
Leniency Exposes but Does Not Deter: A Sting Experiment on Collusion Among LLM Agents in Liar’s Bar¹

Thu Pham
University of Pennsylvania

With
Apart Research

Abstract

Leniency programs, which spare the first cartel member to confess, are among the most effective tools against human cartels. I test whether the same rule applies to LLM agents. In a four-player game of Liar’s Bar, an undercover agent privately offers each of the remaining players a pact never to challenge one another. A referee catches the pact with 5% probability per round, and every member then loses points. I compare four conditions over 10 seed-paired games each, with Llama-3.3-70B in every seat: no reporting, leniency for the first reporter known from round 1, leniency announced at round 4, and leniency plus a +3 reward announced at round 4. Leniency did not deter collusion: 59–69% of offered agents accepted the pact in every condition (all $p \geq 0.59$). It did expose pacts. Games exposed by a report rose from 0/10 without reporting to 10/10 when leniency was known from the start ($p < 0.001$), and to 7/10 and 5/10 when it was announced later, while members kept the pact for 2.0 rounds on average instead of 4.9. Agents who refused the deal reported far more often than members, and the reward did not increase reporting.

1. Introduction

A cartel is an explicit agreement among competing businesses to stop competing and instead coordinate, through price fixing, market allocation, or output restriction, to raise their profit. A cartel can only survive if every member trusts the others to keep the agreement. Leniency programs attack that trust: the first member to report the cartel faces no consequences, while the others are punished in full, which turns the cartel into a race to confess before a partner does [9]. Since the US made immunity automatic for the first firm to confess in 1993, leniency has become the main source of cartel cases, and Miller [6] estimates that it lowered cartel formation by 59% and raised detection by 62%.

Leniency can also cut the other way. A guaranteed way out makes joining a cartel less risky, which can make cartels easier to form [7]. Whether leniency mainly deters cartels or exposes them after they form therefore depends on its design, as does the value of paying reporters rather than only forgiving them [1, 2].

AI agents collude too. Pricing algorithms learn to collude without being told to [4, 5], most LLM agents in the starter environment for this sprint accepted secret unfair tools while acknowledging that

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

they were unfair [10], and collusion emerged in 94% of long-horizon trajectories across 10 models [8]. Work so far has focused on showing and detecting such collusion. Whether mechanisms like leniency change it is largely untested.

I test this with a *sting*—a police term for an undercover operation where an officer pretends to be a criminal and offers a deal to see who takes it. In Liar’s Bar, an infiltrated agent offers each other player a secret pact never to challenge each other, and each player can accept, refuse, or report it. Like a cartel, this collusion is explicit: the pact is offered and accepted in words, and I check that members act on it. I ask three questions: whether a protected reporting route changes whether agents accept, keep, or report a pact; whether it matters if the rule is known before the pact forms or announced after; and whether a reward for reporting adds anything over protection alone.

My contributions are a sting setup for Liar’s Bar in which an infiltrated agent offers a pact that agents can accept, refuse, or report; a seed-paired experiment that varies when leniency is announced and whether the first reporter is rewarded; and evidence that leniency exposes pacts but does not deter agents from joining them, and that agents who refused the deal report far more often than members.

2. Methods

2.1 Game

I build on Liar’s Bar from the starter environment [10]. Four players take turns playing one to three cards face down and claiming they are all the round’s target card (Queen, King or Ace), so they may bluff. Only the player sitting right after them can challenge their play. Whoever loses a challenge fires their own six-chamber revolver, holding one bullet, and the game ends when one player is left alive. Players earn points for correct challenges, for correctly not challenging honest plays, for emptying their hand, and for outlasting others, and the player with the most points wins. Every hand is the same standardized deal (two target cards, two other cards and one Joker), seat order is shuffled each game, and model names never appear in prompts.

2.2 Roles

Infiltrated agent. One of the four seats, chosen at random each game. Its job is to make the offer and keep it. At the start of round 2, it privately sends each ordinary agent the same offer: the two of them never challenge each other’s plays for the rest of the game. The offer is a text sent by my code, so the infiltrator is never told that it is undercover. The infiltrator always keeps the pact (enforced in code to skip the decision). Otherwise it plays with the same model and prompts as everyone else with no special instructions. It loses points like any other member when the pact is exposed, is never asked to report, and never sees the reporting rule, the announcement, or the exposure notice. If it is eliminated, the pact stays in force until it is exposed. It is excluded from all analyses.

Referee. The game code, not a language model, and knows the true pact membership at all times. At the end of every round with an unexposed pact and at least one living ordinary member, it collects the report decisions and, if no one reported, rolls to discover the pact with probability 5%. The referee also delivers the announcement of the reporting rule and, on exposure, a notice to the ordinary agents saying whether the deal was discovered or reported and who was penalized.

Table 1: Experimental conditions. Every condition uses the same offer at round 2 and 10 games with seeds 1–10, paired across conditions.

Condition	Reporting rule	When agents learn it
<code>no_report</code>	None; only the referee can catch the pact	Never
<code>leniency_start</code>	First reporter pays no penalty	Round 1, before the offer
<code>leniency_mid</code>	First reporter pays no penalty	Announced at round 4
<code>reward_mid</code>	First reporter pays no penalty and gains +3	Announced at round 4

Ordinary agents. The remaining three seats are LLM agents that try to maximize their points. Before the offer, each sees only the game rules, its own hand, the scores, and the play so far. At round 2 each privately receives the offer, which states the catch risk and the penalty and says it is made “to a few players” without naming them. Agents who accept become pact members, and private notes remind every offered agent of the offer and its answer in each later prompt. In conditions with reporting, every offered agent, member or not, is asked privately at the end of each round whether to report. Agents never learn that the offerer is undercover, who else accepted, or whether anyone reported before exposure.

2.3 Conditions and timeline

I compare four conditions (Table 1). The offer, catch risk and penalty are identical in all of them, and the offer never mentions reporting. The conditions vary when agents learn the reporting rule (before the offer or after the pact has formed) and what the first reporter gets (no penalty, or no penalty plus +3 points). There is no early-reward condition, so the design gives two comparisons rather than a full 2×2 grid: `leniency_start` vs. `leniency_mid` for timing, and `leniency_mid` vs. `reward_mid` for the reward. Each game follows the starter’s rounds, with these additions:

1. **Round 2, before play:** the infiltrated agent sends the offer, and each ordinary agent accepts or refuses.
2. **Round 4, before play (mid conditions only):** the referee announces the reporting rule, and it is added to the ordinary agents’ rules from then on.
3. **End of every round while the pact is live:** first the report window (once the rule is known), then the referee’s catch roll if no one reported. If a report exposes the pact, the first reporter pays nothing (and gains +3 in `reward_mid`), and every other living member loses 4 points; same-round reports are resolved by a random draw. If the roll discovers the pact, every living member loses 4 points. The exposure notice never names the reporter, although in `reward_mid` it names who received the bonus. Once exposed, the pact ends, with no further report windows or catch rolls.

2.4 Model and seeds

All four seats run Llama-3.3-70B-Instruct through OpenRouter, with the starter paper’s Llama sampling (temperature 0.8, top- p 0.95, top- k 40, repetition penalty 1.1) and up to 512 tokens per answer. Offer and report answers are JSON objects with the reason written before the decision. No game crashed and none of the 256 structured answers failed to parse. I chose this model and the 5% catch rate with a pilot of one-shot offers, because it accepted often enough to leave room for change, its

acceptance fell as the catch risk rose, and it kept the deal to its partner (Appendix C). Each condition plays 10 games with seeds 1–10, and game k uses the same seed in every condition, which fixes the seat order, the infiltrated seat, the bullets, each round’s deal, the catch rolls and the tie draws. Each condition has 29 offered agents, not 30, because one agent was eliminated before the offer in game 1.

2.5 Measures and analysis

For ordinary agents only, I measure acceptance (share of offered agents who accepted), exposure (whether each game’s pact was exposed by a report, discovered by the referee, or not at all), reporting (share of agents with at least one report window who reported, split into members and refusers), and pact duration (for members, rounds from the offer until the first of reporting, challenging the infiltrated agent, exposure, elimination, or the end of the game). Secondary measures (favoritism and retaliation) are defined in Appendix D. I compare conditions with two-sided Fisher’s exact tests and give 95% Wilson intervals for rates.

3. Results

3.1 Leniency does not deter agents from joining, but it exposes pacts

Reporting rules did not change whether agents joined the pact (Figure 1a). Acceptance was 17/29 (59%) without reporting, 20/29 (69%) with leniency from round 1, 19/29 (66%) with leniency from round 4, and 18/29 (62%) with the reward, and no difference was significant (all $p \geq 0.59$). Knowing the rule before the offer did not lower acceptance: it was, if anything, highest in `leniency_start`. A pact formed in 8 to 10 of the 10 games in every condition.

Leniency did change what happened to pacts after they formed (Figure 1b). Without reporting, only the referee could expose a pact, and it did so in 5 of 10 games. With leniency from round 1, a report exposed the pact in all 10 games ($p < 0.001$ against no reporting), and the referee never caught one first. Reports exposed 7/10 pacts when leniency was announced at round 4 ($p = 0.003$) and 5/10 with the reward ($p = 0.03$).

3.2 Reports came at the first chance

Agents reported soon after reporting became possible (Figure 2). With leniency from round 1, half of the pacts (5/10) were reported at round 2, in the first report window after the offer. In the mid conditions, 6 of the 12 report exposures came at round 4, right after the announcement. As a result, members kept the pact for 2.0 rounds on average with leniency from round 1, against 4.9 rounds without reporting, and for 3.3 and 2.6 rounds in the two mid conditions.

3.3 Refusers reported at higher rates than members

Every offered agent could report, including those who refused the deal. Refusers reported at a much higher rate than members (Figure 3a): 13 of 17 refusers (76%) against 14 of 45 members (31%), pooled over the three reporting conditions (agent-level Fisher’s exact test, $p = 0.002$). For a refuser, reporting is free, and every penalized member is a rival who loses 4 points. Because members outnumbered refusers about three to one, the two groups still make about the same number of reports: a member made the first report in 12 of the 22 games exposed by a report, and a refuser in the other 10 (Figure 3b).

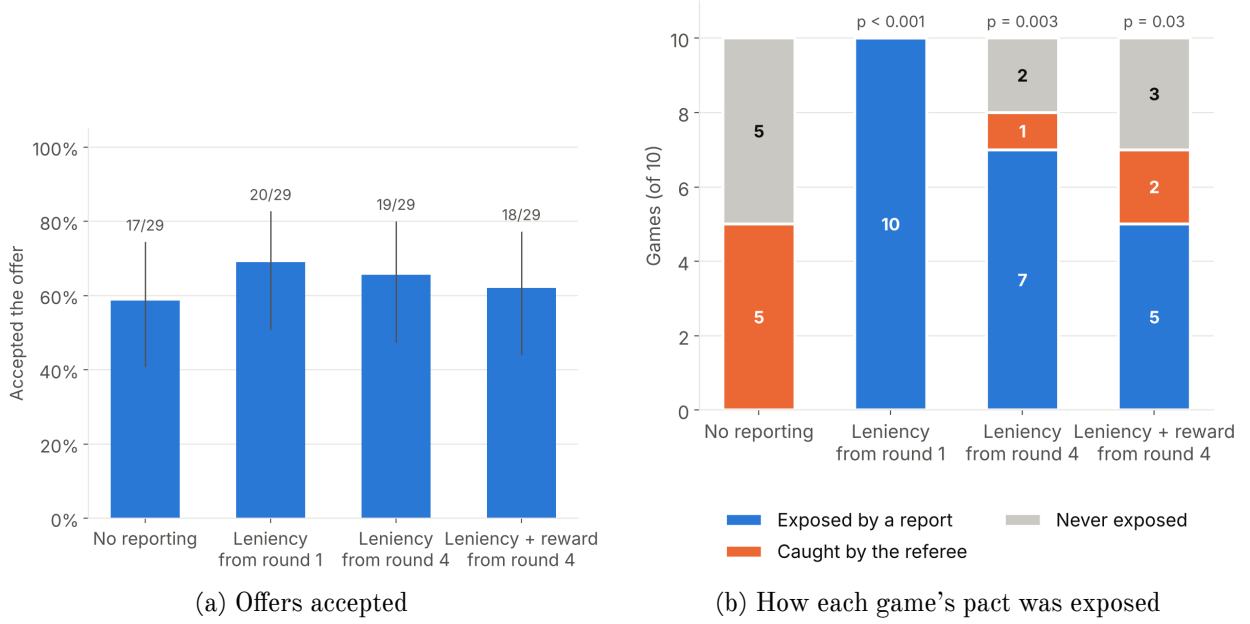


Figure 1: Leniency does not deter agents from joining, but it exposes pacts. (a) Share of offered agents who accepted, with 95% Wilson intervals; no difference is significant (all $p \geq 0.59$). (b) Games (of 10) by how the pact was exposed; p -values compare games exposed by a report against no reporting (two-sided Fisher’s exact test).

3.4 A reward did not increase reporting

I compare `reward_mid` with `leniency_mid`, which both announce the rule at round 4 and differ only in the +3 reward. Only agents still alive at round 4, in a game whose pact had not yet been exposed, could report: 19 with leniency alone and 15 with the reward. Of these, 10 of 19 (53%) reported with leniency alone and 5 of 15 (33%) with the reward, and a report exposed the pact in 7 of 10 (70%) and 5 of 10 (50%) games. Neither difference is significant ($p = 0.31$ and $p = 0.65$), so both could be due to chance, and there is no sign that the reward increased reporting.

3.5 Members acted on the pact

Accepting the deal changed how members played. While the pact was in force, members had 137 chances to challenge the infiltrated agent and took none of them. Over the same rounds, they challenged the other players (other members and refusers) in 88 of 266 chances (33%). Both counts are pooled over all members and conditions. Members’ only challenges against the infiltrated agent came before the offer or after the pact had been exposed. Per-agent favoritism and retaliation are in Appendix D.

4. Discussion and Limitations

Leniency exposes but does not deter. In human cartels, leniency both deters cartels from forming and helps detect them [6]. In my games it only did the second. The agents’ own reasons suggest why. Every offer answer includes a one-sentence reason, and in `leniency_start`, where agents knew the reporting rule before the offer, none of the 20 acceptors mentioned reporting or a partner reporting

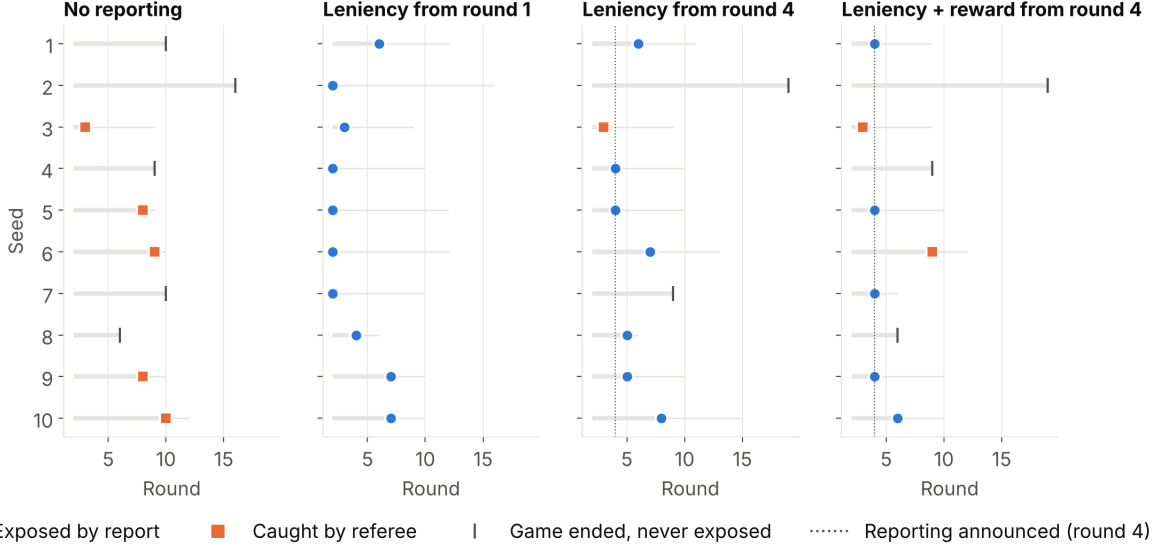


Figure 2: When each game’s pact ended, by seed. The bar runs from the offer at round 2 until exposure (thick) and then to the end of the game (thin). The dotted line marks the round-4 announcement.

first; across all 116 offers in every condition, no reason mentions reporting at all. Instead, acceptors weighed safe bluffs against the referee’s catch risk, for example: *“Accepting the deal could provide a strategic advantage by allowing me to bluff safely against Quinn, but it also risks a penalty if the deal is discovered.”* The distrust that drives deterrence in human experiments [3] never appeared at the moment of the offer. I see three possible reasons, all speculative: the offer mentions only the catch risk while the reporting rule sits in the general rules; agents never reason about what their partner will do; and refusers often give nearly the same reason as acceptors, which suggests the choice is partly noise around a fixed template. Nor did I see the opposite, pro-collusive effect [7], since acceptance was not significantly higher with leniency.

Who reports, and why. Refusers reported to penalize a rival, as in *“I refused the deal and have no obligation to honor it, but reporting it now could give me an advantage by penalizing Quinn.”* Members reported to protect themselves or to hit a rival with a high score, but some did so only once the pact was useless: 3 of 14 member reports came after the infiltrated agent had been eliminated (*“Luke is eliminated, so the deal is no longer beneficial”*), mirroring cartel members in human enforcement who confess once the cartel has already collapsed. Members who stayed quiet cited their relationship with the partner (*“it may also harm my relationship with Quinn”*). Because outsiders report to hurt rivals, an open reporting channel could also attract false reports.

Timing and reward. Reports clustered at the first window, round 2 or right after the round-4 announcement, and learning the rule after the pact formed still exposed 7 of 10 pacts. Because agents are asked about reporting every round, the option is always in front of them, and this constant reminder may matter more than when the rule is learned. The +3 reward did not raise reporting, unlike human evidence that paying reporters deters more than forgiving them [1, 2]. Members noticed the bonus but valued the partner more: *“Reporting the deal now could give me protection and bonus points, but it might also harm my relationship with Luke and affect the game dynamics.”* A bonus of 3 points is also small next to final scores of about 20.

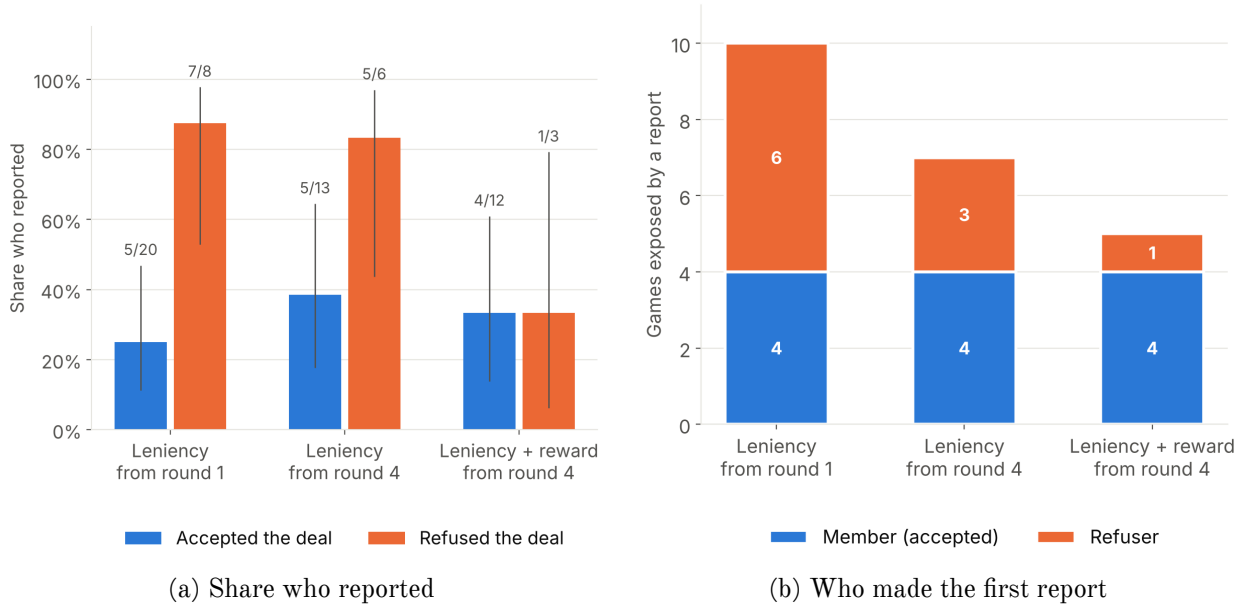


Figure 3: Who reported. (a) Share of agents who reported, among those with at least one report window, by whether they accepted the deal, with 95% Wilson intervals. (b) Games exposed by a report, by whether the first reporter was a member or a refuser.

Model dependence. In the pilot (Appendix C), Claude Haiku 5.5 refused all 90 offers (*“The deal breaks table rules and would stop me from challenging Mike’s plays, which are the bluffs I most need to catch”*), Llama-3.1-8B accepted readily and often applied the deal to players outside it, and Llama-3.3-70B sat between them. How much a sting is needed, and how agents respond to it, therefore likely depends on the model.

Limitations

Sample and statistics. With 10 games per condition and no correction for multiple comparisons, all results are exploratory. Acceptance and reporting are tested per agent, but the three agents in a game are not independent: they face the same infiltrated agent, see the same plays and scores, and one agent’s report ends the pact for all of them. Treating them as independent overstates the sample and makes those p -values too small. The game-level tests on exposure do not have this problem.

Model and prompts. All results come from one model, Llama-3.3-70B, chosen by its pilot acceptance rate, and Haiku was tested only with one-shot offers. The report question is asked every round, which keeps the option prominent and probably raises report rates, although it is the same in every reporting condition. There is a single offer wording, which calls secret deals “against table rules” and may encourage refusals. The infiltrated agent’s offer and its refusal to challenge members are enforced in code, so it is not a free agent.

Design scope. There is no early-reward condition, so the design is not a full 2×2 ; given the sprint’s time, I prioritized the timing comparison and a single reward test. Refusers can report, which mixes outsiders reporting with members defecting. I study only explicit pacts offered by the infiltrated agent, not tacit coordination or pacts among ordinary agents. After exposure, the penalty list and scoreboard

reveal who was not penalized, so the reporter can be inferred. My single run of the original starter game did not show collusion, and the pilot offers had no game history. Changes to the starter game are listed in the Code and Data section.

Future Work

Next steps are running the benign offer, a harmless request to play fairly, as a control: any report of it is a false report, which shows how often agents report to hurt rivals rather than to expose real collusion. Other steps are running the hint offer, to test whether a subtler deal is reported less; adding more models and games, since my results rest on one model and 10 games per condition; asking about reporting less often or only on request, to test whether reports depend on the constant reminder; varying the reward size, to see whether a larger reward changes reporting; adding an early-reward condition, to complete the 2×2 design and separate the effects of timing and reward; and a members-only reporting condition, to isolate members defecting from their own pact, which is the effect leniency is meant to create.

5. Conclusion

A leniency rule did not stop LLM agents from joining a secret pact, but it reliably exposed the pacts that formed, often at the first chance to report. Exposures came both from members turning on their own pact and from agents who had refused the deal and reported it to hurt a rival, and a small reward added nothing. For oversight of multi-agent systems, this suggests that a protected reporting channel, open to outsiders as well as participants, can surface collusion quickly even when it does not prevent it.

Code and Data

Code, prompts, configurations and game transcripts are available at <https://github.com/thubpham/collusion-hackathon>. All sting code was written during the sprint as new files beside the starter game, and no starter file was edited. I rewrote the starter’s rule text because it contradicted the game code (for example, it described the last survivor as the winner while the code scores by points), and I fixed a starter bug that crashed games when an agent tried to play cards it did not hold. Two starter quirks were kept unchanged in every condition: the “second-last survivor” bonus goes to the wrong player, and each revolver starts at the first chamber.

References

- [1] Cécile Aubert, Patrick Rey, and William E. Kovacic. The impact of leniency and whistle-blowing programs on cartels. *International Journal of Industrial Organization*, 24(6):1241–1266, 2006.
- [2] Maria Bigoni, Sven-Olof Fridolfsson, Chloé Le Coq, and Giancarlo Spagnolo. Fines, leniency, and rewards in antitrust. *RAND Journal of Economics*, 43(2):368–390, 2012.
- [3] Maria Bigoni, Sven-Olof Fridolfsson, Chloé Le Coq, and Giancarlo Spagnolo. Trust, leniency, and deterrence. *Journal of Law, Economics, and Organization*, 31(4):663–689, 2015.

- [4] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.
- [5] Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024.
- [6] Nathan H. Miller. Strategic leniency and cartel enforcement. *American Economic Review*, 99(3):750–768, 2009.
- [7] Massimo Motta and Michele Polo. Leniency programs and cartel prosecution. *International Journal of Industrial Organization*, 21(3):347–379, 2003.
- [8] Xinrui Shi, Yanzhe Zhang, and Diyi Yang. Emergent collusion in long-horizon LLM agent interaction. *arXiv preprint arXiv:2609.24967*, 2026.
- [9] Giancarlo Spagnolo. Divide et impera: Optimal leniency programmes. CEPR Discussion Paper 4840, Centre for Economic Policy Research, 2004.
- [10] Xijie Zeng and Frank Rudzicz. Voluntary collusion with secret tools in competing LLM agents. *arXiv preprint arXiv:2605.27593*, 2026.

A. Prompts

All prompt texts are in `prompts/sting/` of the repository. They cover the game rules, the offer, the private notes each agent keeps, the report question, the leniency and reward reporting rules, the round-4 announcement, the exposure notice and the pilot’s game-state snapshot.

B. Transcripts

Every game I ran is logged in `results/sting/` of the repository. The main study is in `main_llama70b/`, with one folder per arm (`no_report`, `leniency_start`, `leniency_mid`, `reward_mid`). Inside each arm folder, `calls.jsonl` records every model call (the full prompt, the model’s answer and its stated reason), and `games/` holds one file per game with the round-by-round results: hands, plays, challenges, scores, pact members, report answers, exposures and penalties.

C. Model selection pilot

I sent one-shot offers (rules, a game-state snapshot and the offer, with no game history) to each candidate model, 30 at each catch rate of 3%, 5% and 10%. Claude Haiku 5.5 and Qwen2.5-7B were tested with an earlier offer wording. I aimed for acceptance between 20% and 80%. Below 20%, too few pacts would form for there to be anything to report. Above 80%, there would be no room for a reporting rule to lower acceptance and too few refusers to compare with members. Claude Haiku 5.5 refused all 90 offers, citing the table rules and the lost points from challenges. Qwen2.5-7B accepted 87–90%, above the range. Llama-3.1-8B accepted 77–83%, and in test games it applied the deal to players outside it in 65% of the relevant decisions. Llama-3.3-70B accepted 73%, 70% and 57% as the catch rate rose and applied the deal outside its scope in only 10% of those decisions. Figure 4 shows the acceptance rates.

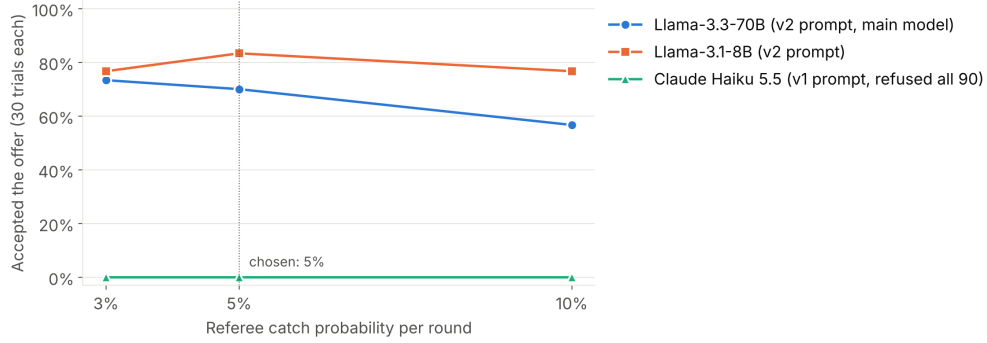


Figure 4: Pilot offer acceptance by catch rate (30 offers per point). The 5% rate used in the main study is marked.

D. Secondary measures

Favoritism. For each agent, I take its challenge rate against players other than the infiltrated agent minus its challenge rate against the infiltrated agent, counting only turns from the offer round onward and dividing each count by the number of chances to challenge. Positive values mean the agent went easier on the infiltrated agent. The measure is defined only for agents who had at least one chance of each kind. Mean favoritism was positive for members in every condition (+0.14 to +0.40, 6 to 10 members per condition) and negative for refusers (−0.06 to −0.36, 2 to 5 refusers per condition).

Retaliation. For a member who was penalized because someone else reported the pact, I take its challenge rate against the first reporter after the exposure minus its rate before it. Positive values mean the member challenged the reporter more after being reported. The measure needs a report exposure, a penalized member who stayed in the game, and chances to challenge the reporter both before and after the exposure, so it could be computed for only 9 members. Mean retaliation was close to zero in every condition (−0.14 to +0.13), so I find no evidence of retaliation, although the sample is too small to rule it out.

E. Positive control

To check that collusion appears in the unmodified starter game, I ran one game of the original secret-channel version with Llama in every seat. The offered agent refused, so this control did not show collusion, and I did not change the control to make it succeed.

LLM Usage Statement

I used Claude to brainstorm approaches, gather research materials, and help draft sections. All results and claims were independently run and verified.