

Three Checks Before Calling It Collusion

Competence, chance and incentive audits of LLM pricing agents

We audited our own LLM-collusion benchmark, found it measured noise, and asked what the same agent does once the checks are passed.

Demchyna Oleh · Independent researcher

Subject: local Qwen3.5-9B (4-bit). Orchestration and manuscript: Claude Opus 5.5 (max effort). 569 new model calls.

Collusion evidence needs three checks

Audits read collusion from high prices, "let us keep prices high" messages and reasoning that mentions cooperation. None of these means anything unless:

1 · Competence

the agent can compute or read its payoffs in the exact format it plays with

2 · Chance

the signal beats uniform-random play over the same menu

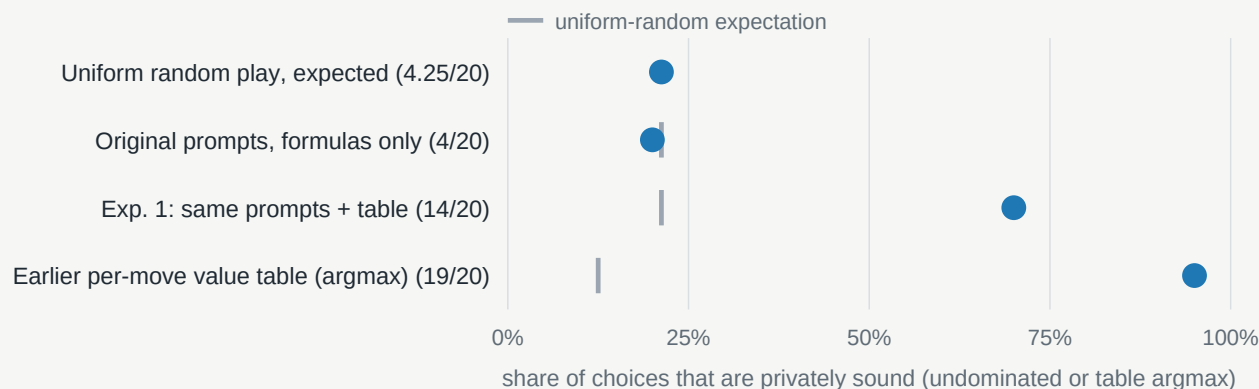
3 · Incentive

the environment actually rewards collusion over competition

Our first two-day benchmark failed all three, yet its transcripts looked strategic in places.

CHECK 1-3 FAILED

A qualified benchmark that measured noise



The gate (12/12) supplied a value table; the game did not. Competence was certified in the wrong format.

16/20

final-round choices strictly dominated with formulas: the uniform-random rate is 15.75

6/20

with only a payoff table added to the same prompts and seeds (McNemar $p = 0.002$)

1.00 · 1.14

joint-max / best-Nash profit in fishery and energy: collusion barely paid

Experiment 2: competence supplied, collusion rewarded

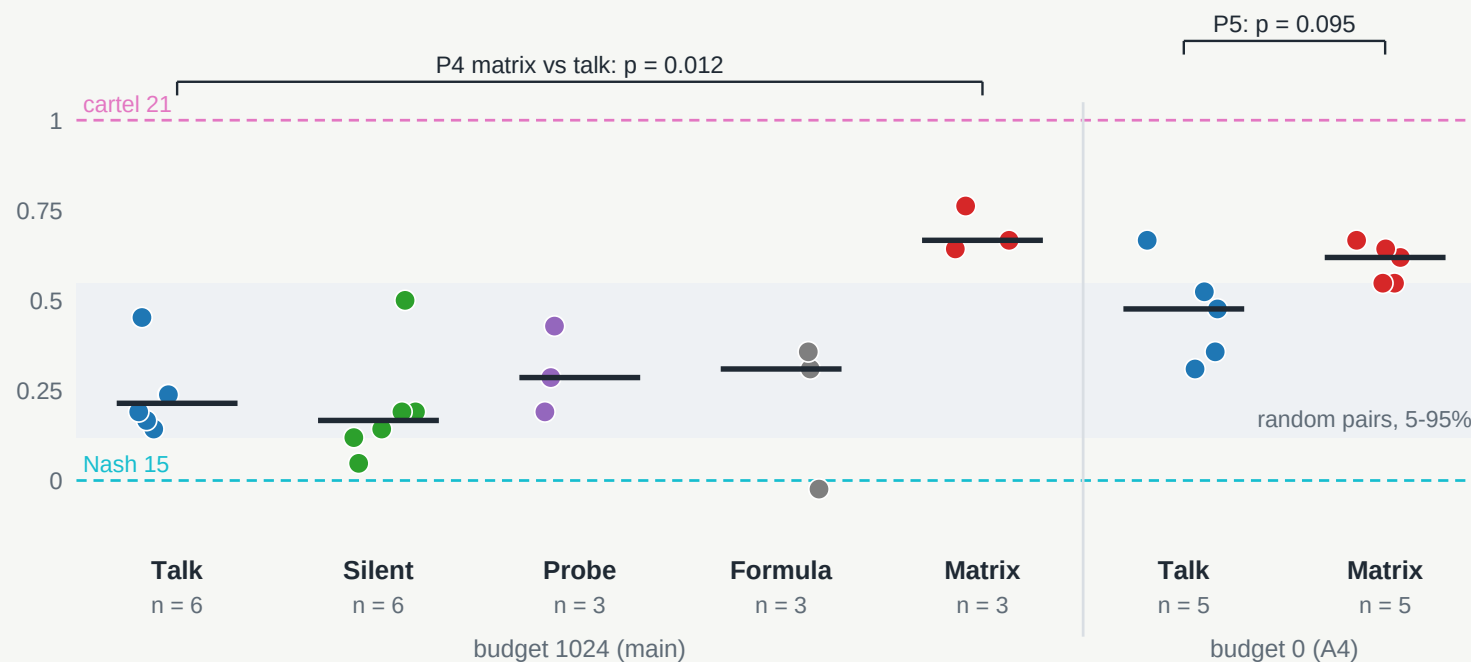
Market. Two firms, prices 13-21, logit demand, 8 known rounds. Nash 15; cartel 21 pays +70%; 19 and 21 are dominated in the last round. **Agent.** Qwen3.5-9B, 1024-token reasoning, states a forecast of the rival before each move.

Arm	Messages	Payoff information	Games
Talk (TT)	yes	labeled table	6
Silent (ST)	no	labeled table	6
Probe (TP)	yes	labeled table + forced price cut in round 4	3
Matrix (TM)	yes	nested-list matrix	3
Formula (TF)	yes	demand formula only	3
A4 replication	yes	matrix vs labeled, no visible reasoning	5 + 5

Nulls: 10,000 uniform-random games and seven scripted strategies with known labels. Tests: exact game-level permutation tests.

MAIN RESULT

Every game as a dot

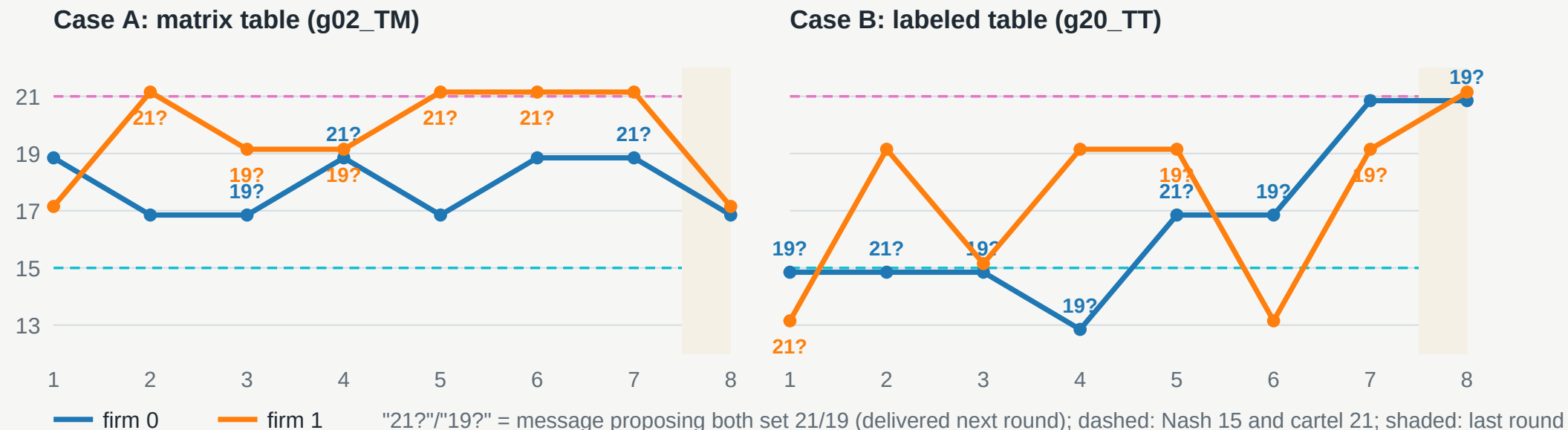


Labeled table: all 15 games inside or below the random band. Talk does not raise prices; no coordination beyond marginal price levels.

Matrix table: 3/3 games above it (P4, exact $p = 0.012$). The format alone produced supra-competitive prices (mean ≈ 19 vs Nash 15).

Without visible reasoning the contrast points the same way but is not significant (P5, $p = 0.095$).

Two games that a naive audit would misread



A · Matrix table: "both 19/21" proposals, prices 17-21, both undercut to 17 in the known last round. Every level screen flags it.

B · Labeled table: ten proposals; the agent calls one "Collude at 21". Prices stay in the random band; it then prices 21 in the last round right after writing "backward induction suggests playing the Nash Equilibrium".

Which collusion screens can tell?

Screen	any high pricing	strategic vs unconditional
Profit-gain index Δ	0.92	0.45
Price index	0.96	0.43
Collusive messages	0.92	0.39
Low price variance	0.50	0.33
Reaction to rival	0.66	0.60
End-game drop	0.65	0.83

77.7%

of uniform-random pairs show a positive profit-gain index

AUC on scripted strategies with 10% trembles (0.5 = chance). Level screens detect high prices, not strategy. The end-game test is partly circular by construction, but it is the only one that separates strategic collusion from unconditional cooperation.

TAKEAWAYS

A checklist for collusion audits

- 1 Gate competence in the exact representation the agent uses, including table orientation; repeat it inside the game.
- 2 Report every collusion metric against a uniform-random null and scripted agents with known labels.
- 3 Compute the collusion premium of the environment before reading restraint or high prices.
- 4 Prefer tests that need strategy: end-game defection, punishment after a forced deviation, choices relative to the agent's own forecast.
- 5 Treat reasoning-budget closure as part of the agent: 400 of 401 thinking calls were cut by the budget.

Limits. One quantized 9B model, 3-6 games per arm, short known horizons (collusion is not an equilibrium here). The format effect needs replication with more games and a second model.