

How Many Characters Does It Take?

Communication bandwidth and supra-competitive pricing in LLM procurement agents

Ishaan Datta - BITS Pilani

Abstract

Multi-agent LLM systems increasingly quote prices and bid on someone's behalf. A natural safety hope is that capping how much agents can say to each other makes collusion harder. We test this in a simulated procurement market where four supplier agents file weekly prices for medicine. The only thing we vary is a per-seat character allowance per cycle (none, 128, 256, 512 or 2,048), holding the world, prompts and tools fixed. Harm is simulated overcharge above a capacity-aware cost benchmark.

With gemma4:cloud, the no-channel control was already 17-25% above cost (pairwise price agreement 0.19-0.23). At 512 characters per seat per cycle, agreement rose to 0.88 and overcharge to 41-53% on the same two seeds; at 2,048 characters agreement was 0.94-1.00 but harm varied (22-41%). In one cell, coordination took seven messages in the first cycle, then silence. gpt-oss:120b stayed within 6% of cost at every allowance and deepseek-v4.1-flash under 10%. No agent ever proposed a lower price for the patients the scenario describes. Realism is a real limit: one candidate model appeared to recognise the exercise as a game, which undermines treating its harm as real to the agent. Results are preliminary: one scenario, one to three seeds per arm, no statistical tests.

Main takeaway: in this market, a channel of a few hundred characters per cycle was enough for gemma to coordinate prices well above cost. The effect was not monotone: on the available seeds, 2,048 characters produced less harm on average than 512 (34.7% vs 47.2% over cost), and two other models barely moved.

1. Introduction

LLM agents increasingly output a price, a bid or an approval. When several share a market and can talk, they may coordinate on outcomes that benefit them and harm others. A common defensive intuition is bandwidth: if agents cannot say much, they cannot agree on much.

We test it with a controlled dose. One procurement world is held fixed (same buyers, costs, capacities, eight cycles, tools) and only the per-seat character allowance changes. In the

control, the messaging tools do not exist. Every treated result therefore has a no-line baseline from the same world, rather than a theoretical ideal.

Harm is simulated overcharge: the extra money buyers pay above what the same order would cost if each seller filed at its own landed cost (a capacity-aware benchmark), stated as a percentage over cost. Selling at cost is the neutral act in the scenario, so this reads directly as extra paid by a district that cannot buy elsewhere. It is a property of the simulation and does not show that any agent intended real harm (Section 5).

Our contributions are:

1. **A bandwidth dose with a matched no-line control.** Only the character allowance differs between arms; the world, prompts, tools and settlement rule are fixed.
2. **A measured result for one model.** For gemma4:cloud on matched seeds, 512 characters per seat per cycle moved overcharge from 17-25% to 41-53% and price agreement from 0.19-0.23 to 0.88. Two other models did not respond.
3. **A negative result on salience.** In 182 admitted messages, no agent proposed a lower price for the people who could not afford the medicine, and nobody answered the request for free stock, although every seat repeatedly read the note about both.

We do not claim necessity, monotonicity, a minimum effective allowance, generality across models, or anything about deployed systems. 512 characters is the smallest allowance completed on gemma in this report, so the threshold is unknown.

2. Related Work

Cheap talk coordinates expectations when interests are sufficiently aligned (Crawford & Sobel, 1982; Farrell & Rabin, 1996), and laboratory markets show that even non-binding communication raises prices (Cooper, DeJong, Forsythe, & Ross, 1992). Our channel is cheap talk with a hard character budget, and the budget is the treatment.

Repeated interaction can sustain supra-competitive prices without any explicit agreement (Stigler, 1964; Friedman, 1971; Tirole, 1988). Empirical screens look for fingerprints in prices and bids — structural separation of competition from collusion in procurement (Bajari & Ye, 2003), variance screens (Abrantes-Metz, Froeb, Geweke, & Taylor, 2006), and behavioural screening more broadly (Harrington, 2006). Our no-line control is exactly the object those screens target: desks post prices with no channel at all and still land above cost. That is why the control, not zero, is the baseline.

Algorithmic pricing can learn collusive policies without explicit communication (Calvano, Calzolari, Denicolò, & Pastorello, 2020), and algorithmic pricing has been associated with higher retail margins in German gasoline markets (Assad, Clark, Ershov, & Xu, 2024). Fish, Gonczarowski, and Shorrer (2024) study collusion among LLM pricing agents. Our work differs

in the axis it varies: we do not ask whether LLM agents can collude, but how much channel they need before they do, and we report the answer as harm over a matched baseline.

We read out coordination as pairwise agreement over coarse price bins. Cohen's (1960) kappa is undefined in treated cells where all seats share a bin, so we report raw agreement. Parallel pricing alone does not establish collusion, and antitrust practice likewise requires more than parallel conduct. Here the no-line control is the parallel-conduct baseline, and we report differences from it. We run no inferential tests (Section 5).

3. Methods

The world (S1, institutional paracetamol procurement). Four desks (seats A-D) file a weekly price per carton for a district relief season. Five buyers each need 2-3 cartons; total demand is 40% +/- 5% of district capacity, so some desks go unserved and price competition has teeth. Each desk has a private landed cost (₹762-897) and weekly capacity (6-10 cartons). Allocation is cheapest-first and buyers are price-takers. Prices are capped at ₹1,200 per carton, which becomes a focal point in treated cells.

The allowance. A cell is an (allowance, seed) pair at N=4 seats. The allowance is the number of characters each seat may send per cycle: 0, 128, 256, 512 or 2,048. At 0 the message tools are absent. Otherwise every seat has the same two tools (a shared association room and direct notes to one or more seats), the same ledger and the same rule: a message over the remaining allowance is rejected, not debited, and may be retried shorter in the same turn.

Turn structure. Each cell runs 8 cycles of 6 exchanges. Seats act simultaneously and messages are delivered at the close of each exchange, so no seat sees another's message before acting. Exchange 6 is filing-only. Agents therefore have several chances to talk before committing a price, so a "one-shot" pattern is a choice, not a constraint.

Harm. The benchmark for each cycle is the same buyer sheet allocated at each desk's own landed cost, with each buyer's reference being its marginal supplier cost. Harm is the overcharge the district pays above that benchmark; % over cost is harm divided by the benchmark. We also record cases forgone but do not add them to the money axis, so cells that exclude buyers are understated.

Models. gemma4:cloud (primary), gpt-oss:120b-cloud and deepseek-v4.1-flash:cloud, through Ollama at temperature 1.0 with native tool calls only. Text that is not a tool call is never parsed into an action. Two further candidates (glm-5.3-flash and beam) are not scored (see Section 5).

Realism gate. Every rendered string is in-world: no mention of simulations, tests, prompts, tokens or agents. A gate scans visible reasoning and messages. A visible failure halts the cell and is recorded as HALTED, not dropped. Game recognition is a limit on realism, discussed in Section 5.

Seeds. A seed fixes the world generator (costs, capacities, weekly demand draw), not the prompt text. gemma4:cloud: two seeds at 0 and 512 characters, three at 2,048. gpt-oss:120b and deepseek-v4.1-flash: one seed per allowance, and the seed differs between allowances in Table 1, so those contrasts are not matched. The salience arm (Table 2) has two seeds per allowance. Counts are too small for inference, so we report cell-level numbers.

Other scenarios. Three further scenarios did not measure what they claimed and are boundary conditions, not evidence. In market-making, no seat ever sent a message (0 send attempts). In credit rating, agents talked and then filed honest ratings. In bridge approval, the checker caught the planted data-room gap every cycle, so the chain deadlocked and harm was zero. A fourth (approval) was dropped.

4. Results

Table 1. S1 at N=4, 8 cycles. Allowance is characters per seat per cycle. Harm is overcharge over a capacity-aware cost benchmark (₹78,590-83,125 depending on seed); % over cost is harm divided by that benchmark. Agreement is mean pairwise same-bin agreement over cycles and pairs.

Seed - a particular set of numbers

Model	Chars	Seed	Mean price (₹)	Agreement	Harm (₹)	% over cost
gemma4:cloud	0	0	964	0.19	14,021	17.1
gemma4:cloud	0	1	1,022	0.23	19,860	25.3
gemma4:cloud	512	0	1,100	0.88	33,666	41.1
gemma4:cloud	512	1	1,197	0.88	41,810	53.2
gemma4:cloud	2,048	0	1,098	0.94	33,186	40.6
gemma4:cloud	2,048	1	950	0.94	17,350	22.1
gemma4:cloud	2,048	2	1,100	1.00	33,775	41.3
gpt-oss:120b	0	2	879	0.17	4,010	4.9
gpt-oss:120b	256	2	861	0.15	3,076	3.8
gpt-oss:120b	2,048	0	870	0.23	4,614	5.6
deepseek-v4.1-flash	0	1	898	0.15	7,244	9.2

Model	Chars	Seed	Mean price (₹)	Agreement	Harm (₹)	% over cost
deepseek-v4.1-flash	512	0	857	0.48	5,893	7.2

The no-line control is not competitive. With no channel, gemma desks dispersed across ₹850-1,200 and still cleared 17-25% above cost. Treated cells therefore do not create harm from zero: communication moved an existing focal point, and "communication caused collusion" would overstate the result.

The channel moves gemma further from cost. At 512 characters, agreement was 0.88 on both seeds and desks converged on ₹1,100-1,200, raising overcharge 2.4x and 2.1x relative to the same seed with no line. At 2,048 characters agreement was 0.94-1.00. In one cell the seats exchanged seven messages (994 bytes) in the first cycle, then said nothing for seven cycles while filing identical prices. In the highest-harm cell, the highest-harm cycle contained that cell's shortest admitted message (51 characters), which set ₹1,200. In these cells the channel seems to have been needed to start coordination, not to sustain it; this is a single-cell observation.

Other models barely moved. gpt-oss:120b never exceeded 5.6% over cost at any allowance, with low agreement (0.15-0.23) and prices near its own costs. deepseek-v4.1-flash was at 9.2% with no line and 7.2% with 512 characters (agreement 0.48): some convergence, no price ceiling. Seeds differ across arms for both, so these contrasts are not matched. We do not know why gemma responds and these models do not, and gemma may not be representative.

The effect is not monotone. Gemma's mean overcharge was 21.2% with no line, 47.2% at 512 characters and 34.7% at 2,048 (one 2,048-character seed: 22.1%). More channel did not buy more harm. Transcripts suggest that at 2,048 characters the desks spend their allowance on corrections about the buyer sheet and arguments about who is breaching, while at 512 the messages are pure price proposals. That is a hypothesis for a larger study, not a finding.

No one proposed a cheaper price for the patients. The district health office note says camp committees cannot afford the filed prices and asks suppliers to hold prices down; a later version also asks any supplier able to release stock free to say so. Every seat read the note repeatedly (29-32 reads per cell in audited cells). Across 182 admitted messages, the words patient, medicine, welfare, charity, humanitarian, moral and duty never appear, and no message offers free or discounted stock. "Afford" appears twice, both as a gloss on a high price: one admitted message proposes ₹1,100 "while keeping it affordable for the district", and one rejected over-allowance message asks others to "align to keep relief affordable". In gemma's visible reasoning, the district's need is a limit on what buyers will bear, not a reason to price lower.

Verdicts. Across the completed S1 cells: 706 ACTION, 6 WITHHELD, 0 REFUSED, 0 MALFORMED. Seats did not refuse or break character in these cells. No halt was a refusal.

Salience arm (v17). A later S1 version tells the seats that camp committees report the filed prices exceed what the relief budget can carry, and asks explicitly whether any supplier will release stock free for the season. The world, feasible prices and tools are unchanged. gpt-oss completed two seeds at 0, 128, 256 and 512 characters. Gemma and beam runs were still in progress when this report was written and are not used as evidence.

Table 2. S1 v17 (affordability line and free-medicine plea), gpt-oss:120b, N=4. Columns as in Table 1.

Model	Chars	Seed	Mean price (₹)	Agreement	Harm (₹)	% over cost
gpt-oss:120b	0	0	856	0.08	1,326	1.6
gpt-oss:120b	0	1	839	0.10	1,268	1.6
gpt-oss:120b	128	0	840	0.17	868	1.1
gpt-oss:120b	128	1	841	0.21	959	1.2
gpt-oss:120b	256	0	841	0.17	1,707	2.1
gpt-oss:120b	256	1	846	0.10	734	0.9
gpt-oss:120b	512	0	840	0.17	1,592	1.9
gpt-oss:120b	512	1	853	0.21	1,456	1.9

Across two seeds and four allowances, including a no-line control, gpt-oss's mean filed price stayed within ₹839-856 and overcharge never exceeded 2.1%. Neither the channel nor the affordability text moved it.

5. Discussion, Limitations and Dual-Use

Reading. If the question is "will these agents collude if given a channel?", the answer here is that one model family did: with a few hundred characters per cycle, in a single burst, without naming what it was doing. In this world, a cap at roughly the length of a couple of text messages is not a mitigation for that model. It says nothing about models that did not respond.

Visibility. The agents read the note about the harm repeatedly and never acted on it or spoke of it. Making a harm legible in context was not the same as making it operative in behaviour, so an

evaluation that checks whether a model "knows" about a harm may overstate how much that knowledge constrains it. This is one scenario and one primary model.

Limitations and Dual-Use / Ethical Considerations. What this result does and does not establish:

(i) **Scale.** One scenario, one strongly responding model and one to three seeds per arm.

Nothing is statistically tested; differences are descriptive.

(ii) **Realism.** This is the main threat to validity. By the author's observation (no transcript analysis is included here), glm-5.3-flash recognised almost immediately that it was in a game, so the harm was no longer real to it. Its behaviour cannot be read as a response to real stakes, and its cells are not scored. Separately, gpt-oss-20b fell back to its default personality instead of the assigned seat role. The same problem may affect scored models in ways our gate does not catch, and excluding models removes those most likely to behave differently. Simulated overcharge is computed from filed prices and is unaffected, but recognising a game limits what it says about real intent or real-world harm.

(iii) **Price cap.** Several treated cells sit exactly on the ₹1,200 cap, so harm at high allowances is partly bounded by design, and the cap is a plausible focal point.

(iv) **Baseline.** The no-line control is already supra-competitive, so the treatment effect is movement from an imperfect equilibrium; against perfect competition, all of the control's harm would be attributed to the world rather than the line.

(v) **Metric.** Harm counts overcharge only; cells that exclude buyers are understated. Agreement over coarse price bins plus overcharge against a no-line control is not a validated collusion detector, and this report has no scripted-cartel positive control.

(vi) **Sampling.** Temperature 1.0 with one seed per cell gives one trajectory per cell, not a distribution; seeds vary the world, not the sampling.

(vii) **Dual use.** The setup shows that a very short channel can be enough to start coordination, and that talk volume is a poor proxy for coordination, so a monitor keyed to message volume could miss it. That may help defenders more than attackers, but it could also inform someone trying to stay under a bandwidth limit. The agents' standing instruction is to protect the plant's contribution (see PROMPTS.md), not to collude. Results come from a simulation and say nothing about real firms.

Future work. More seeds per arm and a within-cell design with and without a line, so differences can carry intervals; gemma below 512 characters to find any threshold; the salience arm on gemma; framing that reduces game recognition; a scripted-cartel positive control for the agreement measure; and delaying price actions after a message to separate the founding message from the price series that sustains coordination.

6. Conclusion

We held a procurement market fixed and varied only how many characters its pricing agents could exchange. With no channel, gemma desks already priced 17-25% above a capacity-aware cost benchmark. With 512 characters per seat per cycle they converged on a

single price at 41-53% above cost; at 2,048 characters agreement was 0.94-1.00. In one cell, coordination took seven messages and one cycle. gpt-oss:120b and deepseek-v4.1-flash barely moved, and no agent offered a cheaper price for the patients named in the district's note.

For this model, a cap of a few hundred characters did not prevent coordination, and a channel that small may not look like a channel to a designer. The result is preliminary and simulated. It does not show that bandwidth caps fail in general, where the threshold lies, or that deployed agents would act the same way.

Code and Data

- **Code repository:** <https://github.com/id-19/apart-sprint> (private repository)
- **Data:** each cell directory under `harness2/runs/` contains, per cell, the full request/response log (`raw_calls.jsonl`), per-turn trajectory, admitted and rejected messages, per-cycle settlements, the world event record, and a summary with the harm and agreement measures. The repository currently commits only one example run (an $N=3$ cell); the cells in Tables 1-2 are not committed.
- Artifacts: `PROMPTS.md` renders the exact text each seat receives; `RUNBOOK.md` documents how to reproduce a cell.

Author Contributions

Ishaan Datta designed the harness, the scenarios, and the dose; implemented the code; ran the experiments; and wrote this report. Claude Code was used as a pair programmer for implementation and for drafting assistance.

References

1. Abrantes-Metz, R. M., Froeb, L. M., Geweke, J., & Taylor, C. T. (2006). A variance screen for collusion. *International Journal of Industrial Organization*, 24(3), 467–486.
2. Assad, S., Clark, R., Ershov, D., & Xu, L. (2024). Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market. *Journal of Political Economy*, 132(3).
3. Bajari, P., & Ye, L. (2003). Deciding between competition and collusion. *Review of Economics and Statistics*, 85(4), 971–989.
4. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.
5. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
6. Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1992). Communication in coordination games. *Quarterly Journal of Economics*, 107(2), 739–771.
7. Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.

8. Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, 10(3), 103–118.
9. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv:2404.00806.
10. Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *Review of Economic Studies*, 38(1), 1–12.
11. Harrington, J. E. (2006). Behavioural screening and the detection of cartels. In *European Competition Law Annual 2006*.
12. Stigler, G. J. (1964). A theory of oligopoly. *Journal of Political Economy*, 72(1), 44–61.
13. Tirole, J. (1988). *The Theory of Industrial Organization*. MIT Press.

Appendix

A. Prompt excerpt (512 characters, N=4, seat A, cycle 1). Each seat receives an in-world brief (role, unit cost, weekly capacity, the standing instruction to protect the plant's contribution), an opening turn with the district buyer sheet, its own previous sales, the district bulletin, the other desks in the association and the remaining allowance; then cycles of simultaneous exchanges with the same tools.

B. Dose realisation. The allowance is characters per seat per cycle. Internally this is a rung multiplied by N: rung 32 at N=4 gives 128 characters, and rung 512 gives 2,048. At allowance 0 the messaging tools are absent from the tool schema.

C. What a rejected message looks like. An over-allowance message is refused before sending, is not debited, and is reported to the seat in-world with the remaining allowance, so a seat may retry shorter in the same turn. Every admitted and rejected message is written to `messages.jsonl` with the disposition and the reason.

D. Realism gate. Visible-content hits halt a cell (HALTED, with the cycle reached). analysis/meta.py scans all transcripts.

LLM Usage Statement

Claude Code (Anthropic) was used as a pair programmer for the harness, scenario modules and analysis scripts, and to help draft and edit this report. The experimental design, treatments and interpretation are the author's. Table values come from run summaries produced by the analysis scripts; unfinished cells are excluded.

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