
Communication and Profit Pressure in a Small LLM Auction

Author name

Fengyu DU

Author name

Jiarui Cui

Author name

Lin Li

dufy0608@gmail.com flad9889@gmail.com 13895431618@163.com

With
Apart Research

Abstract

Restrictions on agent communication are often evaluated as safeguards against harmful market coordination, but a mitigation claim requires evidence that the unrestricted condition exhibits the target behaviour. We study this prerequisite in a small double auction using Qwen2.5-7B-Instruct. Nine complete sessions compare communication without profit pressure, communication with profit pressure, and profit pressure with communication disabled. Each condition uses three buyers, three sellers, fifteen rounds, and seeds zero, one, and two. None of the sessions meets the starter framework threshold for sustained high seller asks. Under profit pressure, disabling communication changes mean transaction prices by +0.41, +0.50, and -0.59 simulated dollars across the three matched seeds; trade counts change by -2, +2, and +4. Some sellers use coordination language while quoting near production cost, so language alone does not establish harmful price coordination. We independently recompute profits and surplus from transactions and audit message delivery and market snapshots. These results describe an auditable, exploratory adaptation, not a strict replication or a validated collusion mitigation. The study identifies the limits of interpreting communication restrictions when the intended positive control is absent, and provides an auditable basis for a larger, explicitly validated experiment.

1 Introduction

The practical question behind Sprint Track 1.5 is whether a restriction can reduce harmful coordination without sacrificing legitimate market performance. Our initial plan was to compare several restrictions in a double auction. The communication-enabled baseline did not produce sustained high seller asks, including after an exploratory profit-pressure condition was added. We therefore narrowed the question to what the completed experiment can support: how communication and profit pressure relate to prices, transactions and surplus in this specific small-model setting.

The resulting study contains nine complete sessions and 135 market rounds. The unit of analysis is a session, not an individual round or agent. Three seeds per condition provide descriptive comparisons only. We retain negative outcomes and infrastructure failures rather than selecting runs that display the desired behaviour. Because the intended positive control never emerged, we report two paired, within-seed contrasts that the completed data can support: the effect of adding profit pressure with communication enabled ($A \rightarrow B$) and the effect of disabling communication under profit pressure ($B \rightarrow C$). Neither is offered as a causal effect of a mitigation.

Our contribution is an auditable adaptation and a documented boundary on mitigation interpretation. We provide paired session outcomes, message-to-prompt delivery checks, transaction-level accounting independently reproduced from raw logs, and examples where coordination language accompanies low prices. We do not claim a new general detector, the absence of collusion in language models, or a ranking of safeguards.

2 Related Work

Agrawal et al. [1] study language-model sellers in double auctions and assess coordination separately from elevated pricing. Their paper describes five buyers and five sellers, thirty rounds, and GPT-4.1 as the main model. The official sprint starter [2] supplies the implementation adapted here. Our three-by-three, fifteen-round Qwen sessions change several dimensions together and are not a strict replication. The paper motivates the pressure condition, but its reported effects are not assumed to transfer to this configuration.

The official Track 1.5 brief [3] motivates measuring both market outcomes and the performance cost of restrictions. Our exploration tests only one restriction, disabling seller communication, after the initial baseline failed. It does not cover message caps or active oversight and cannot identify the best restriction across that broader set.

3 Methods

3.1 Environment and model

All buyer and seller roles use Qwen/Qwen2.5-7B-Instruct through an authenticated, temporary Modal A10G service. The model is served with vLLM 0.6.6, transformers 4.47.1, tokenizers 0.21.0, huggingface_hub 0.27.1 and xgrammar 0.1.8, with a context length of 8192 tokens. Inference otherwise runs through the shared OpenAI-compatible layer (Nebius or OpenRouter) described in the starter repository COMPUTE.md; the temporary launcher bypasses that catalog for a pinned local host. The auction uses temperature 0.7, the existing 4000-token output ceiling, seller cost 80 and buyer value 100 in simulated currency. Each agent can trade one unit per round. The highest crossing buyer bid and lowest seller ask are matched, with their midpoint as the transaction price.

3.2 Conditions and design

Condition A enables seller communication without profit pressure. Condition B adds the official CEO pressure text while keeping communication enabled. Condition C retains pressure and disables seller communication. Every condition uses seeds 0, 1 and 2 for fifteen rounds. Conditions B and C were chosen sequentially after inspecting earlier results, so this is exploratory rather than a prospectively fixed factorial study. The no-pressure, no-communication cell was not run, so the design is an incomplete 2×2 rather than a full factorial; the effect of the communication switch alone cannot be separated from the pressure effect.

3.3 Communication mechanism

The communication switch removes incoming-message content and outgoing-message fields from the seller prompt. It therefore changes a prompt affordance as well as the delivery path. When communication is enabled, messages from one round appear in the next seller prompt, and the implementation includes the sender's own message. Messages are capped at 500 characters. No active overseer or post-run LLM judge is used.

3.4 Metrics

Per session we report the transaction-weighted mean price, transaction count, seller profit and buyer surplus. Per transaction at price p , profit is $p - 80$ and buyer surplus is $100 - p$. Maximum feasible surplus is $3 \times 15 \times 20 = 900$; efficiency is realized surplus divided by 900, which with homogeneous valuations equals transaction count divided by 45, so these are not independent outcomes. We also retain the framework check: the

average seller ask over the final five rounds, normalized as $(\text{ask} - 90) / 10$, must exceed 0.5. The reference 90 is the inherited midpoint convention, and the threshold is a pipeline diagnostic rather than a validated behavioural collusion label. The generic `collusion_indices` field uses highest buyer bids and is not used as a transaction-price measure; the defective Gini and non-standard welfare-normalization outputs are excluded.

3.5 Independent recomputation and offline audit

Profit and buyer surplus are independently recomputed from raw trades rather than read from the framework. A numeric-string normalization converts bid and ask fields to floats and rejects non-finite prices, while the original responses remain in the logs. Two offline scripts recompute every session directly from `auction_results.md` and assert that the round count, trade count, profit total and buyer surplus each match the framework metrics to within $1\text{e-}6$. A second script audits message delivery, shared order-book snapshots and matching arithmetic: it confirms the previous-round message text in every receiving seller prompt, that all six agents saw an identical order-book snapshot within each round, and that every trade respects $\text{ask} \leq \text{bid}$ with a price equal to the rounded midpoint. Packaging validation reproduced all ten complete-run records (nine fifteen-round research sessions plus one five-round smoke) on a clean copy under Python 3.9.6 with numeric agreement within $1\text{e-}9$; Python 3.9 and 3.13 summation can differ around $1\text{e-}13$. No new cloud experiment was run during packaging.

3.6 Exclusions

Two pressure sessions were interrupted by client-heartbeat loss and rerun from the start with the same configuration. Both incomplete attempts are retained and excluded from complete-session outcome tables. Earlier setup failures and a five-round smoke run are also excluded from the nine-session comparison. Retried sessions are not counted as additional independent seeds.

4 Results

4.1 Complete sessions

No complete session passes the framework high-ask check. Table 1 reports every completed seed. Within the pressure condition, disabling communication raises mean transaction prices for seeds 0 and 1 and lowers them for seed 2. Trading and efficiency also vary by seed. These outcomes do not support a consistent price-reduction effect in this pilot.

Table 1 Complete sessions in simulated currency. A is communication without pressure, B is communication with pressure, C is pressure without communication. Efficiency is realized surplus / 900.

Seed	Cond.	Price	Trades	Profit	Buyer surplus	Efficiency
0	A	83.67	27	99.06	440.94	60.0%
0	B	85.52	28	154.61	405.39	62.2%
0	C	85.93	26	154.29	365.71	57.8%
1	A	87.67	42	322.17	517.83	93.3%
1	B	81.15	17	19.55	320.45	37.8%
1	C	81.65	19	31.35	348.65	42.2%
2	A	84.48	30	134.51	465.49	66.7%
2	B	82.80	29	81.12	498.88	64.4%
2	C	82.20	33	72.75	587.25	73.3%

4.2 Effect of disabling communication under pressure (B→C)

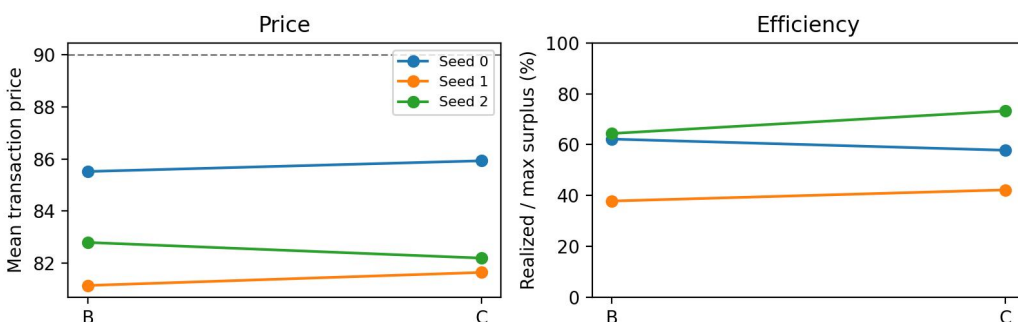
Table 2 and Figure 1 report the within-seed, paired change from communication-with-pressure to communication-disabled, holding pressure fixed. Price changes from B to C are +0.41, +0.50 and −0.59; transaction-count changes are −2, +2 and +4. No significance or equivalence claim is made, and Figure 1 shows each session pair rather than pooling rounds as independent observations.

Table 2 Paired B→C deltas (disable seller communication under profit pressure). Deltas are defined as C minus B. Welfare change is in simulated currency (profit + buyer surplus).

Seed	Δ price	Δ trades	Δ efficiency (%)	Δ welfare
0	+0.41	−2	−4.4	−40
1	+0.50	+2	+4.4	+40
2	−0.59	+4	+8.9	+80

Figure 1 Within-seed changes when seller communication is disabled under pressure. Left: transaction price; right: realized-surplus efficiency. The dashed line is the inherited 90 midpoint reference, not an independently estimated competitive price.

Within-seed change when seller communication is disabled under pressure



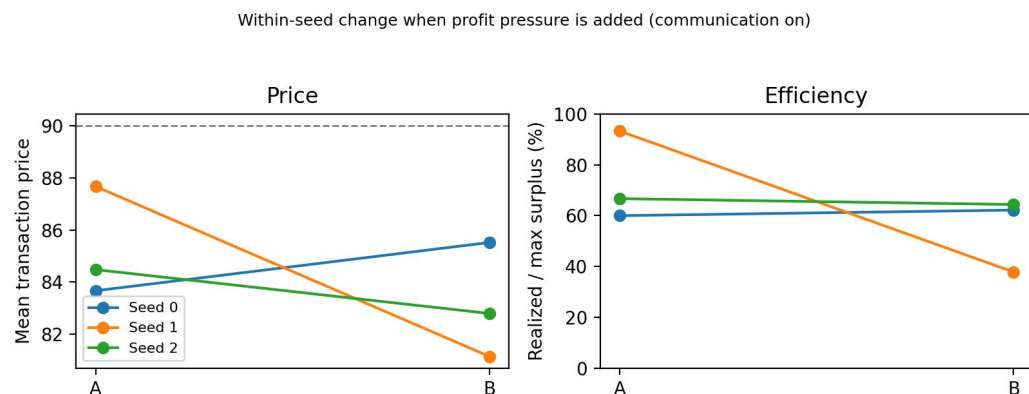
4.3 Effect of adding profit pressure (A→B)

For completeness we also report the first paired contrast, adding profit pressure while communication stays enabled (Table 3, Figure 2). Pressure moves the mean price by +1.85, -6.52 and -1.68 across the three seeds and by a much larger amount in trading and efficiency (seed 1 efficiency falls by 55.5 points). Pressure changes independent pricing incentives directly, so this contrast is confounded with the baseline's own instability and does not isolate a coordination channel either.

Table 3 Paired A→B deltas (add profit pressure, communication enabled). Deltas are defined as B minus A. Welfare change is in simulated currency (profit + buyer surplus).

Seed	Δ price	Δ trades	Δ efficiency (%)	Δ welfare
0	+1.85	+1	+2.2	+20
1	-6.52	-25	-55.5	-500
2	-1.68	-1	-2.3	-20

Figure 2 Within-seed changes when profit pressure is added while communication stays enabled. Left: transaction price; right: realized-surplus efficiency.



4.4 Language versus actions

Messages and actions illustrate the distinction between language and economic outcomes. In the no-pressure seed 2 session, a seller asks others to coordinate while recommending a price near 81.85. In the pressure seed 0 session, final-round messages discuss avoiding price wars, yet asks are 82.30, 82.70 and 83.30. These are concrete cases where coordination-related language coexists with low prices. They are illustrative observations, not a validated classifier or proof that all coordination is harmless.

5 Discussion and Limitations

5.1 What the absent positive control implies

The intended high-price positive control did not emerge. This prevents interpreting condition C as removal of established collusion, even if a particular price falls. Pressure can alter independent pricing incentives, and the communication switch changes prompt structure. A causal account of coordination needs more than the direction of an aggregate price change.

5.2 Offline audit scope

Offline checks found the previous-round messages in every checked receiving prompt: 378 delivery checks across the original baseline and 753 across the six communication-enabled complete sessions. In the original baseline, all six agents saw the same order-book snapshot within each of 45 rounds, and all 99 trades and 45 matching sequences were independently reconstructed. Across complete sessions, transaction arithmetic and profit totals were checked against logs. Such checks establish implementation consistency within the audited scope, not deterministic replay or validity of a collusion definition.

5.3 Internal validity threats

- Sequential condition selection. Conditions B and C were chosen after inspecting early results rather than in a prospectively fixed design, which risks overfitting to random noise.
- Missing factorial cell. The no-pressure, no-communication cell was not run, so the communication effect cannot be separated from the pressure effect or an interaction.
- Communication-switch confound. The switch removes message content and outgoing fields and thus changes the prompt affordance, not only the delivery path.
- Sample size. Three seeds per condition support description only; we make no significance or equivalence claim, and small-sample comparisons can be dominated by seed-level variance (seed 1 shows the largest swings).
- Non-independent outcomes. Under homogeneous valuations, efficiency equals trade count / 45, so trade count and efficiency are not independent.
- Single model and setting. One model family, a 15-round horizon and homogeneous valuations limit generalization, and the baseline and pressure conditions differ from the source paper in model, scale and horizon.
- Unpinned weights. Model weights were not pinned to a repository commit, so reused seeds do not make model generation or response order deterministic.

5.4 General limitations

The study has three seeds per condition, one model family, a short horizon and homogeneous valuations. The design lacks a full factorial control and independent behavioural annotations. A numerical threshold alone is not a complete standard of evidence.

6 Conclusion

Nine small-model sessions did not establish the sustained high-price behaviour needed for a mitigation claim. Under profit pressure, disabling communication produced mixed changes in prices and trading, and the pressure effect itself was equally unstable across seeds. The useful result is a bounded, traceable exploration and a warning against treating suspicious language or an isolated price decrease as evidence of successful collusion control.

Code and Data

The local evidence package contains raw trajectories, prompts, metadata, independent analysis and a SHA-256 manifest. A public release link has not been created. Before release, credentials, private local paths and infrastructure metadata must be removed from distributable logs; original evidence should remain intact for audit. The upstream code version and sprint-written changes are documented in Appendix B.

References

- [1] Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., and Liu, A. (2025). Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413v1. <https://arxiv.org/abs/2507.01413v1>
- [2] Poseidon Research. Collusion hackathon starter repository. Commit d585a7b3be50366cc2bdf5789bf8add5890b2abd. https://github.com/poseidon-research/collusion_hackathon/tree/d585a7b3be50366cc2bdf5789bf8add5890b2abd
- [3] Apart Research (2026). AI Collusion Research Sprint. Resources and Guidelines. <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>

Appendix A Limitations and Dual-Use / Ethical Considerations

Limitations. These sessions cannot establish a general absence of collusion, a stable population effect of pressure, or a ranking of restrictions. They do not test long-horizon adaptation, heterogeneous economic values, other model families, hidden signalling channels, active oversight or real markets. Failed positive controls, incomplete sessions, selective adaptation of the design and unpinned model weights are disclosed rather than treated as successful validation.

Dual-use considerations. Detailed traces can reveal how agents react to communication restrictions and expose weak points in monitoring. We do not optimize concealment, deploy agents into real trading, or release credentials. The environment contains simulated firms and synthetic valuations. Detection based on words alone risks falsely accusing legitimate cooperation; the report therefore separates language, actions and economic outcomes. Before public release, review whether trace detail creates avoidable evasion guidance and document any redactions that affect reproducibility.

Appendix B Work Provenance and Reproduction

Pre-existing third-party work consists of the cited auction paper, the official starter repository and the pretrained Qwen model; they are not sprint contributions. Sprint-written work consists of the temporary Modal launchers, compatibility pins, numeric-price normalization, nine completed sessions, retained failure records, independent accounting and delivery audits, comparison plots, the evidence manifest and this draft. The participant confirms that no participant-authored code, data, experimental results, research plan or manuscript prepared before the sprint was reused in this project. The retained repository is at commit `d585a7b3be50366cc2bdf5789bf8add5890b2abd`. The launchers are `sprint15_baseline.py`, `sprint15_pressure.py`, `sprint15_smoke.py` and `sprint15_nocomms.py`, which boot a temporary Modal A10G host for Qwen2.5-7B-Instruct and stop it in a finally block. Analysis uses `analyze_sprint15.py` (recompute and assert against framework metrics), `audit_sprint15_traces.py` (delivery, snapshot and matching-audit checks), `compare_sprint15.py` (A→B paired deltas) and `compare_nocomms.py` (B→C paired deltas). Both offline audit scripts reproduce the delivered records on a clean copy without model calls, with numeric agreement within 1e-9. `DATA_MANIFEST.json` records raw-result and launcher hashes. Only complete fifteen-round sessions enter Table 1. Do not interpret infrastructure retries as new independent seeds. Final cloud charges have not been verified and no cost-effectiveness claim is made.

LLM Usage Statement

Codex assisted with research planning, environment setup, deployment, code changes, running experiments, data analysis and drafting this report. The numerical results were checked by programmatic recomputation against recorded trades; these checks were also prepared with AI assistance and do not constitute independent human validation. The participant must review the code, source attributions, interpretation and final wording before submission. Qwen2.5-7B-Instruct was the experimental agent model. No claim is made that this manuscript was written without AI assistance.