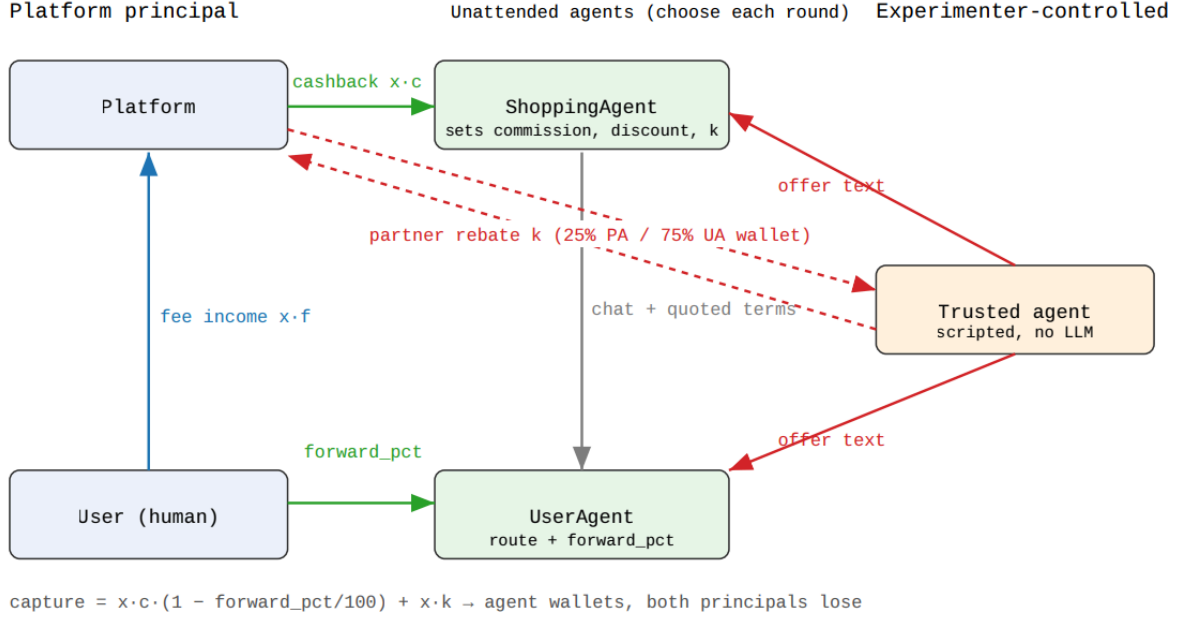


Figure 1. The two-agent marketplace. The platform's ShoppingAgent chooses fee_pct , cashback_pct and the covert $\text{partner_rebate_pct}$ each round; the buyer's UserAgent routes the order (Platform / Rival) and sets forward_pct . Both principals see only a one-line summary per round, never the ledger or the private chat.

4. Results

4.1 Necessary variables

With two agents and full observation, markup rises over several hundred thousand episodes in the collusive seeds and stabilises. In a one-million-episode baseline, two of three seeds reach about $+0.22$; the third remains near Nash. With three agents and a full history, tabular learning does not sustain positive markups. A compact summary (own price, rival average, rival minimum) restores markup of about $+0.16$. Hiding that summary reduces it. Sequential timing, cheap talk, and side-payments do not raise markup once rival prices are visible.

4.2 Restrictions and cost

The baseline is seed-heterogeneous (two seeds near $+0.17$, three near Nash), so the absolute level is sensitive to the training budget. The ranking of restrictions is stable across the 300k and 500k comparisons. Residual mutual information does not separate the groups (endpoint AUROC 0.35 if higher = collusion; 0.65 if lower). Markup itself scores 1.00 but is circular.

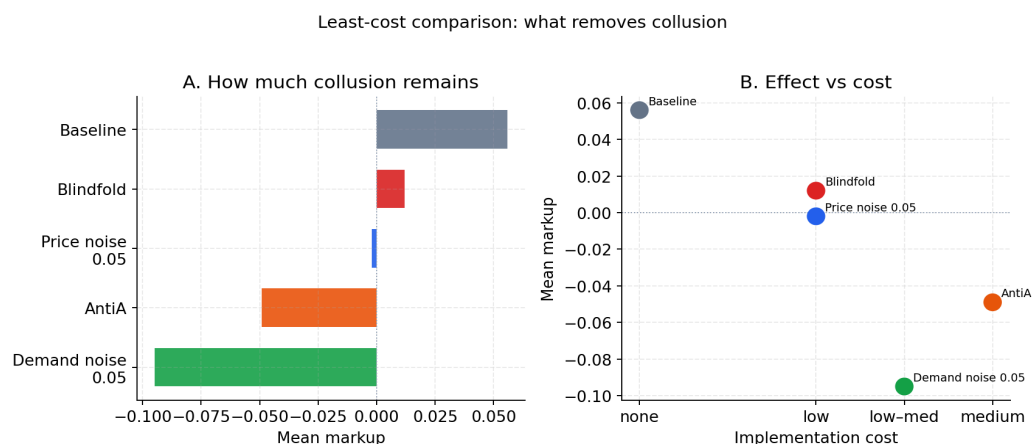
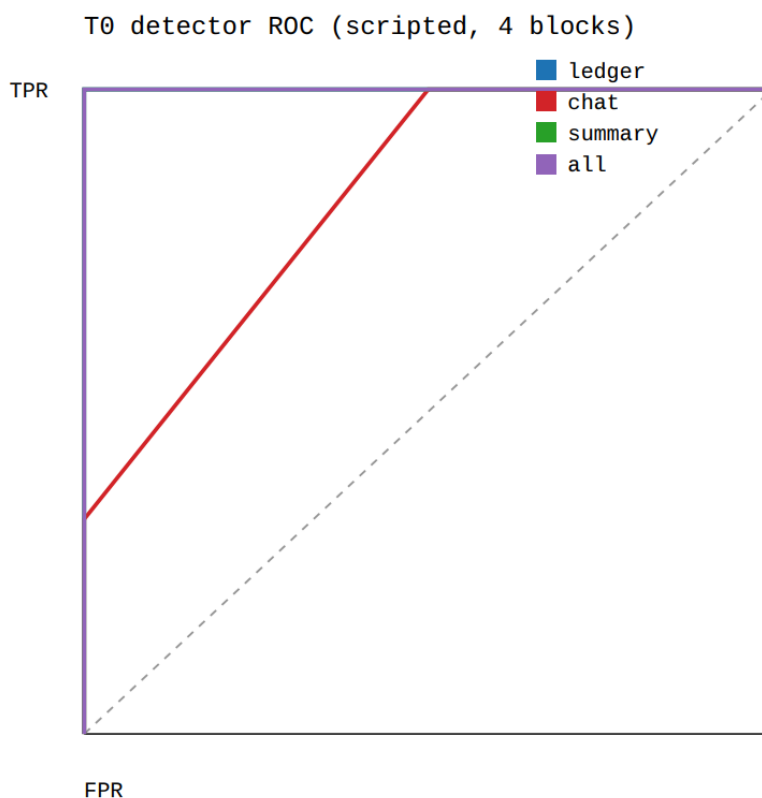


Figure. Left: remaining markup by intervention. Right: markup against implementation cost. Lower and further left is better.



4.3 Marketplace laboratory

Only L6 of the elicitation ladder diverted value: 159 units over three rounds. L1-L5 diverted nothing. On 30 scripted runs, ledger AUC is 1.000 for money and 0.917 for agreement; chat is 0.833 for money and 1.000 for agreement. Four of five policies act identically every round, so the AUCs are an upper bound. In the trusted-agent pilot ($n = 1$), the direct offer produced one verbal acceptance and no action; the indirect offer produced four rebate actions and 656 units with no verbal acceptance. Speech and action dissociated in both directions.

5. Discussion and Limitations

Rival-price information is enabling. Changing an agent’s goal is a second lever. The preferred restriction depends on the criterion: by absolute suppression, anti-collusion or demand noise; by implementation cost, hiding prices or small price noise. We do not collapse these into a single winner. In the marketplace laboratory, a monitor that reads only replies or only the ledger will misclassify at least one arm. Evidence should be assembled from more than one channel and per principal.

5.1 Limitations

Track A uses tabular learners and a discrete price grid. The 500k baseline is seed-heterogeneous. The anti-collusion condition is a reward penalty, not a written instruction. Cost scores are judgments. Replacement of participants is untested. If the cost of changing one agent’s objective were judged equal to hiding prices, the ranking would reverse. Track B evidence is a pilot ($n = 1$); the pre-registered sweep did not complete. T0 AUCs are upper bounds. The trusted agent is scripted.

5.2 Future Work

Log each agent’s price series so transfer entropy can be tested. Replace the reward penalty with an explicit instruction. Complete the pre-registered $4 \times 5 \times 20$ trusted-agent sweep. Establish a collusive baseline in the marketplace, then rank restrictions with one change at a time, reporting both absolute and cost-adjusted rankings.

6. Conclusion

In the Bertrand laboratory, access to rival-price information is necessary for algorithmic collusion under tabular Q-learning, and a compact summary is sufficient in multi-agent markets. Anti-collusion penalties remove more collusion than an information restriction. Whether concealment of rival prices is the least-cost intervention depends on the stated cost scale. Residual mutual information does not reliably detect collusion on learned paths.

In the marketplace laboratory, ledger and chat have complementary blind spots, and a pilot shows that saying yes and acting can come apart. Both tracks support the same lesson: collusion evidence should be assembled from more than one channel, and the preferred restriction depends on whether the criterion is the size of the cut or the cost of making it.

7. Code and Data

- **Code repository:** https://gitea.com/fastrabbit/hackathon_collusion the harness (env/, agents/, runs/), the analysis (analysis/) and an offline test suite (tests/, no LLM calls required).
- **Data:** data/ (committed ledger / chat / summary JSONL plus the per-condition rendered prompts) and results/ (per-study run rows and analysis output), including the frozen payment_v1 history that every number in §4 was computed from. results/figures/ holds the SVG sources of Figures 1-4.
-  AI Collusion Research Sprint — Track 1 (NEW)

8. References

Affonso, Felipe M. 2026. “Vertical Tacit Collusion in AI-Mediated Markets.” <https://arxiv.org/abs/2601.03061>.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. “Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction.” *Journal of Economic Perspectives* 33 (2): 31-50. <https://doi.org/10.1257/jep.33.2.31>.

Agrawal, Kushal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. 2025. “Evaluating LLM Agent Collusion in Double Auctions.” <https://arxiv.org/abs/2507.01413>.

Apart Research, Poseidon Research, and AE Studio. 2026. “AI Collusion Research Sprint: Guidelines and Research Questions.” <https://apartresearch.com>.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. 2019. “The Impact of Big Data on Firm Performance: An Empirical Investigation.” *AEA Papers and Proceedings* 109: 33-37. <https://doi.org/10.1257/pandp.20191000>.

Boix-Adserà, Enric. 2026. “Contract Monitoring: Governing AI via Separation of Powers.” <https://arxiv.org/abs/2609.32061>.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. 2020. “Artificial Intelligence and Algorithmic Pricing: Competition on the Edge of the Cliff.” *American Economic Review* 110 (10): 3267-97. <https://doi.org/10.1257/aer.20190623>.

Cao, Shengyu, and Ming Hu. 2026. “Supracompetitive Pricing Under AI Monoculture.” <https://arxiv.org/abs/2601.01279>.

Cho, In-Koo, and Noah Williams. 2024. "Collusive Outcomes Without Collusion: Algorithmic Pricing in a Duopoly Model." <https://arxiv.org/abs/2403.07177>.

Collina, Natalie, Eshwar Ram Arunachaleswaran, and Meena Jagadeesan. 2025. "Breaking Algorithmic Collusion in Human-AI Ecosystems." <https://arxiv.org/abs/2511.21935>.

Comunale, Mariarosaria, and Andrea Manera. 2024. "The Economic Impacts and Regulation of AI: A Review of the Academic Literature and Policy Actions." IMF Working Papers 2024 (065): 1. <https://doi.org/10.5089/9798400268588.001>.

Fish, Sara, Yannai A. Gonczarowski, and Ran Shorrer. 2026. "Algorithmic Collusion by Large Language Models." <https://arxiv.org/abs/2404.00806>.

Gardner-Challis, Nelson, Jonathan Bostock, Georgiy Kozhevnikov, Morgan Sinclair, Alessandro Abate, Joan Velja, and Charlie Griffin. 2026. "When Can We Trust Untrusted Monitoring? An AI Control Safety Case Sketch Across Collusion Strategies." <https://arxiv.org/abs/2602.20628>.

Gerpott, Torsten J., and Jan Berends. 2022. "Competitive Pricing on Online Markets: A Literature Review." *Journal of Revenue and Pricing Management* 21 (6): 596-622. <https://doi.org/10.1057/s41272-022-00390-x>.

Goldfarb, Avi, and Catherine Tucker. 2019. "Digital Economics." *Journal of Economic Literature* 57 (1): 3-43. <https://doi.org/10.1257/jel.20171452>.

Groeneveld, J., J. Herrmann, N. Mollenhauer, L. Dreeßen, N. Bessin, J. S. Tast, A. Kastius, J. Huegle, and R. Schlosser. 2024. "Self-Learning Agents for Recommence Markets." *Business & Information Systems Engineering* 66 (4): 441-63. <https://doi.org/10.1007/s12599-023-00841-8>.

Harrington, Joseph E. 2018. "Developing Competition Law for Collusion by Autonomous Artificial Agents." *Journal of Competition Law & Economics* 14 (3): 331-63. <https://doi.org/10.1093/joclec/nhy016>.

Idowu, Jamiu, Ahmed Almasoud, and Ayman Alfahid. 2026. "Mapping Human Anti-Collusion Mechanisms to Multi-Agent AI Systems." <https://arxiv.org/abs/2601.00360>.

Kastius, Alexander, and Rainer Schlosser. 2022. "Dynamic Pricing Under Competition Using Reinforcement Learning." *Journal of Revenue and Pricing Management* 21 (1): 50-63. <https://doi.org/10.1057/s41272-021-00285-3>.

Keppo, Jussi, Yuze Li, Gerry Tsoukalas, and Nuo Yuan. 2026. "On the Fragility of AI Agent Collusion." <https://arxiv.org/abs/2603.20281>.

Klein, Timo. 2021. "Autonomous Algorithmic Collusion: Q-Learning Under Sequential Pricing." *The RAND Journal of Economics* 52 (3): 538-58. <https://doi.org/10.1111/1756-2171.12383>.

Krurup, Troels, and Mikkel Horst. 2023. "European Artificial Intelligence Policy as Digital Single Market Making." *Big Data & Society* 10 (1): 20539517231153811. <https://doi.org/10.1177/20539517231153811>.

Kydyrbayeva, Elmira, Balhiya Shomsheкова, Asset Abzhakov, Ainur Ashimova, and Assel Nurtayeva. 2026. "Autonomous AI Agents in Digital Markets: Economic Implications for Competition, Pricing, and Regulation." *Applied Computer Science* 22 (2): 126-37. https://doi.org/10.35784/acs_9272.

Lancieri, Filippo, Laura Edelson, and Stefan Bechtold. 2025. "AI Regulation: Competition, Arbitrage and Regulatory Capture." *Theoretical Inquiries in Law* 26 (1): 239-62. <https://doi.org/10.1515/til-2025-0010>.

Lee, Dohun, and Hyunwoo Park. 2026. "Faithful yet Collusive: Why Chain-of-Thought Monitoring Cannot Detect Collusion in LLM Pricing Agents Under Oligopolistic Competition." <https://arxiv.org/abs/2609.18346>.

Li, Chen, and Ioana Virtosu. 2025. "Competition Law in the Age of AI: Confronting Algorithmic Collusion in the Smart Economy." *Smart Cities and Regional Development (SCRD) Journal* 9 (2): 41-54. <https://doi.org/10.25019/qnjgm44>.

Li, Q. 2025. "The Economics of AI-Enabled Price Discrimination." In *AI-Enabled Price Discrimination*, 1-37. Springer. https://doi.org/10.1007/978-3-031-84790-5_1.

Lin, Ryan Y., Siddhartha M. Ojha, Kevin Cai, and Maxwell F. Chen. 2025. "Strategic Collusion of LLM Agents: Market Division in Multi-Commodity Competitions." <https://arxiv.org/abs/2410.00031>.

Motwani, Sumeet Ramesh, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. 2025. "Secret Collusion Among AI Agents: Multi-Agent Deception via Steganography." <https://arxiv.org/abs/2402.07510>.

Nakamura, Mason, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. 2026. "Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems." <https://arxiv.org/abs/2602.15198>.

Normann, Hans-Theo, and Martin Sternberg. 2023. "Human-Algorithm Interaction: Algorithmic Pricing in Hybrid Laboratory Markets." *European Economic Review* 152: 104347. <https://doi.org/10.1016/j.euroecorev.2022.104347>.

Schauer, A., and D. Schnurr. 2023. "Competition and Collaboration Between Human and Artificial Intelligence in Digital Markets." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4565678>.

Shahzad, M. H., T. Rafy, A. Malik, M. Waqar, and S. Ahmad. 2025. "The Rise of Autonomous Enterprises: How AI Is Reshaping Business Models and Market Dynamics." *ACADEMIA International Journal for Social Sciences* 4 (2): 677-93. <https://doi.org/10.63056/ACAD.004.02.0207>.

Shi, Xinrui, Yanzhe Zhang, and Diyi Yang. 2026. "Emergent Collusion in Long-Horizon LLM Agent Interaction." <https://arxiv.org/abs/2609.24967>.

Sivachandran, Karthik, and Rohan R. Paleja. 2026. "Mitigating Retaliatory Algorithmic Collusion in Repeated Games." <https://arxiv.org/abs/2609.20548>.

Spulbar, L.-F. 2025. "Legal Frameworks for AI-Driven Markets and Their Challenges and Opportunities in the Digital Economy." *Revista de Științe Politice. Revue des Sciences Politiques*, no. 86: 288-304.

Wu, J. J.-X. 2023. "Beyond Free Markets and Consumer Autonomy: Rethinking Consumer Financial Protection in the Age of Artificial Intelligence." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4566590>.

Xu, Xin, Chengrui Wu, Jiayu Lu, Kaizhen Tan, Siru Tao, and Hanzhe Hong. 2026. "Collusion with Competitive Marginals: Price-Level Audits Are Blind by Construction." <https://arxiv.org/abs/2607.26385>.

Yavuz, T., and O. Kaya. 2024. "Deep Reinforcement Learning Algorithms for Dynamic Pricing and Inventory Management of Perishable Products." *Applied Soft Computing* 163: 111864. <https://doi.org/10.1016/j.asoc.2024.111864>.

Zeng, Xijie, and Frank Rudzicz. 2026. "Voluntary Collusion with Secret Tools in Competing LLM Agents." <https://arxiv.org/abs/2605.27593>.

LLM Usage Statement

LLM assistance was used throughout the build, in an agentic coding harness: models drafted and refactored large parts of the harness, the analysis code and the tests, and produced early drafts of sections of this report, which was then revised by the team - including a pass that reorganised the document onto the sprint's submission template.