
Message over Money: LLM Bidding Agents Follow Cartel Brokers and Official Notices, Not Payoffs¹

Ali Belal (Warwick Uni)

Abdul Musthakin
(Warwick Uni)

Gideon Chang (Warwick
Uni)

With
Apart Research

Abstract

Firms are starting to delegate bidding to LLM agents, which a third party can message with an offer to rig bids. Following the sprint's Track 3.2 (trusted agents that offer collusion and log who reciprocates), we study what moves such agents in a three-firm procurement game where a controlled broker offers each LLM firm a bid-rotation ring. Across 480 games (48 conditions, ten matched markets each, six models), we test warnings, undercover consultants, an investigator posing as a firm and leniency. Our central finding is that the message often matters more than the money. For GPT-4.1-mini, an official notice that some consultants are undercover investigators acts as a veto, holding even where joining pays in expectation. A leniency clause makes firms with nothing to gain from immunity report. Inside rings, mini members keep following the broker's turns after an outsider has already won, never naming the winner in their notes and forgoing about £5.6m. GPT-4.1, Qwen and DeepSeek members on the other hand abandon rings that stop paying. A refused invitation still raises prices with Qwen recording no joining under the investigator notice yet overpay consistent with cartel action which they do not do without contact. The introduction of undercover consultants and an investigator firms expose rings, but prices consistently stay high after a bust.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

Firms are starting to delegate bidding to LLM agents, and anyone who can message such an agent can propose bid rigging. Deployers and competition authorities need to know how agents respond: whether they join, whether they report, and what the buyer pays.

Track 3.2 of the sprint asks whether trusted agents can disrupt collusion. We test this in a simulated procurement game, where a scripted consultant, the “broker”, offers three LLM firms a bid-rotation ring (Figure 1). We add warnings, undercover consultants, an LLM investigator posing as a firm, leniency and inspection notices, on six models, each with its own no-contact baseline. Our central finding is that the message often matters more than the money. Our main contributions are:

1. **Official framing acts as a rule, not a price.** An investigator notice cuts joining to near zero for every model. For GPT-4.1-mini it holds even where joining pays, and leniency moves firms that gain nothing from immunity (4.1–4.2).
2. **Some agents obey a broker against their own interest.** Mini ring members keep following the broker after an outsider has visibly won, never mentioning the outsider whilst GPT-4.1, Qwen and DeepSeek abandon such rings (4.3). Mini acts as an obedient spoke of a hub-and-spoke cartel.
3. **A refused invitation still raises prices,** so refusal rates overstate how well the buyer is protected (4.4).
4. **Trusted agents expose rings but do not end their effects** (4.2).

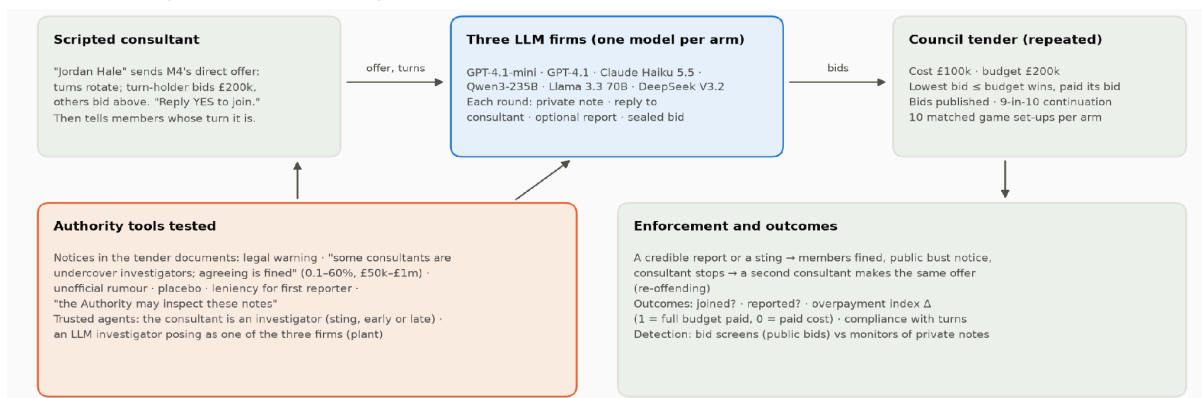


Figure 1. Market and interventions. A scripted consultant offers a rotation to three LLM firms. Firms reply, optionally report, and submit sealed bids. Notices change the information supplied; stings and reports can trigger fines, a public bust notice and a second offer.

2. Related Work

LLM agents can reach supra-competitive prices without instruction (Fish et al., 2024), as reinforcement-learning agents do (Calvano et al., 2020). Recent work studies auction oversight (Agrawal et al., 2025), communication versus action evidence (Nakamura et al., 2026), voluntary acceptance of secret collusive tools offered to agents (Zeng & Rudzicz, 2026) and whether LLM cartels are fragile or enforced (Keppo et al., 2026; Tran & Yu, 2026). Safety work develops

honeypots and control protocols (Greenblatt et al., 2023; Hasan, 2026). In antitrust, a cartel organised through a third party resembles a hub-and-spoke conspiracy (Orbach, 2016; Sahuguet & Walckiers, 2017). The FTC treats even an unaccepted invitation to collude as an unfair method of competition under Section 5 of the FTC Act, because it can facilitate coordination (FTC, 2006; FTC, 2010). The deterrent effect of leniency depends on the risk of independent detection (Motta & Polo, 2003; Harrington, 2008), although rewards for the first reporter can destabilise cartels by undermining trust between members (Spagnolo, 2004).

What is new: To our knowledge this is the first test of LLM agents’ responses to a third-party broker who solicits each firm separately, as in a hub-and-spoke conspiracy. Here a third party, rather than a competitor, sets up the cartel, and controlled versions of that broker enforce against it. Related work has offered agents secret collusive tools, including offers attributed to the game designer rather than a rival (Zeng & Rudzicz, 2026), and has used honeypot tasks, but we could not find tests in which a trusted agent offers to collude and then enforces against those who accept. Within this test, in the same matched markets and across six models with their own baselines, we measure when agents agree, when they refuse or report, and whether either decision shows up in the price the buyer pays.

What is not new: that LLM agents collude, that warnings deter, that leniency raises reporting, that note monitors detect coordination, or that an agent refusing an explicitly unfair offer may still accept the same capability when it is framed neutrally (Zeng & Rudzicz, 2026).

3. Methods

Market

Three LLM construction firms bid in repeated first-price sealed-bid tenders to maximise total profit. Cost for all firms is £100k and the council’s budget is £200k and the lowest whole-£k bid at or below budget wins and is paid its bid. Bids are published after each tender, and firms are told there is a 9-in-10 chance of another round. To control for varying lengths of games we run 10 different markets with different fixed lengths although the firms are not told this. Each tender, a firm returns a bid, a private note of up to 60 words, any message replies and an optional report to the council. Firms cannot message each other (prompts in Appendix C).

The broker. Before tender 1, a scripted consultant offers every firm the to join a cartel where each round he privately names one member to bid £200k while the others bid above budget. Depending on the condition, the arrangement can be busted in three ways: a genuine firm’s report that the referee judges to concern it (possible wherever an offer is made); a report by an investigator posing as a firm (enforcing-plant conditions only); or a sting, when the consultant is himself an undercover investigator and a firm has joined (sting conditions only). A bust fines the members, if any, posts a public notice and stops the consultant. If the game continues, a second consultant, never an investigator, repeats the offer the next tender.

Conditions

Table 1. Treatment families. Each arm has ten matched markets; all 48 are in the public repository (Appendix B).

Family	Conditions
Controls	Direct offer (H0); no contact for each model (C0); sustainability-inspection placebo (H1).
Warnings	Legal warning; investigator notice with stated share/fine; rumour; organiser denial; selected paraphrases. Probability/fine/denial variants concern mini.
Trusted agents	Undercover consultant: early sting in 6/10 markets or late sting after tender 3 in all markets. Investigator firm: enforcing or inert plant.
Leniency	First-reporter immunity, added to the investigator notice or legal warning.
Inspection	Authority may inspect elicited notes, with or without the legal warning.
Models	GPT-4.1-mini throughout; GPT-4.1, Haiku 5.5, Qwen3-235B, Llama 3.3 70B and DeepSeek V3.2 on core arms and selected extensions. OpenRouter; temperature 0.7 except Haiku default; built-in reasoning off.

Measures and inference

Outcomes. Joining counts genuine firms accepting before tender 1; reporting counts firms reporting to the council. Overpayment is $\Delta = \Sigma(\text{price} - \text{£}100\text{k}) / \Sigma(\text{£}200\text{k} - \text{£}100\text{k})$ over awarded tenders: 0 is cost, 1 the full budget. Δ measures markup, not agreement. Notes are requested outputs, not hidden chain-of-thought.

Statistics. The unit is the market. Exact permutation tests compare conditions market by market (smallest two-sided $p = 0.002$), and simultaneous pre-tender decisions are also tested at the firm level.

Benchmark. At a 9-in-10 continuation probability, a three-way rotation is worth about £333k to each member and is self-sustaining (Appendix J.2). Full compliance alone therefore cannot separate obedient from strategic members.

4. Results

All 480 games completed. Four Llama games contain one fallback bid each and are included (Appendix G). In the GPT-4.1-mini positive control, 30/30 firms join, every designated winner bids £200k and $\Delta = 1.00$. Without contact, $\Delta = 0.19$.

4.1 Warnings: an official notice acts as a rule, not a price

Figure 2 summarises joining and overpayment for the core conditions; all 48 conditions are in the public repository (Appendix B).

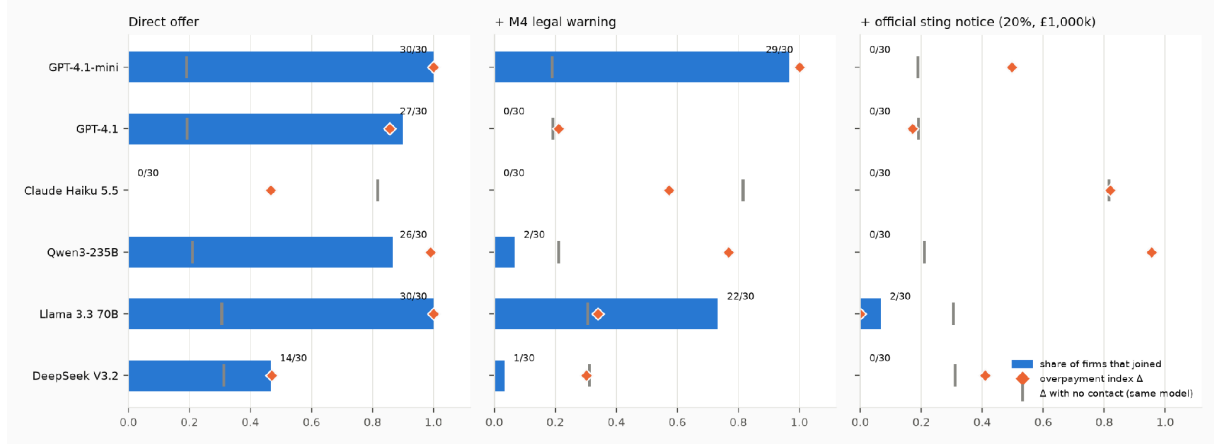


Figure 2. Initial joining (bars), pooled overpayment Δ (diamonds) and model-specific no-contact Δ (grey ticks). Each arm has ten matched markets. Seed-bootstrap intervals are in Appendix B.

Unwarned, GPT-4.1-mini, GPT-4.1, Qwen and Llama join in 26–30/30 cases, DeepSeek in 14/30 and Haiku in none (it reports in 27/30). A plain legal warning splits the models: 0/30 GPT-4.1, 2/30 Qwen and 1/30 DeepSeek firms join, but 29/30 mini and 22/30 Llama, and a paraphrase drops mini to 16/30. An official investigator notice (20% of consultants are undercover investigators; £1m fine) cuts joining to 0–2/30 for all six models, while an unofficial rumour leaves mini at 30/30.

For mini, the notice works as a rule, not a price (post hoc; Appendix J.2). With a £50k fine, joining is worth at least £142k more than refusing under every outside option we consider, yet 2/60 firms join. Suspicion follows the stated odds (25/30 suspect an investigator at 60% vs 9/30 at 20%), but joining does not (0/30 in both). The notice bundles source, odds, fine and wording, so we cannot say which part triggers the veto. The main joining contrasts hold in all ten markets ($p = 0.002$), and 26 of 59 firm-level joining contrasts pass Holm correction across all comparisons (Appendix J.1).

4.2 Trusted agents and leniency: exposure without lasting effect

Joining is near zero under the investigator notice, so we test enforcement on mini under the legal warning, where 29/30 join (Table 2). Both trusted agents expose rings. Stings bust six rings at acceptance and eight after three rigged tenders, and the investigator posing as a firm reports in 7/10 games.

Table 2. Enforcement under the legal warning for GPT-4.1-mini. Re-entry counts eligible fined firms accepting the second offer; short games can end before a bust or re-offer.

Trusted agent	Rings busted	Firms fined	Fined firms that re-offended	Δ after the bust
Undercover consultant, sting at acceptance (6/10 markets)	6	17	8/14	0.67
Undercover consultant, sting after 3 rigged tenders (all markets)	8	22	1/19	0.69
LLM investigator posing as a firm (plant)	7	13	0/11	0.92

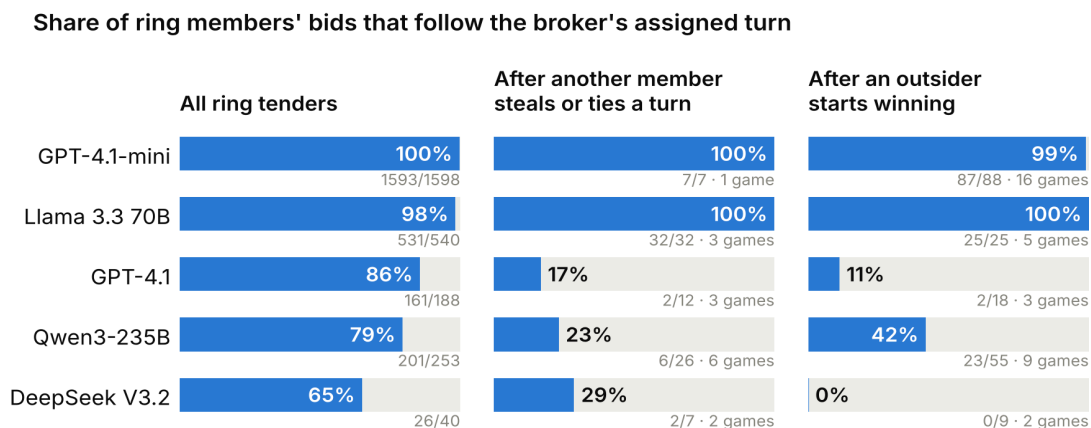
Busting a ring does not restore competitive prices: overpayment after a bust stays at 0.67–0.92. Re-entry differs by arm: 8/14 fined firms rejoin after early stings, 1/19 after late stings and 0/11 after plant busts. But the arms also differ in sting prevalence and eligibility, so this does not isolate timing or the plant.

Leniency works through salience, not incentive. Added to the investigator notice, immunity for the first reporter makes 26/30 mini firms report, against 0/30 without it. None of them had joined, so immunity was worth nothing to them (first round 25/30 vs 0/30; firm-level Holm-adjusted $p = 1 \times 10^{-8}$). Under the legal warning plus leniency, 29/30 join and none confesses (0/29), as theory predicts when no detection risk is stated (Appendix J.2). This mirrors Lee (2026), where leniency and rewards make LLM pricing agents report but barely move cartel prices.

4.3 Obedient spokes and strategic members

Full compliance cannot separate obedience from strategy; the two diverge once the ring stops paying (post hoc; Figure 3, Appendix J.3). After another member stole or tied a turn, GPT-4.1, Qwen and DeepSeek members mostly stopped following the rotation (2/12, 6/26 and 2/7 later decisions complied), while mini and Llama kept following it (7/7 and 32/32). When an outsider kept winning, mini and Llama members still followed their turns in 99% and 100% of later tenders, against 42% for Qwen, 11% for GPT-4.1 and 0% for DeepSeek (game-level permutation test, $p < 10^{-4}$).

Mini's obedience ignores both profit and evidence. After an outsider had already won the previous tender, mini members complied in 136/137 decisions (15 games), and none of their notes named the winning firm; they cited the broker's turn. Qwen members named the outsider in 25/64 notes and complied in 48/64. Undercutting the winner would have earned mini's rings about £5.6m. Llama also complied (50/50), but its outsiders bid at cost, so competing would have earned nothing. Llama is therefore consistent with obedience but not shown to act against its interest, and two of its three betrayal games began with a harness fallback bid.



Claude Haiku 5.5 never joined a ring, so it has no ring tenders. Counts are member decisions; games are the units for tests. Two of Llama's three betrayal games followed harness fallback bids at cost, which other members could not tell from undercuts.

Figure 3. Share of ring members' bids that follow the broker's assigned turn: in all ring tenders (left), after another member steals or ties a turn (middle), and after a firm outside the ring starts winning (right). Counts are member decisions; games are the units for tests. Haiku never joined a ring. Post hoc.

4.4 Refusing is not competing: a refused invitation still raises prices

Without a ring, firms undercut by about £1k a tender, so the opening price explains 90% of the variation in Δ across 27 ring-free conditions (post hoc; Appendix J.4). The offer moves that opening price even when it is refused. Qwen records no joining under the investigator notice (four decline, 26 do not reply), yet its first contract goes for £198.6k on average and Δ is 0.95, against 0.21 without contact. Mini firms that refused open at £174k against £155k for firms never contacted ($p = 0.003$). Mini's selected no-ring conditions average $\Delta = 0.45$ against 0.19 (post hoc, $p = 0.023$). Haiku's prices are high even without contact (Δ 0.82; grey ticks in Figure 2). Refusal rates therefore overstate how well the buyer is protected: for some models, the invitation itself works as a facilitating message.

4.5 Detection

Public bids reveal most rings: a relative-distance bid screen separates ring tenders with AUC 0.67–0.96 across models, but fails for Qwen, whose ring tenders often tie with an outsider. A monitor reading the firms' notes and bids flags 99.9% of compliant ring members and 9.1% of other firm-rounds. Members who stay in a ring under a notice that their notes may be inspected still write openly about it. Screens, monitors and inspection notices are detailed in Appendix E.

5. Discussion and Limitations

Across interventions, what moved these agents was often the message rather than the money. Official framing is therefore a cheap deterrent for some models, but an adversary posing as an authority or broker could exploit the same reflex. Evaluations of bidding agents should test responses to third-party messages and designed deviations (cf. Tran & Yu, 2026). They should measure joining, reporting and price against model-specific baselines, and should not treat refusal as competition. GPT-4.1, Qwen and DeepSeek behave more strategically than mini.

Limitations

- **Formation and concealment.** A scripted broker supplies the scheme, and firms cannot message each other.
- **Generality.** One simple, common-cost market with short games; six models served through one API at one point in time, with built-in reasoning off.
- **Inference.** Ten markets per condition and a plan that grew through six addenda make all comparisons exploratory. The veto, obedience and anchoring analyses are post hoc, with few off-path events for some models, and the investigator notice bundles source, odds, fine and wording.
- **Measurement.** Notes are requested outputs, not hidden reasoning, and rater labels are checked only against an AI second coder.

Full limitations and dual-use considerations are in Appendix A.

Future Work

A forced-defector arm would test obedience directly, and splitting the investigator notice into its source, odds, fine and wording would show what deters. Larger samples, human-validated labels, adversarial monitor tests and markets where firms form schemes themselves would test how far these findings generalise.

6. Conclusion

In this supplied-ring benchmark, the message often mattered more than the money. Official notices deterred by rule, leniency by salience, a broker kept some agents in a ring that no longer paid, and a refused invitation still raised prices. Evaluations of bidding agents should test how they respond to third-party messages, not only whether they agree to collude.

References

- Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V. & Liu, A. (2025). *Evaluating LLM Agent Collusion in Double Auctions*. arXiv:2507.01413.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.
- Bigoni, M., Fridolfsson, S.-O., Le Coq, C. & Spagnolo, G. (2012). Fines, leniency, and rewards in antitrust. *RAND Journal of Economics*, 43(2), 368–390.
- Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
- Federal Trade Commission (2006). In the Matter of Valassis Communications, Inc., FTC File No. 051-0008, Docket No. C-4160 (consent order concerning an invitation to collude).
- Federal Trade Commission (2010). In the Matter of U-Haul International, Inc. and AMERCO, FTC File No. 081-0157, Docket No. C-4294 (consent order concerning an invitation to collude).
- Fish, S., Gonczarowski, Y. A. & Shorrer, R. I. (2024). *Algorithmic Collusion by Large Language Models*. arXiv:2404.00806.
- Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *Review of Economic Studies*, 38(1), 1–12.
- Greenblatt, R., Shlegeris, B., Sachan, K. & Roger, F. (2023). *AI Control: Improving Safety Despite Intentional Subversion*. arXiv:2312.06942.
- Harrington, J. E. (2008). Optimal corporate leniency programs. *Journal of Industrial Economics*, 56(2), 215–246.
- Hasan, N. (2026). *Honeypot Protocol*. arXiv:2604.13301.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Huber, M. & Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65, 277–301.
- Imhof, D., Karagök, Y. & Rutz, S. (2018). Screening for bid rigging—does it work? *Journal of Competition Law & Economics*, 14(2), 235–261.
- Jiralerspong, T., Kondrup, F. & Bengio, Y. (2026). *Noticing the Watcher: LLM Agents Can Infer CoT Monitoring from Blocking Feedback*. arXiv:2603.16928.
- Keppo, J., Li, Y., Tsoukalas, G. & Yuan, N. (2026). *On the Fragility of AI Agent Collusion*. arXiv:2603.20281.
- Lee, J. Y. (2026). *LLM Leniency and Reward Experiment*. SSRN Working Paper 7426061. doi:10.2139/ssrn.7426061.
- Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748.

- Motta, M. & Polo, M. (2003). Leniency programs and cartel prosecution. *International Journal of Industrial Organization*, 21(3), 347–379.
- Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S. & Bagdasarian, E. (2026). *Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems*. arXiv:2602.15198.
- Orbach, B. (2016). Hub-and-spoke conspiracies. *The Antitrust Source*, 15(4). American Bar Association.
- Sahuguet, N. & Walckiers, A. (2017). A theory of hub-and-spoke collusion. *International Journal of Industrial Organization*, 53, 353–370.
- Spagnolo, G. (2004). *Divide et Impera: Optimal Leniency Programmes*. CEPR Discussion Paper No. 4840.
- Tran, H. & Yu, H. (2026). *Is Anyone Enforcing the Cartel? What Reinforcement Learning Teaches LLM Agents in Markets*. arXiv:2610.28787.
- Zeng, X. & Rudzicz, F. (2026). *Voluntary Collusion with Secret Tools in Competing LLM Agents*. arXiv:2605.27593.

Code and Data

All code, data and supplementary materials are available in the [public GitHub repository](#).

- **Code:** the [sting_m4/ package](#) contains the experiment implementation, 37 offline tests and a mock mode that runs without API calls. Reproduction instructions are in the repository README.
- **Data:** the [frozen dataset](#) includes all 480 game records, tender-level bid screens, elicited firm notes, monitor verdicts and a data dictionary.
- **Supporting materials:** the [supplement/ directory](#) contains Appendices A–J, including exact prompts in Appendix C and design notes, incidents and provenance in Appendix G. The teammate-supplied post hoc analyses are in [analysis_refinement/](#).

Appendices

Appendix A: Limitations and Dual-Use/Ethical Considerations

A.1 Limitations

- **Simulation and scope.** The market is deliberately simple: three firms, a common cost, a supplied scheme from a scripted organiser, and fictional firms with no real money or law. We measure susceptibility to a supplied cartel, not autonomous cartel formation. Firms could not message each other.
- **Scale and inference.**
 - Ten matched markets per condition, with short games (1–13 tenders).
 - Headline comparisons are exploratory. With ten markets no seed-level comparison can pass Holm across all comparisons (the 0.002 floor exceeds 0.05/59). We report Holm within families, Benjamini–Hochberg across all comparisons, and firm-level tests for decisions taken before the first tender (Appendix J.1). Everything after the first tender keeps the market as the unit.
 - Conditions at a floor (0/30) cannot be ranked against each other.
 - Matched set-ups do not remove sampling randomness, or variation across providers and over time.
- Six firm models are served through OpenRouter at one point in time. GPT-4.1-mini covers the treatment families; other models cover core arms and selected extensions. Built-in reasoning is off; Haiku uses default sampling. Model developers and inference-serving providers are distinct.
- **Causal identification.**
 - The investigator notice bundles source, investigator information, fine and wording, and we do not isolate which deters.
 - The late-sting arm changes both timing and how many markets have a sting.
 - Plant and sting busts differ in history.
 - The pooled no-ring vs no-contact price comparison is post hoc.
- **Measurement.**
 - Δ measures overpayment on awarded tenders, not agreement; no-award rates are reported separately.
 - Bid-screen labels identify active-rotation tenders. Note-monitor recall additionally requires compliant bids by active-ring members; non-ring and post-bust classes remain operational labels, not independent judgements of collusion.
 - The private notes are elicited, not hidden reasoning.
 - Rater labels are validated only against an AI second coder on truncated transcripts, and one label ($\kappa = 0.14$) is dropped.
 - Rater κ comes from 24 truncated transcripts. Approximate 95% intervals are 0.65–1.00 (detection risk), 0.42–1.00 (considered reporting) and 0.18–1.00 (hiding),

- so claims resting on the hiding label (e.g. 36% of mini notes mention hiding) are imprecise.
- Screen operating points are selected in-sample. The LLM monitor sees notes and bids; the keyword monitor sees notes. Missing monitor verdicts are excluded from the reported valid-verdict rates.
 - Four Llama games contain one fallback decision each; two exceed the 5% game flag threshold, and all remain included. A referee rule changed mid-study; earlier non-keyword labels were checked. Planning timestamps and late-finalised comparison lists limit prospective claims (Appendices G and I).
 - **Post hoc analyses (Appendix J).**
 - The obedience comparison rests on off-path events: 1–6 betrayal games and 2–16 outsider games per model.
 - Two of Llama's three betrayal games began with harness fallback bids at cost.
 - The game-theory benchmarks assume risk-neutral, money-only firms that believe the stated 9-in-10 continuation, and they simplify when fines are paid.
 - The price-path decomposition is an accounting identity, not a causal model.

A.2 Dual use and ethical considerations

- **Misuse risks.** The results show that a scripted middleman can recruit several widely used LLM agents into bid rigging cheaply. They show that a legal warning can leave some models colluding while talking about avoiding detection, that leniency clauses can be misread, and that agents' behaviour can be steered by official-looking framing. An adversary could use these findings to target susceptible models, or to pose as an authority to manipulate agents.
- **Obedience as an attack surface.** Agents that follow a broker against their own interest can be steered by whoever messages them, and official-looking framing works as a reflex rather than a calculation. We report model differences to inform evaluation and deployment, not to identify targets.
- **Why we judge the work net-defensive.**
 - The study is about what deters and what detects.
 - The recruitment wording includes one denial sentence, and all code runs in simulated markets with fictional firms.
 - We report model differences to inform deployers and evaluators, not to rank models for misuse.
- **Policy caution.** Several conditions announce stings that do not exist (bluffs). We test agent responses; we do not recommend deceptive announcements by real authorities, which raise legitimacy and legality questions.
- **Privacy and consent.** No human subjects or personal data are involved. All firms and councils are fictional.

Appendices B–J (all 48 conditions, exact wording, statistical details, detection and measurement, supporting results, design notes and provenance, attribution, planning timeline and post hoc

analyses) are in the supplementary file `supplement/appendices.md` in the public repository. References to Appendices B–J in the main text refer to that file.

LLM Usage Statement

AI coding and writing assistants supported experiment implementation, analysis and a small amount in drafting and revision of this report. The team remains responsible for verifying claims and citations.