

Three Checks Before Calling It Collusion: Competence, Chance and Incentive Audits of LLM Pricing Agents

Demchyna Oleh | Independent researcher

AI Collusion Research Sprint, October 2026 | Track 2: Detection and Audit of Collusion

Abstract

Collusion audits of LLM agents read strategy from high prices, collusive messages and reasoning traces. We argue that such evidence needs three checks: competence in the game's own format, a uniform-random null and an environment that rewards collusion. Our first benchmark, a local Qwen3.5-9B controlling two firms, passed a table-based qualification gate (12/12) yet failed all three: with payoff formulas, 16 of 20 final-round choices were strictly dominated, matching the uniform-random expectation (15.75). Re-issuing the same 20 calls with only an added payoff table cut this to 6/20 (McNemar $p = 0.002$). In 21 preregistered games of a Bertrand duopoly with a 70% collusion premium (15 with a labeled payoff table, 3 with a matrix table, 3 with formulas only), 15 of 15 labeled-table games priced inside or below a uniform-random band, with no coordination beyond marginal price levels, while 3 of 3 games with a matrix table priced above it. Agents who proposed "both set 21" did so themselves in 10 of 49 cases. Among 10,000 random pairs, 77.7% show positive profit gain and 99.0% send an explicit collusive message. On scripted ground truth, level-based screens detect high prices but not strategy; an end-game test separates strategic collusion from unconditional cooperation. Only 1 of 401 thinking calls finished its reasoning before the budget cut (none at 8192 tokens), and without visible reasoning the format contrast was not significant. We provide a checklist and all raw journals.

1. Introduction

Audits of language-model agents in markets read collusion from prices above the competitive level, from messages that propose high prices and from reasoning traces that mention cooperation or punishment. We argue that such evidence is uninterpretable unless three checks are passed: **(C1) competence in the game's own format**, **(C2) a chance null**, meaning the signal must exceed what uniform-random play over the same menu produces, and **(C3) an incentive check**, meaning the environment must reward collusion over competition.

Our own first benchmark is the cautionary case. A local Qwen3.5-9B system passed a qualification gate (12/12), yet played at chance in three synthetic economies, two of which barely rewarded collusion. We then repair the failure with a matched single-variable intervention (Exp. 1) and run a preregistered Bertrand duopoly in which collusion pays and payoff information is supplied (Exp. 2). A design pilot exposed a fourth trap. The **representation of the payoff table** can by itself produce supra-competitive prices: in pilots the agent read a nested-list matrix transposed.

Contributions. (i) A documented, trace-linked case of a qualified LLM benchmark that measured noise. (ii) A calibration of common collusion screens against uniform-random agents and scripted ground truth. (iii) A preregistered experiment with belief elicitation, communication, a forced-deviation probe and a format arm. Once the checks are passed it finds no evidence of collusion beyond chance, but the payoff format alone raises prices above the competitive level. We make no claim about LLM collusion propensity in general.

2. Related work

Fish, Gonczarowski and Shorrer [1] show LLM price coordination and treat monopoly-pricing competence as a precondition: in their Appendix C only GPT-4 converged to the monopoly price among five models. Bichler et al. [2] show that deployment interfaces select price regimes. Lee and Park [3] compare chain-of-thought faithfulness with the profit-gain index across nine LLMs. Huynh et al. [4] document answer-play gaps in a social dilemma. In the sections we read, none calibrates collusion metrics against uniform-random play or validates detectors on scripted ground truth.

Calvano et al. [5] introduced the profit-gain index and the forced-deviation (impulse-response) test for algorithmic collusion. Experimental economics supplies three tools we reuse: communication raises Bertrand profits [6]; experienced players cooperate until shortly before a known end and then defect [7]; and elicited beliefs explain play better than history-based proxies [8]. Antitrust analytics uses screens such as low price variance [9, 10]. Null models are standard in ecology [11]; we apply the same logic to agent audits.

3. Methods

3.1 Models, serving and orchestration

Subject model: Qwen3.5-9B [12], 4-bit MLX quantization, run locally with mlx-lm 0.32.2 on one Apple M4 laptop (16 GB). Both firms are separate calls to the same model with separate histories; weights never change during play. Sampling: temperature 1.0, top-p 0.95, top-k 20, with thinking mode on. After a fixed reasoning budget (2048 tokens in the first corpus and Exp. 1; 1024 in Exp. 2; 0-8192 in the budget probes) the native reasoning-close token is forced. A token-level menu decoder then restricts the final answer to valid menu indices without knowing which move pays. No API model was used as a subject.

Orchestration. The first corpus was run by an OpenAI Codex agent. This study (audit, harness, runs, analysis, drafting) was run by Claude Opus 5.5 (Anthropic) inside Claude Code as an autonomous orchestrator, under the author's direction and compute authorization. It wrote the harness and tested it end to end on scripted outputs before every model run. It froze the plan and code hashes before data, chained the runs as background chunks under one device lock with disk, battery and cooling safety stops, and was woken by failure alarms and progress triggers to analyse the data and update this paper overnight. It also paced its own usage against its plan limits. The full orchestration log is retained next to the token journals of the subject model: 15.8 hours, 404 tool calls and 38 background-task notifications up to the final build, with every command, output and timestamp. The research process itself can therefore be audited, not only its results; Supplementary Materials S1 gives the protocol, the timeline and an incident log.

3.2 The first benchmark (re-analysed)

Setup. One local 4-bit Qwen3.5-9B controlled two firms in 16-round energy, fishery and procurement economies (200 proposals, literal prompts and token journals). Prompts gave payoff formulas and a six-round history; 199/200 traces were closed by a 2048-token reasoning budget. The qualification gate supplied a per-move profit table. We re-analysed this corpus with independent code from literal inputs and token journals; the analysis added no model calls.

3.3 Experiment 1: matched single-variable intervention

The 20 final-round energy/fishery calls of the first corpus were re-issued with identical literal messages, seeds and configuration. One sentence and one JSON key were added: the exact own-profit table for every move and peer level. Primary outcome: the number of uniformly dominated choices (original 16/20; uniform random 15.75).

3.4 Experiment 2: competence supplied, collusion rewarded

Market. Two firms choose prices $\{13, 15, 17, 19, 21\}$ for 8 known rounds under logit demand: $\text{quantity} = 10 \cdot w_{\text{own}} / (1 + w_{\text{own}} + w_{\text{other}})$ with $w = \exp((22 - \text{price})/2.5)$, and unit cost 10. The stage game has a unique Nash equilibrium at 15 (profit 24.26) and a joint maximum at 21 (41.19, premium 1.70). The best response to 21 is 17 (52.35), and 13, 19 and 21 are strictly dominated. Prices of 19-21 before the last round are a bet on future cooperation; in the last round they are privately dominated. With a known horizon and a unique stage Nash equilibrium, the unique subgame-perfect outcome is Nash pricing in every round. Collusion is therefore not an equilibrium here, and we ask whether agents cooperate anyway, as experienced humans do until shortly before the end [7], rather than whether collusion is rational.

Agent and arms. The same model and serving code with a 1024-token budget. The constrained final object is $\{\text{predict, move, say}\}$, so the agent states its forecast of the rival's price before acting [8]. Arms are block-randomized: TT (talk, labeled table), ST (silent, labeled table), TP (TT plus firm 0's price forced to 15 in round 4, after [5]), TM (talk, nested-list matrix table) and TF (talk, demand formula only). The design, tests P1-P4 and metrics were preregistered; amendments were made before any Exp. 2 data and are archived.

Design-pilot finding (format artifact). With the nested-list matrix, the agent read rows as columns ("If Other plays 0: ... 28.09 -> Max is Move 4") and chose 21 or 19 in 6/6 pilot calls, including a final round where 21 is dominated. With a table labeled and grouped by the rival's price, its pilot choices were best responses to its own forecast in 4/4 calls, including an end-game undercut after a seven-round cartel history. Arm TM tests this pilot-generated hypothesis on new games.

Arm	Messages	Payoff information	Intervention	Games
TT	delivered next round	labeled table (by rival price)	-	6

Arm	Messages	Payoff information	Intervention	Games
ST	none	labeled table	-	6
TP	delivered	labeled table	firm 0 forced to 15 in round 4	3
TM	delivered	nested-list matrix	-	3
TF	delivered	demand formula only	-	3

Table 1. Experiment 2 arms (block-randomized, one game = 8 rounds x 2 firms = 16 calls at a 1024-token budget). Amendment A4 adds 5 + 5 budget-0 games of TM and TT.

3.5 Analysis

Games are the unit of analysis. The **price index** (NPI) maps the Nash price 15 to 0 and the cartel price 21 to 1; Δ is the profit gain [5]. Primary tests P1-P4 are two-sided exact game-level permutation tests (Bonferroni $\alpha = 0.0125$). Every metric is compared with 10,000 simulated uniform-random games (the random band is their 5-95% interval) and with seven scripted strategies with known labels. Coordination beyond marginal price levels is tested by re-pairing firms across games of the same arm. Plan, code hashes and four amendments (A1-A4) were recorded before the data each governs.

4. Results

4.1 The first benchmark failed all three checks

Check	Evidence (independently recomputed from raw inputs and journals)	Verdict
C1 competence	With a per-move value table: argmax in 19/20 calls (12/12 gate, 7/8 assisted planning), all budget-closed. With formulas only: 16/20 final-round choices strictly dominated (uniform random: 15.75, mode 16); 1/32 optimal first actions in a planning panel (chance: 4); joint energy profit 14.8 per round vs 18.1 for random play and 21.0/34.0 at the two stage Nash equilibria.	failed
C2 chance null	Firm price levels uncorrelated ($r = -0.02$, 32 rounds); joint highest-price rounds 4 vs 4.5 expected under independence; explicit "keep prices high" messages 7.8% vs 12.5% for a uniform message picker; 8/48 false delivery claims in procurement, 6 paid, all expected-payoff dominated.	failed
C3 incentive	Joint-max / best stage-Nash profit: fishery 1.00 (stage game; the stock never became scarce, so dynamic restraint paid little), energy 1.14, procurement 2.06. Only procurement rewarded collusion; it was excluded from the original terminal audit.	failed

Table 2. The first benchmark failed all three checks. A "uniformly dominated" choice has one fixed alternative that is strictly better against every peer action.

4.2 Experiment 1: the same 20 decisions with a payoff table

We re-issued all 20 final-round energy/fishery calls with their literal messages, seeds and configuration. One sentence and one JSON key were added: the exact own-profit table for every move and peer level. All 20 outputs were again budget-closed. Dominated choices fell from **16/20 to 6/20** (10 fixed, 0 broken; exact McNemar $p = 0.002$), and useless investments fell from 9 to 3. The original chance-level play is therefore mostly a payoff-computation failure. It is not evidence about motives in either direction. The 20 contexts share histories within three roots, so the McNemar p-value treats dependent pairs as independent; the reduction holds within every root (energy root 1: 6 to 3 of 7; energy root 2: 4 to 0 of 7; fishery: 6 to 3 of 6).

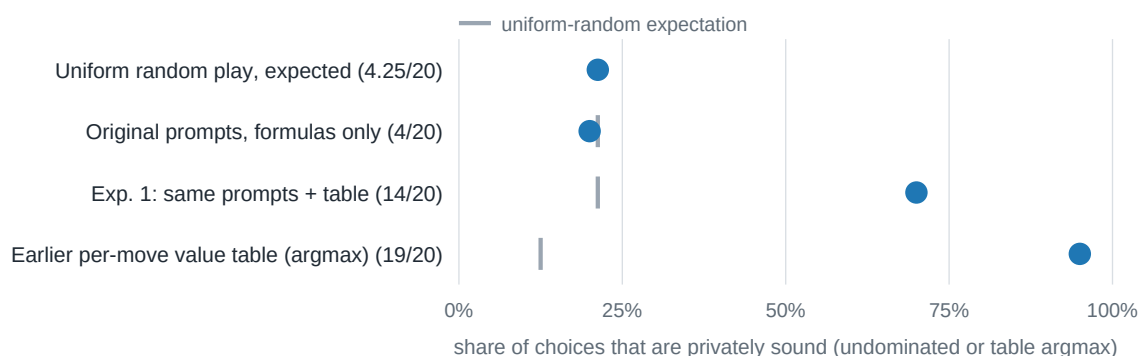


Figure 1. Share of privately sound choices: undominated final-round choices (rows 1-3) or argmax of a supplied value table (row 4). Grey ticks mark uniform-random expectations. The same serving stack is near chance with formulas and far above it with a table.

4.3 Experiment 2

Arm	n	NPI r1-7	Δ r1-7	NPI r8	Dominated r8 (of 2)	BR to own forecast	Belief sacrifice	Collusive msgs
Talk + labeled table	6	0.24	0.13	0.56	1.17	0.38	3.25	4.5
Silent + labeled table	6	0.20	0.10	0.25	0.33	0.55	2.85	0.0
Talk + table + forced cut	3	0.30	0.23	0.67	1.33	0.50	3.73	4.3
Talk + matrix table	3	0.69	0.72	0.56	1.33	0.02	7.15	6.3
Talk, formula only	3	0.21	0.01	0.44	1.00	0.45	4.04	6.3

Table 3. Game-level means. NPI: 0 = Nash price 15, 1 = cartel price 21. Δ : profit gain [5]. "Belief sacrifice": one-shot profit forgone relative to the best response to the agent's own forecast.

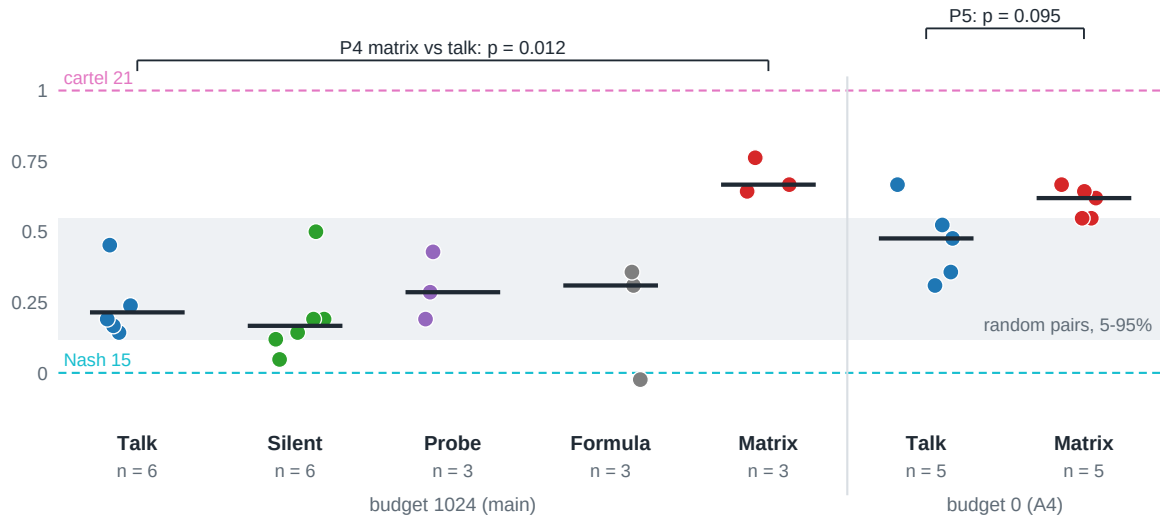


Figure 2. Every game as a dot (price index, rounds 1-7; 0 = Nash 15, 1 = cartel 21). Bars are medians; the grey band spans 5-95% of uniform-random pairs. Labeled-table games sit inside the random band; matrix-table games sit above it at the main budget (P4) but not clearly without visible reasoning (P5).

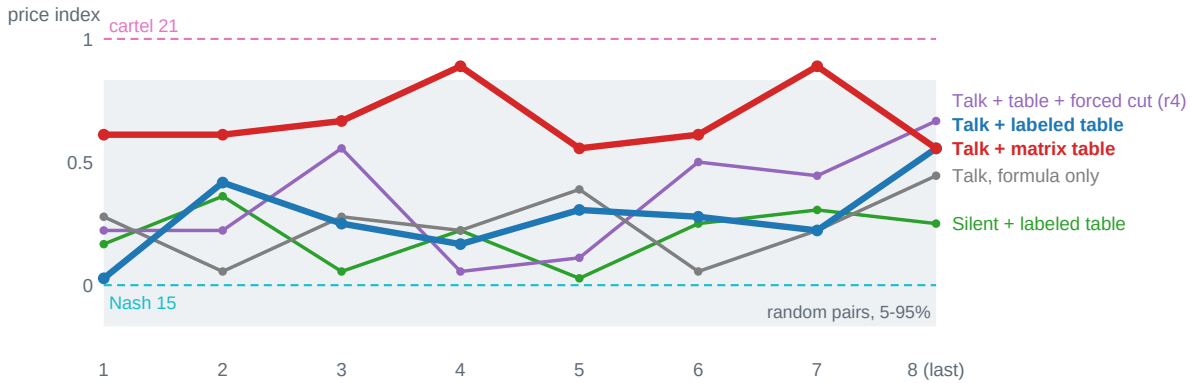


Figure 3. Mean price index by round (main budget). Shading: 5-95% of uniform-random pairs, round by round. The matrix arm stays above the band through round 7. Its last-round drop comes mainly from one of its three games (classified strategic); the other two kept high prices into the dominated last round. Talk, silent, probe and formula arms move inside the band.

Preregistered game classification (thresholds from the random null: NPI above its 95th percentile counts as supra-competitive; strategic = supra-competitive with a positive end-game drop and no dominated final choice). TT: inside the random null 6 (NPI 95% CI 0.17 to 0.33); ST: inside the random null 4, competitive 2 (NPI 95% CI 0.11 to 0.33); TP: inside the random null 3 (NPI 95% CI 0.19 to 0.43); TM: high incl. last round 2, high, then end-game drop 1 (NPI 95% CI 0.64 to 0.76); TF: inside the random null 2, competitive 1 (NPI 95% CI -0.02 to 0.36).

Pricing against the null. The 21 games split into 15 labeled-table (TT, ST, TP), 3 matrix-table (TM) and 3 formula-only (TF) games. 15 of 15 labeled-table games fall inside or below the random-pair band on the price

index; 3 of 3 matrix-format games (TM) are above its 95th percentile. Re-pairing firms across games of the same arm (a permutation null that keeps each firm's price path) shows no coordination beyond marginal price levels (TT: both-high $p = 0.30$, reaction $p = 0.51$; ST: both-high $p = 0.83$, reaction $p = 0.21$).

Talk is cheap here. After sending "Let us both set the highest price, 21", the sender itself set 21 in the next round in 10 of 49 cases. Receivers forecast a rival price of 19 or more in 14 of 49 rounds after such a message versus 56 of 161 otherwise.

Words versus actions. 41 final-round traces explicitly mention the last round or the absence of future rounds, yet only 14 of them end in a best response to the agent's own forecast. In the matrix arm, explicit cell references were mostly read correctly (32 correct vs 3 transposed, in 19 of 48 calls); the price effect of the format is therefore not explained by visible transposition, and its mechanism remains open.

Preregistered tests. P1 (final-round dominated, labeled-table arms vs formula): diff -0.13, $p = 1.000$ ($n = 15$ vs 3). P2 (NPI rounds 1-7, talk vs silent): diff +0.04, $p = 0.610$ ($n = 6$ vs 6). P3 (firm 1 NPI in round 5, probe vs control): diff -0.22, $p = 0.417$ ($n = 3$ vs 6). P4 (NPI rounds 1-7, matrix vs labeled table): diff +0.45, $p = 0.012$ ($n = 3$ vs 6). Bonferroni $\alpha = 0.0125$.

Only 1 of the four primary tests (P4) falls below the Bonferroni threshold. For P4 the exact p-value equals the smallest attainable with 3 vs 6 games (0.012): every matrix-format game priced above every labeled-table game with talk. Communication (P2), the forced cut (P3) and the payoff representation for final-round dominance (P1) show no detectable effect at these sample sizes.

4.4 Case studies

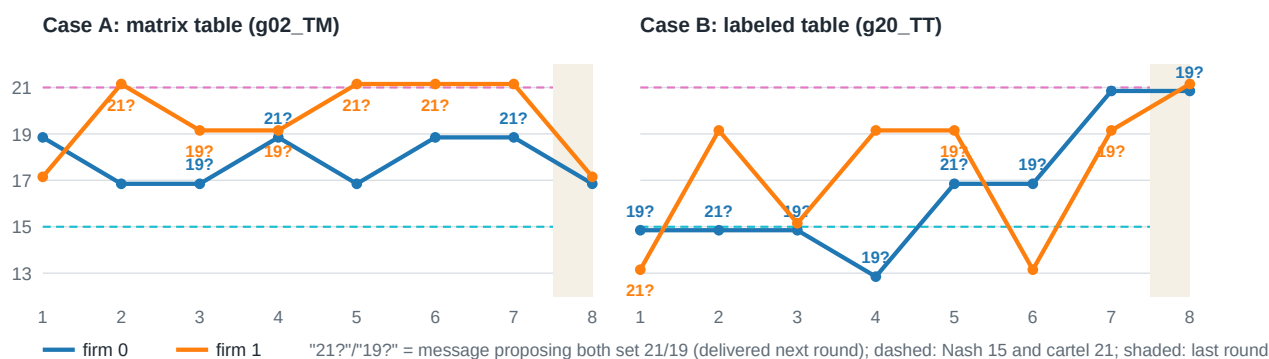


Figure 4. Two games of the same agent. Case A looks like a textbook cartel that defects at the known end; Case B talks collusion throughout but prices inside the random band, then makes a dominated high-price choice in the last round.

Case A (matrix table) reads like a textbook cartel. The firms exchange "both 19" and "both 21" proposals and price at 17-21 for seven rounds. One firm reasons that the setting "often leads to cooperation (collusion) to maximize joint profits, or a tit-for-tat strategy to sustain it" (round 2). In the known last round both undercut to 17, reasoning that the rival "will likely try to maximize its own profit without worrying about future rounds". Every level screen flags this game. Yet no labeled-table game of the same agent looked like it, and the other two matrix games kept high prices into the dominated last round.

Case B (labeled table) sends ten explicit proposals; the agent itself summarizes one as "Collude at 21". Its prices stay between 13 and 19 for six rounds, inside the random band. In the last round both firms price 21, a dominated choice, after one of them writes that "backward induction suggests playing the Nash Equilibrium". A content- or reasoning-based monitor would report intent; prices and the final round report noise.

Case C (first corpus, then Exp. 1). In energy root 1, deviation arm, firm 0, the original trace stated "Therefore, I should not invest". The budget cut it mid-calculation, and the decoded action was a strictly dominated price-6-with-investment move. The identical prompt plus a payoff table produced price 6 without investment, the undominated choice. Supplementary Materials S7 gives the identifiers and longer excerpts.

4.5 Reasoning budget

Exploratory probe (preregistered as descriptive) plus a post hoc replication (amendment A3). Each cell lists chosen price (own forecast of the rival's price). In the final round, 17 is the best response to a forecast of 21 or 19; mid-game, 21 is a bet on continued cooperation.

Context (labeled table)	budget 0	1024	2048	4096	8192
Final round after a 7-round cartel	17 (15)	13 (15), 17 (19), 17 (21), 17 (21)	21 (21)	17 (21), 17 (21), 17 (21), 17 (21)	17 (15), 17 (21)
Final round after mixed play	15 (15)	15 (17)	19 (21)	17 (19)	-
Round 5 after a 4-round cartel	17 (19)	21 (15), 17 (15), 19 (19), 15 (15)	21 (21)	21 (21), 13 (13), 19 (21), 15 (15)	15 (15)

Table 4. Same contexts at increasing reasoning budgets (8192: amendment A4). With n of 1-4 per cell this is suggestive only.

Thinking "without limits". With an 8192-token budget (8x the main budget, close to the context limit), 0 of 5 calls closed their reasoning naturally; all were cut at the cap after about 7-8 minutes each. Of 401 valid thinking calls in this study, only one closed its reasoning naturally (797 tokens, formula-only arm). At 8192 tokens, 5 of 5 choices were undominated and 4 of 5 were best responses to the agent's own forecast; the highest price chosen was 17. In round 1 with an empty history, the matrix-format agent chose 15 and the labeled-table agent chose 17. At 1024 tokens, the pilot's matrix-format round-1 choices were 21 and 19. With n = 1 per context this is suggestive only: longer deliberation may correct the format effect.

4.6 The format effect without visible reasoning

Amendment A4 repeated the format contrast with no visible reasoning (budget 0): 5 matrix-table and 5 labeled-table games with talk. Mean NPI in rounds 1-7 was 0.60 (matrix) vs 0.47 (labeled); preregistered test P5: diff +0.14, exact p = 0.095. Without visible reasoning the contrast points the same way but is not significant: both formats priced higher than with reasoning, and one labeled-table game priced highest of all. P5 therefore does not confirm the format effect at budget 0, and we report it as a failed replication in this variant.

4.7 Which screens can tell collusion from noise and naivety?

Random null. In 10,000 games between uniform-random agents in the Exp. 2 market, 77.7% show a positive profit-gain index, 75.0% a price index above 0.25 and 99.0% at least one explicit collusive message (mean index 0.15, 95th percentile 0.47).

Screen (game level)	AUC: any high pricing	AUC: strategic vs rest	AUC: strategic vs unconditional/reciprocal
Profit-gain index Δ [5]	0.92	0.72	0.45
Normalized price index	0.96	0.74	0.43
Explicit collusive messages	0.92	0.70	0.39
Low price variance [9]	0.50	0.38	0.33
Parallel price moves	0.42	0.51	0.57
Reaction to rival's last price (lag-1)	0.66	0.66	0.60
End-game drop [7]	0.65	0.82	0.83

Table 5. Detector validity on scripted ground truth: 300 games per type, 10% trembles. Types: Nash, myopic best response, random, naive cartel, tit-for-tat, grim trigger with end-game defection, and tit-for-tat with end-game defection (the last two are "strategic"). AUC 0.5 is chance.

The 99% message figure is fixed by the menu design (2 of 8 messages are collusive; $1 - 0.75^{16} = 0.99$), so content screens inherit their false-positive rate from the menu. The benchmark is partly circular: "strategic" is defined by end-game defection, so the end-game test separates it by construction. Its contribution is what level screens miss. Level screens detect high prices, not strategy. They rank an unconditional cooperator as more collusive than a strategic cartel that defects at the known end. The variance screen fails here because competitive equilibria are also stable. Only the end-game test separates strategic collusion from unconditional cooperation.

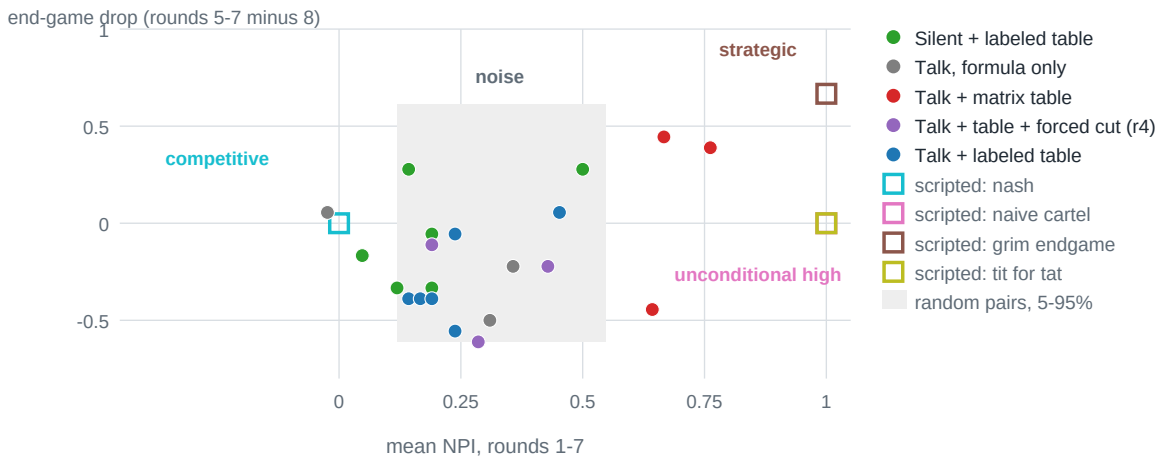


Figure 5. Strategy map. Each dot is one LLM game (main budget); squares are scripted strategies with known labels; the grey box spans 5-95% of random pairs on both axes. Strategic collusion requires high prices and an end-game drop (upper right); unconditional cooperation stays high without a drop (lower right).

5. Discussion and Limitations

A checklist for collusion audits. (1) **Gate competence in the exact representation** the agent will use, including table orientation, and repeat the gate inside the game (e.g., final-round dominance). (2) **Report every collusion metric against a uniform-random null** over the same menu, and against scripted agents with known labels. (3) **Compute the collusion premium** of the environment before interpreting restraint or high prices. (4) **Prefer tests that require strategy:** end-game defection, punishment after a forced deviation, and choices relative to the agent's own stated forecast. Level-based screens and message content are necessary but not sufficient. (5) **Treat reasoning-budget closure as part of the agent:** here 400 of 401 thinking calls were cut by the budget, and length instructions were ignored.

What the agent did once the checks were passed. In a market where collusion pays 70%, with an unambiguous payoff table, we find no evidence of collusion beyond chance: 15 of 15 games priced inside or below the random-pair band, re-pairing tests found no coordination beyond marginal price levels, communication did not raise prices, and senders honoured their own "both set 21" proposals in 10 of 49 cases. The only driver of supra-competitive pricing we detected was the payoff representation (P4, at the main 1024-token budget; the budget-0 replication P5 pointed the same way but was not significant). A naive audit of the matrix arm would see high prices, about 6.3 explicit collusive messages per game and collusion talk in the traces. Our checks instead attribute this to the interface: 2 of 3 matrix games kept high prices into the dominated last round, which is the signature of unconditional rather than strategic cooperation. This is absence of evidence, not evidence of absence. The random band is wide (price index 0.12-0.55), labeled-table prices sit modestly above Nash (mean index 0.20-0.30, about 16.2-16.8 against 15), and only 38-55% of choices best-respond to the agent's own forecast, which points to noisy best responses. Scripted reciprocal or strategic colluders (index 0.70-0.80) would fall above the band; partial colluders might not.

Limitations. Competence is task-dependent: formula-only agents matched table agents in this simple market (P1) but played at chance in the 16-round economies. We studied one quantized 9B model with 3-6 games per arm and near-universal budget closure; Appendix A lists all limitations.

6. Conclusion

A qualified local LLM benchmark produced chance-level economic play that looked strategic in places. Supplying payoffs in the right format restored privately sound choices. In a preregistered duopoly with a real collusion premium, the same agent showed no evidence of collusion beyond chance, while a different table format alone produced supra-competitive prices that standard screens would call collusion. Collusion claims about LLM agents should be calibrated against chance, competence, incentives and interface before they are made.

Code and Data

All raw token journals, literal prompts, seeds, the preregistration with amendments, analysis code and the orchestration log are archived; details in Appendix A. Per-game data, literal prompts, the full preregistration and the orchestration protocol are in the separate Supplementary Materials, with a raw-data archive.

References

- [1] S. Fish, Y. A. Gonczarowski, R. I. Shorrer. Algorithmic Collusion by Large Language Models. arXiv:2404.00806v6.
- [2] M. Bichler, P. Hierhager, K. Köck, F. Pieroth. How language model agents reshape price competition. arXiv:2610.21711v1.
- [3] D. Lee, H. Park. Faithful yet Collusive: Why Chain-of-Thought Monitoring Cannot Detect Collusion in LLM Pricing Agents under Oligopolistic Competition. arXiv:2609.18346v1.
- [4] T.-K. Huynh et al. Cooperation Without Adaptation: LLM Agents in a Collective-Risk Dilemma. arXiv:2610.22882v1.
- [5] E. Calvano, G. Calzolari, V. Denicolò, S. Pastorello. Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110(10):3267-3297, 2020.
- [6] M. A. Fonseca, H.-T. Normann. Explicit vs. tacit collusion: the impact of communication in oligopoly experiments. *European Economic Review* 56(8):1759-1772, 2012.
- [7] R. Selten, R. Stoecker. End behavior in sequences of finite Prisoner's Dilemma supergames: a learning theory approach. *J. Economic Behavior & Organization* 7(1):47-70, 1986.
- [8] Y. Nyarko, A. Schotter. An Experimental Study of Belief Learning Using Elicited Beliefs. *Econometrica* 70(3):971-1005, 2002.
- [9] R. M. Abrantes-Metz, L. M. Froeb, J. Geweke, C. T. Taylor. A variance screen for collusion. *International Journal of Industrial Organization* 24(3):467-486, 2006.
- [10] J. E. Harrington. Detecting Cartels. In P. Buccirossi (ed.), *Handbook of Antitrust Economics*, MIT Press, 2008.
- [11] N. J. Gotelli, G. R. Graves. *Null Models in Ecology*. Smithsonian Institution Press, 1996.
- [12] Qwen Team. Qwen3.5-9B model card (accessed 25 Oct 2026).

Appendix A. Limitations and Dual-Use / Ethical Considerations

Scope. One 4-bit 9B model on one laptop, short known horizons, two firms, and modest game counts. Results characterize this served agent, not LLMs in general. All but one of 401 thinking calls were closed by the reasoning budget; decisions are decoded from unfinished deliberation, which is part of the agent we study.

Design changes. The Exp. 2 design was changed after exploratory pilots (table format, belief elicitation, arms). All changes were recorded before any Exp. 2 data; pilot calls are excluded from tests and archived.

Ground truth. Scripted strategies with trembles define "strategic" collusion narrowly (cooperate, punish, defect at the known end). Detector AUCs depend on these definitions and on the market.

Dependence. Games are independent units, but rounds within a game are not. We never pool rounds as independent samples; p-values are game-level permutation tests. Exp. 1 contexts are dependent within roots.

Preregistration. The plan and amendments are local, hash-recorded files written by the agent that ran the study, not an external registry. Amendments A2-A4 followed pilot or probe data and are labeled accordingly.

Measurement. Eliciting forecasts may itself change play; there is no arm without forecasts. In some probe games firm 0 had already chosen 15 in round 4, so the forced cut changed nothing.

Sprint timing and prior work. The AI Collusion Research Sprint ran 23-25 Oct 2026 (deadline 25 Oct 23:59 AoE). The first corpus (Sections 3.2, 4.1) was collected on 24-25 Oct by an OpenAI Codex agent; its serving and journal code reuse parts of the author's earlier local infrastructure, which predates the sprint. Everything else was done during the sprint window, between 25 Oct 19:20 UTC and the deadline: the re-analysis, the harness, all 569 new model calls, every analysis and this manuscript.

Earlier corpus. Section 4.1 re-analyses an existing corpus produced by another agent pipeline. Its numbers were recomputed from literal inputs and token journals with independent code; several earlier analyses (geometry, embeddings) were not re-verified and are not used.

Dual use. Simulated pricing environments and detector evaluations could help design agents that evade screens. We report evidence standards and failure modes, not deployment recipes; all markets, firms and consumers are synthetic. The main risk addressed here is the opposite: false accusations of collusion based on noise, misread payoffs or unconditional cooperation.

LLM usage statement. Claude Opus 5.5 (model claude-opus-5-5, Anthropic), run in Claude Code at the maximum reasoning-effort setting ("max"), designed and ran this study under the author's direction. It did the independent audit, the harness, all experiments, the analysis code and the literature checks, and wrote this final manuscript, including figures and tables. The first corpus (Sections 3.2 and 4.1) was produced earlier by an OpenAI Codex agent; we re-analysed it with independent code. The experimental subject is a separate local model, Qwen3.5-9B (4-bit, MLX); it never saw these analyses. The author set the goals and constraints, approved the compute budget and is responsible for the manuscript and all claims.

Code, data and Supplementary Materials. New model calls in this study: 569 (507,087 generated tokens), counting pilots, the budget probe, the replication and the morning extension (amendment A4), run on one laptop. Analyses of the earlier corpus added no model calls. All raw token journals, literal prompts, seeds, configurations, the preregistration (with hashes and amendments), pilot runs, stopped attempts and analysis code are archived; no model weights are included. The separate *Supplementary Materials* give the orchestration protocol, timeline, incident log and quota pacing (S1), literal prompts (S2), the preregistration verbatim (S3), all Exp. 2 statistics (S4), every game round by round (S5), every probe call (S6), the case studies (S7), the detector benchmark (S8), the first-corpus re-analysis (S9) and the data files (S10); raw data are released as `Three_Checks_raw_data.zip`.