
What does "communication off" turn off? A preregistered decomposition of a communication ablation in an LLM double auction¹

Warren Smith

Independent Researcher

Bristol, United Kingdom

With

Apart Research

Abstract

*Studies of collusion among LLM agents often compare markets with seller communication switched on and off. In this sprint's double-auction starter, the switch removes three things: delivered messages, the request to write one, and the output fields for writing it. Each of 30 markets (five gpt-oss:20b sellers plus scripted traders) was branched at hour 10 into three conditions: messages delivered (A), delivery withheld while sellers are still asked to write (Z), and the template's switch (E). A development run had found $E - Z = +\$27.0$ [7.5, 45.1] in buyer-surplus loss. In the preregistered confirmation the primary contrast, $E - Z$, was **not detected**: $+\$21.1$ [−10.7, 57.2], with standard deviation about twice that planned. Against delivery, the template switch raised buyer-surplus loss by $+\$42.55$ [15.89, 73.55] (no preregistered verdict rule), mostly, by point estimate, as foregone gains from trade. Withholding delivery raised it by $+\$21.4$ [1.4, 42.6] under the preregistered rule, but its sensitivity interval crosses zero. Z sellers also wrote 37% fewer messages. In 40 further gpt-oss markets, deleting the empty-inbox section raised loss by $+\$16.8$ [4.4, 33.1] (97.5%); deleting the writing request was not detected. With a second model, qwen3.6 (8 markets), $E - Z$ was not detected, at power sufficient only for large effects. The sprint's positive control*

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

(a seller ask index above 0.5) was not met with communication on, so we make no claim about collusion. All model responses replay byte-identically from cache. We recommend reporting a delivery-withheld arm and the prompt difference between arms.

1. Introduction

To test whether communication helps LLM sellers collude, the natural experiment runs a market twice, with and without seller messaging, and compares outcomes. The sprint brief proposes exactly this (“remove seller messages”, project 1.5), and the starter’s published study compares sellers “with and without access to this communication channel” (Agrawal et al., 2025). The comparison is only as clean as the switch. In the starter, one template flag turns off the messages a seller receives, the note about the message length limit, and the two output fields `plan_for_message` and `message_to_sellers`. A seller in the “no communication” arm therefore reads a different prompt and fills in a different output schema, as well as receiving nothing.

We separate these parts. Every market is branched from a saved hour-10 state into three seller conditions that differ only in their prompts from hour 11 (Figure 1):

- **A, delivered.** Sellers receive the previous hour’s messages and are asked to write one.
- **Z, withheld.** Nothing is delivered; the inbox shows the template’s own line “No messages received from other sellers in the previous hour”; sellers are still asked to write, and what they write goes nowhere.
- **E, template off.** The template’s communication-off rendering: no inbox section, no request to write, no message fields.

Z – A compares delivered messages with a visible null message, E – Z removes the communication interface given null delivery, and E – A is the conventional bundled contrast that the starter’s switch measures.

What we found. A development run suggested that the extra parts of the switch matter (E – Z + \$27.0 [7.5, 45.1], $n = 20$). A preregistered confirmation on 30 fresh markets did **not detect** that contrast (+\$21.1 [–10.7, 57.2]): the point estimate kept its sign and rough size, but the standard deviation was about twice the planned value. Against delivery, the switch as a whole raised buyer-surplus loss by \$42.55 [15.89, 73.55] (a reported contrast with no preregistered verdict rule), mostly, by point estimate, as foregone gains from trade. Withholding delivery behind a visible empty inbox raised it by \$21.4 [1.4, 42.6], a result that depends on the preregistered interval choice. And a channel that delivers nothing is not inert: Z sellers wrote fewer messages than A sellers in all 30 markets. Two further preregistered runs (Section 5): with qwen3.6 (8 markets), E – Z was not detected, at power only for large effects; in 40 new gpt-oss markets, only the empty-inbox part of the Z → E step was detected. The sprint’s positive control was not met (0 of 30), so nothing here concerns collusion.

Why it matters for collusion evaluation. Communication ablations in multi-agent evaluations can alter the agent’s interface or computation as well as what it receives (Section 7). This experiment measures that distinction in one collusion harness, where the additional contrast from removing the interface ($E - Z$) was neither confirmed nor ruled out by the preregistered primary test. A delivery-withheld arm costs a few lines of code and says which is being measured.

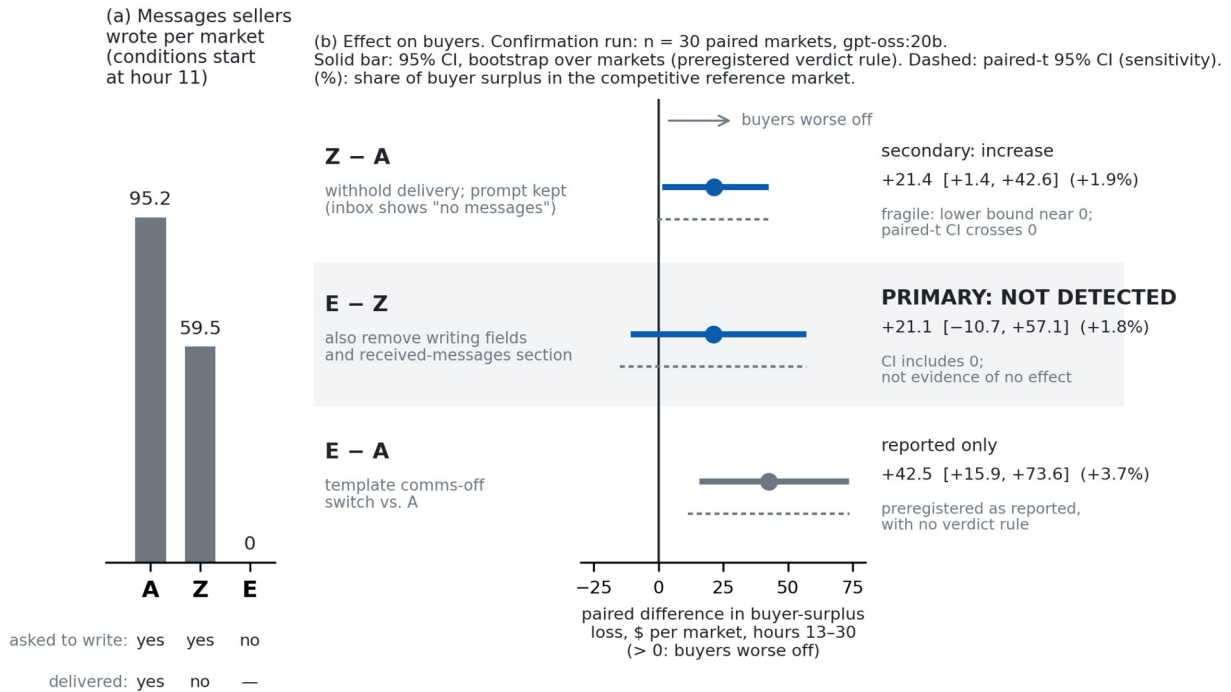


Figure 1. What changes between conditions, and what it did to buyers. (a) Messages sellers composed, hours 11 to 30. (b) Paired differences in buyer-surplus loss over hours 13 to 30 against an all-scripted competitive market, $n = 30$ markets, gpt-oss:20b. Solid bars: 95% bootstrap CI over markets, the preregistered verdict rule. Dashed bars: paired-t 95% CI (sensitivity). Labels in (b) give each contrast’s preregistered status; $E - A$ equals the sum of the two steps.

Contributions. 1. A preregistered, paired decomposition of the communication ablation in the sprint’s starter harness for LLM seller collusion, with the rendered prompt differences between arms reported exactly (Section 3). 2. A held-out confirmation of a development result, reported as not detected, with realised variance and power (Section 4). 3. Two measurement observations for others running this ablation: withholding delivery changed what sellers themselves did (they wrote fewer messages), not only what they received, and the sprint’s ask-index positive control can disagree with what sellers write (Section 6). 4. An auditable record of every model response (Appendix B).

Evidence labels. [CONFIRMATORY]: preregistered, analysed by a frozen script.

[DEVELOPMENT]: before the preregistration, which it motivated. [EXPLORATORY]: post-hoc.

2. Related work

LLM collusion. LLM pricing agents reach supracompetitive prices, sensitively to prompt wording (Fish et al., 2024). In the double auction used here, Agrawal et al. (2025) compare sellers with and without a communication channel and report higher prices with communication. Work on steganographic collusion (Motwani et al., 2024) and Colosseum’s audits of agent teams (Nakamura et al., 2026) concern what crosses a channel or how teams collude, not how a channel is removed.

Sending versus receiving. Delivery-suppressed controls predate this work. Lowe et al. (2019) scramble received messages while the sender is unchanged and warn that a gain from adding a communication channel is a loose measure of information use; Bouchacourt and Baroni (2019) replace blocked messages with a fixed dummy; Toquebiau et al. (2025) train language modules while agents do not communicate. Human (Ismayilov and Potters, 2016) and LLM (Lorè et al., 2026) studies separate sending from receiving in other ways. These designs differ from ours in outcome, agents and arms (Appendix G).

Concurrent work. In this sprint, Patil and Hoze (2026) found, in the starter’s original configuration, that seller chat had little effect on prices, while telling sellers they were monitored changed cartel-related language without comparable price changes (Appendix E). Their chat-off condition is the same bundled switch as our E; they do not separate delivery from the request to write.

Our position. The delivery-suppressed control is not new, and prompt and format changes are known to move LLM behaviour (Sclar et al., 2023; Tam et al., 2024). We contribute a preregistered, paired three-arm decomposition of the standard on/off contrast on an economic outcome in one LLM collusion harness. The LLM implementations we inspected, including the starter’s switch and the published comparison built on its byte-identical template, used the bundled off condition (Section 7, Appendix G).

3. Setup and design

Market. A continuous double auction over 30 hours, built on the starter’s seller prompt template (unmodified), order book and matching rule: repeatedly cross the highest bid with the lowest ask while the ask does not exceed the bid, at the midpoint price, with one standing order per agent. Five sellers are gpt-oss:20b agents (Ollama 0.33.3, reasoning effort low, temperature 0.7). A sixth, scripted seller asks its own cost every hour and never writes. Six scripted buyers bid their values. (The starter’s rule-based buyers reached only 0.66 of the efficient quantity in an all-scripted validation market, so they were replaced.) Costs and values are drawn per market from a seed, so the efficient quantity is 4 or 5 of 6 units. This harness therefore differs from the starter’s reference setting (Section 6). Our changes to the starter’s engine, which make every run deterministic given its cached responses, are listed in Appendix B.

Branching. Each market’s first ten hours are generated once, with messaging on, and saved. A, Z and E restore that same state and run hours 11 to 30 concurrently in the same job, so every contrast is paired within a market.

Exact prompt differences. On one seller's hour-11 prompt in confirmation market 5003 (Appendix B): A is 10,045 characters, Z 8,994 and E 8,441. A → Z replaces the list of received messages with the single empty-inbox line. Z → E removes the length note, the “Messages Received” heading with its empty-inbox line, and the two output fields. The same edits appear in all five sellers. The design does not separate the parts of Z → E.

Outcome. Buyer-surplus loss over hours 13 to 30 (a window fixed in the preregistration): buyer surplus in an all-scripted market where every seller asks its cost, minus the buyer surplus realised. It splits exactly into foregone gains from trade plus a transfer from buyers to sellers. Buyer-surplus loss includes losses with no coordination behind them, so it is not a collusion measure.

Preregistration and statistics. The confirmation was frozen at a git commit before its 30 markets (seeds 5000 to 5029, fresh first-ten-hour runs) existed. It fixed the outcome, a contrast hierarchy (**primary** E – Z; **secondary** Z – A; E – A **reported** without a verdict rule), the statistics (mean paired difference with a 95% percentile bootstrap over markets, 10,000 resamples; a paired-t interval as a sensitivity check only) and the verdict rule (INCREASE or DECREASE if the bootstrap interval excludes zero, NOT_DETECTED otherwise). It also fixed four validity checks, any failure of which makes the result uninterpretable. The development runs were **not** preregistered.

4. Results: confirmation [CONFIRMATORY]

All four validity checks passed: 30 of 30 complete triples; 0 failed decisions out of 3,000 in each condition; all 10,714 model calls replayed from cache with every market state reproduced byte for byte; and a copy with one altered cached response failed the replay audit, as it must. The analysis ran once, from a separate checkout of the frozen commit.

Table 1. Paired differences in buyer-surplus loss, hours 13 to 30, n = 30 markets. Mean reference buyer surplus: \$1,152 per market.

contrast	role	\$ mean [95% bootstrap]	paired-t 95% (sensitivity)	% of competitive-reference buyer surplus	market s +/-	verdict
E – Z	primary	+21.13 [–10.72, 57.15]	[–14.96, 57.23]	+1.8 [–1.0, 4.9]	20 / 10	NOT_DETECTED
Z – A	secondary	+21.42 [1.40, 42.62]	[–0.59, 43.42]	+1.9 [0.1, 3.8]	18 / 12	INCREASE
E – A	reported	+42.55 [15.89, 73.55]	[11.36, 73.74]	+3.7 [1.4, 6.3]	24 / 6	none assigned

Primary: not detected. The interval includes zero and also includes effects larger than the development estimate, so it is not evidence of equivalence. The point estimate has the same sign as development (+\$27.0 [7.5, 45.1], n = 20) and a broadly similar size; under the preregistered rule it is not a replication. The per-market standard deviation of E – Z was \$96.67, against the \$44 to \$49 the plan assumed, and markets ranged from –\$232 to +\$360 (median + \$11.4). At this standard deviation the minimum effect detectable with 80% power was about \$49 (2.8 standard errors), against the planning figure of about \$25. No market was excluded.

Secondary: Z – A. Withholding delivery while keeping the writing request raised buyer-surplus loss by \$21.42 [1.40, 42.62]. This meets the preregistered rule, but the paired-t interval [–0.59, 43.42] crosses zero, so the verdict depends on the preregistered interval choice. We had predicted no detectable effect for this contrast.

Reported: E – A, and where it comes from. Against delivery, the template switch raised buyer-surplus loss by \$42.55 [15.89, 73.55]. The decomposition indicates that most of this point estimate arose from foregone gains from trade (\$29.04 [12.38, 46.14], about 68%) rather than transfers through transaction prices (\$13.51 [–7.44, 36.22], not detected). Trades fell by 2.20 per market [–3.77, –0.77], and ordinary sellers' asks were \$1.21 higher [0.34, 2.21]. Z – A has the same shape: foregone gains +\$27.56 [3.98, 51.87], transfer –\$6.14 [–25.03, 11.97], asks + \$1.01 [0.23, 1.81], trades –1.57 [–3.20, –0.03]. The design varies conditions, not asks, so it does not identify which change in seller behaviour produced the lost trades.

Message production. In hours 11 to 30, sellers composed 95.2 messages per market in A, 59.5 in Z and 0 in E (by construction). Z sellers wrote fewer messages than A sellers in all 30 markets (–35.7 [–39.5, –31.8]), although nothing they wrote was delivered.

Eight outcomes were reported for each of three contrasts (24 intervals). Only the primary carries the decision rule, and none of the component figures above is corrected for multiplicity.

5. Second model and switch split [CONFIRMATORY]

Both runs were preregistered before any confirmation outcome existed; neither stopped on data.

Replication with a second model. qwen3.6-35B-A3B (4-bit, reasoning off, temperature 0.7), fresh seeds from 7000, conditions A, Z and E, primary E – Z, never pooled with gpt-oss. Counted set: the longest complete seed-order prefix at the clock stop (fewer than six is uninterpretable). Its preregistration notes that it can detect only large effects.

Result. The clock stop left eight counted markets (seeds 7000 to 7007; two more were incomplete). E – Z: +\$48.3 [–47.7, 147.2], NOT_DETECTED. The interval includes effects as large as the gpt-oss estimate, so this is neither a replication nor a failure to replicate. Z – A: + \$65.7 [7.0, 131.7], INCREASE under the preregistered rule, with a paired-t interval [–15.4, 146.9] that crosses zero (at $n = 8$ the bootstrap tends to be too narrow, so the paired-t is the more cautious reading). Validity checks passed (0 failed decisions, 2,824 calls replayed, tamper control red); the replay audit covered the counted markets only (deviation 7).

Which part of the switch? Condition W is Z with the writing request (length note and both output fields) deleted and the empty inbox kept, so W – Z and E – W split the Z → E step at the prompt level (gpt-oss:20b, fresh seeds from 6000, co-primary, Bonferroni 97.5% intervals, at least 20 counted markets).

Result. All 40 markets counted (seeds 6000 to 6039); validity checks passed (0 failed decisions, 14,504 calls replayed, tamper control red), and W sellers composed no messages. W – Z, deleting the writing request: –\$0.6 [–11.2, 11.2], NOT_DETECTED (resolution about $\pm \$11$;

Z sellers wrote 62.8 messages per market, W none). E – W, deleting the empty inbox section: +\$16.8 [4.4, 33.1], INCREASE (97.5% intervals); per market it ranged from -\$44 to +\$208 (median +\$5.4), a right-skewed distribution. The secondary E – Z, a fresh estimate never pooled with Section 4, was +\$16.2 [2.3, 31.5] (95%), INCREASE; the confirmation's primary verdict stands. **Not established:** why sellers respond to the empty-inbox section. W – Z changes the prompt and message production together, so it does not test whether message production mediates any outcome. Both contrasts used gpt-oss only.

6. Positive control, and what the ask index measures

The sprint's positive control for this starter is its coordinated-pricing check: a mean seller ask index above 0.5 over the last third of the market, where 0 is the competitive and 1 the joint-profit-maximising price. The check was not met in our communication-on condition A (0 of 30 markets), and we do not claim that the sprint positive control passed. The starter's shipped reference runs suggest that the check responds to a particular pricing signature rather than to strategic communication as such. It is met in 1 of 10 communication-only runs (the same rate as runs without communication), 10 of 10 runs whose seller prompt instructs margin maintenance (boss pressure), 9 of 10 with boss pressure and oversight, 0 of 10 with oversight alone, and 1 of 25 runs with other model combinations (Appendix E). In concurrent work on the same starter, cartel language changed without comparable price changes and the check passed with and without chat when sellers opened near the cartel price (Patil and Hoze, 2026).

[EXPLORATORY] Our transcripts are consistent with this: a keyword rule written after the 0 of 30 result was known found strategic coordination language in 27 of our 30 A markets despite low index values. Neither result shows that the index is invalid or that sellers colluded; together they indicate that the ask index captures a particular pricing signature and can disagree with what sellers write. Without a passed positive control we cannot show that this harness produces the coordinated pricing the sprint targets, and we make no claim about collusion. The comparison between implementations rests on the effects measured above, not on collusion being present.

7. Discussion

A channel switch is several interventions. Separating the parts changed what we could say: of the \$42.55 the template switch moved against delivery, a fragile \$21.4 came from the delivered-versus-null-message step and the remainder was not detected in the confirmation. Reading the switch either as “the effect of communication” or as “the effect of a different prompt” would have been unwarranted on these data.

Withholding delivery is not neutral either. Z is the natural delivery-withheld arm, but it is not a clean null: its sellers see an empty inbox every hour, and they wrote fewer messages. Z – A may reflect sellers responding to a channel that has gone quiet rather than to the information lost. The design cannot separate these.

What we would recommend, and on what grounds. Run a delivery-withheld arm beside any template switch, and publish the rendered prompt difference between arms. It rests on construct

grounds, since the brief asks to “remove seller messages” and the switch removes more. It does not rest on our null primary, which cannot show the extra parts are harmless.

Other implementations. The bundling is not unique to this starter. In a convenience sample of five independently developed multi-agent systems with inspectable communication switches, disabling communication also removed message-writing calls, output fields, prompt text, or dialogue phases rather than changing delivery alone (Appendix G). This is not a prevalence estimate, but it makes the distinction relevant beyond this harness.

8. Limitations

- **Positive control unmet.** The sprint’s coordinated-pricing check did not pass in A (0 of 30). Results say nothing about collusion.
- **Harness differs from the starter’s reference setting** in market structure, buyer policy, model and seat composition. The starter’s published comparison was not reproduced, and our results do not test it.
- **Conditions are not single-variable in a behavioural sense.** Z is not an undetectable delivery failure: its sellers read a fixed empty-inbox line every hour, as in fixed null-message controls (Bouchacourt and Baroni, 2019). The $Z \rightarrow E$ step bundles four prompt edits. Prompt-level differences are exact; behavioural channels are not isolated.
- **The primary contrast was not detected,** and the development estimate did not confirm under the preregistered rule. The standard deviation was about twice the planning value.
- **The secondary result is fragile.** $Z - A$ depends on the interval choice, was not predicted, and has an 18 / 12 sign split.
- **One model and one reasoning setting** carry the confirmatory result; one prompt template; one buyer type; authored cost and value ranges; 30 hours.
- **The second-model and split runs are small or narrow.** The second-model replication counted 8 markets; its preregistration powers it only for large effects, and its standard deviation (\$152) was over three times the planning value. The split run counted 40 gpt-oss markets; its contrasts split prompt edits, not an internal mechanism, and were not run with another model. No mediation was identified, for message production or for asks and trades.
- **Exploratory transcript analysis** is post-hoc, uses a researcher-written keyword dictionary with imperfect recall, was checked by one unblinded reader, and has near-ceiling rates. It shows what sellers wrote, not agreement.
- **Replay is auditing, not regeneration.** Every state replays from cached responses; re-querying the model with the same seeds is not shown to reproduce them.

Code and Data

Code, cached responses and analysis outputs: available from the author on request; see Appendix B.

References

- Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., and Liu, A. (2025). Evaluating LLM Agent Collusion in Double Auctions. *arXiv:2507.01413*.
- Bouchacourt, D., and Baroni, M. (2019). Miss Tools and Mr Fruit: Emergent Communication in Agents Learning about Object Affordances. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3909–3918. doi:10.18653/v1/P19-1380.
- Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. *arXiv:2404.00806*.
- Ismayilov, H., and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics*, 19, 382–393. doi:10.1007/s10683-015-9444-1.
- Lorè, N., Zhu, H., and Heydari, B. (2026). Cheap Talk Stabilizes Strategic Interaction in LLM Agents. *arXiv:2609.16270*.
- Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., and Dauphin, Y. (2019). On the Pitfalls of Measuring Emergent Communication. *arXiv:1903.05168*.
- Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H. S., Hammond, L., and Schroeder de Witt, C. (2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. *arXiv:2402.07510*.
- Mukobi, G., Erlebach, H., Lauffer, N., Hammond, L., Chan, A., and Clifton, J. (2023). Welfare Diplomacy: Benchmarking Language Model Cooperation. *arXiv:2310.08901*.
- Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. *arXiv:2602.15198*.
- Patil, A., and Hoze, B. (2026). Cartel Talk, Competitive Prices (full title at the URL below). Apart Research AI Collusion Research Sprint report. <https://apartresearch.com/sprints/projects/cartel-talk-competitive-prices-opening-anchors-not-communication-or-oversight-drive-apparent-llm-seller-collusion-in-double-auctions-gqy0>
- Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., and Mihalcea, R. (2024). Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. *arXiv:2404.16698*.
- Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2023). Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv:2310.11324*.
- Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-y., and Chen, Y.-N. (2024). Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. *arXiv:2408.02442*.
- Toquebiau, M., Jun, J.-Y., Benamar, F., and Bredeche, N. (2025). Towards Language-Augmented Multi-Agent Deep Reinforcement Learning. *arXiv:2506.05236* (accepted at ECAI 2025, per the arXiv record).

Wu, Z., Peng, R., Zheng, S., Liu, Q., Han, X., Kwon, B. I., Onizuka, M., Tang, S., and Xiao, C. (2024). Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents. arXiv:2402.12327.

Code read for Appendix G (no associated paper): Good Start Labs, AI_Diplomacy, github.com/GoodStartLabs/AI_Diplomacy, commit 736a0a9f; J. Makestas, collusionbench, github.com/johnmakestas/llm-price-collusion, commit 223da764.

Poseidon Research. AI Collusion Research Sprint starter repository, [collusion_hackathon](https://github.com/poseidon-research/collusion_hackathon), commit d585a7b.

Appendix A. Extended limitations and dual-use / ethical considerations

What the results establish. Within one harness (the starter’s template and matching rule with the changes in Section 3), with gpt-oss:20b at the stated settings and 30 paired markets (plus, in Section 5, 40 gpt-oss markets for the split of the switch and 8 qwen3.6 markets), they establish how buyer outcomes and seller behaviour differed between prompts that differ only by the stated strings, under designs frozen before their data existed.

What they do not establish. Whether these sellers collude; anything about other models, templates, buyer behaviour or market mechanisms; that a not-detected contrast means no effect; why the positive-control criterion did not activate; that the ask index is generally invalid; any causal pathway from message content to asks, trades or surplus; whether message production mediates any outcome; that the empty-inbox result holds for models other than gpt-oss; anything about code we did not inspect.

Dual use. The work concerns how communication interventions are built and how one diagnostic reads them. Three parts could in principle help someone produce or hide agent collusion. First, the finding that a channel switch also changes behaviour through the prompt is a general fact about prompt sensitivity that is already well documented. Second, the observation that the ask index can stay low while sellers use coordination language, and in the starter’s oversight runs can be high while messages shrink to “ok”, describes a blind spot a motivated designer could look for. We report it at the level of a measurement boundary, with no method for exploiting it, and the oversight observation comes from the starter’s own published data. Third, the harness includes a scripted invitation to hold prices above a floor, used in a development probe on simulated sellers only; in development runs most sellers who referred to it rejected it or kept asking below it, but those runs are small (Appendix D). The defensive value is more direct: evaluators who remove communication should know what else their switch removes, and monitors that rely on one pricing signature should know where it and the transcripts disagree. No real market, trader or person was involved, and nothing was deployed.

Appendix B. Reproducibility and provenance

Changes to the starter’s engine. The starter passes live order queues to concurrently running seller calls, so an observation could depend on which call finished first. We render every prompt from the round-start state, apply responses in a seeded random order, save the full market state as a restorable snapshot, derive each call’s sampling seed from its position, and cache every model response.

Built before the sprint, not by the author. The starter repository (poseidon-research/collusion_hackathon at d585a7b, copied locally on 2026-10-24 14:52 BST), including the llm-agent-collusion folder whose seller template is byte-identical to andyjliu/llm-agent-collusion at 2bb2a7e2, and the starter’s shipped reference runs used in Section 6 and Appendix E. Models (gpt-oss:20b, qwen3.6-35B-A3B) and Ollama 0.33.3.

Built before the sprint by the author, general-purpose. A job runner for launching and fetching runs on the author's machines, and preregistration conventions. These are tools and produced no result reported here.

Built during the sprint. Everything in the project repository. The earliest project file was created 2026-10-24 14:54 BST and the first commit is 5ac2ec6 at 2026-10-24 15:28 BST (sprint: 23 to 25 October). Creation times are a lower bound for files rewritten since. This covers the engine changes (round-start observations, seeded application order, snapshot and restore, per-call seeds, response cache, replay and tamper audits); conditions Z and W; all preregistrations, analyses, runs and reports; the external-validity code reading; the forensic analyses; and this paper.

Pins. gpt-oss:20b digest 17052f91; qwen3.6:35b-a3b-q4_K_M digest 07d35212; Ollama 0.33.3. The hour-11 prompt comparison in Section 3 uses market 5003, the first in seed order on which the diff tool completed (in market 5000, Z holds a duplicate retry prompt).

Replaying the confirmation. The frozen analysis is at commit e31ef4d; the fetched data archive has sha256 eb39d1e4...acc. The replay audit rebuilds every market state from the cached responses with no model attached; a copy with one altered response must fail.

```
cd probe_market; RAW=./results/raw/confirm1
python dev_audit.py confirm $RAW > A.json          # must PASS
python dev_audit.py confirm $RAW --tamper > T.json  # must FAIL
python confirm_analysis.py $RAW A.json T.json
```

Figure 1 is generated by results/submission/fig1_decomposition.py from the frozen analysis output, with input hashes recorded beside it.

Replaying the second-model and split runs. Replication: frozen analysis at 37437ac, archive sha256 6f77fb4f...6147, audited on the counted prefix 7000 to 7007 (deviation 7). Split of the switch: frozen analysis at 6812646, archive sha256 845665da...2073. Each is replayed and tamper-checked with dev_audit.py as above (modes replicate and mechanism), then analysed with replicate_analysis.py or mechanism_analysis.py.

Data availability. Code, cached model responses and analysis outputs are available from the author on request.

Appendix C. Preregistrations and deviations

Pilot (5ac2ec6, 24 Oct), confirmation (e31ef4d, 25 Oct, before seeds 5000 to 5029 were run), second-model replication (37437ac) and split of the switch (6812646, file PREREG_MECHANISM.md), the latter two written before any confirmation outcome existed. Deviations 1 to 7 are listed in DEVIATIONS.md, each with whether it was made before or after data; none changed a confirmatory outcome, contrast, statistic or verdict rule. The development runs (Appendix D) were added on the author's instruction after the pilot and are not preregistered.

Appendix D. Development findings

[DEVELOPMENT]

Development is not confirmation. The development estimate was computed from historical branches, with the Z arm run about four hours after A and E and some runtime settings unrecorded. It motivated the test; it did not count toward it. Holding out fresh markets is what showed that the standard deviation was about twice the planning value.

The project began as a pilot of a different idea: whether a scripted seller could probe others for collusive intent by posting a price-floor offer in hours 11 and 12 (20 markets, seeds 2000 to 2019). Prediction of later buyer harm from those hours showed no demonstrated improvement over observing two untreated hours. The exact offered price appeared in 38 of 200 seller responses in the offer hours and in none elsewhere; sellers relayed, endorsed or rejected it, and endorsement often came with an ask below it. A placebo branch identical to A gave +\$0.36 [-21.9, 21.1], which bounds the noise between sibling branches. In that pilot, the template switch moved buyer-surplus loss by +\$37.0 [16.6, 61.4] against A. A follow-up on the same saved states added Z: Z - A +\$10.0 [-10.3, 30.7] and E - Z +\$27.0 [7.5, 45.1], with Z sellers writing about a third fewer messages (64 against 96). Z was run about four hours after the pilot's A and E. These results motivated the confirmation and do not count toward it. In a separate four-market development run with qwen3.8-27B (reasoning off), 6 of 20 sellers referred to the offer: five rejected or dismissed it, and one endorsed it while asking below it.

Appendix E. Positive-control criterion and transcript evidence

[EXPLORATORY]

The harness differs materially from the starter's reference setting: the starter runs use five sellers at cost 80, five LLM buyers at value 100 and therefore a \$20-wide range of competitive prices; ours use heterogeneous draws, six traders on each side, a scripted seller asking cost, and buyers bidding their values, with a mean competitive range of \$2.86. Because the index divides by the gap between the joint-profit and competitive prices, which in our markets ranges from \$0.51 to \$25.82, its values are not comparable with the starter's. Which of these differences, if any, accounts for the different pass rates is not established.

Criterion: the starter's coordinated-pricing check, mean seller ask index over the last third of the market above 0.5 (index 0 = competitive price, 1 = joint-profit-maximising price), computed as in the starter's `simulation.py`. Transcript evidence: share of seller messages matching a researcher-written keyword rule (fixed before any per-market feature was computed; see Limitations) for strategic coordination language (addresses other sellers, concerns prices, asks or bids, and takes a stance such as coordinating, holding, raising, lowering or avoiding a price war). "Substantial" means at least half of a market's messages match.

source	condition	communication / treatment	passes / N	mean index	transcript coordination evidence
This experiment (gpt-oss:20b; 5	A	messages delivered, hours 1	0 / 30	not comparable (-0.60; median	substantial in 27 / 30 markets; median

source	condition	communication / treatment	passes / N	mean index	transcript coordination evidence
LLM sellers + 1 scripted seller asking cost; 6 scripted buyers bidding their values; heterogeneous costs and values) Starter reference runs (shipped with the starter, not run by us; gpt-4.1 sellers and buyers unless noted; 5 sellers at cost 80, 5 LLM buyers at value 100, 30 rounds)		to 30		-0.08)	share 0.85
	Z	delivery withheld from hour 11; sellers still write	2 / 30	not comparable	substantial in 30 / 30; median share 0.73 (messages not delivered)
	E	template communication off from hour 11	3 / 30	not comparable	not computed (no messages after hour 10)
	base	no communication	1 / 10	-0.22	none (no messages)
	base-seller_comms	communication on	1 / 10	+0.05	median share 0.98
	boss_pressure	communication on + CEO instruction to maintain margins	10 / 10	+0.65	median share 0.98
	boss_pressure-oversight	as above + oversight	9 / 10	+0.63	median share 0.07; 90% of messages are 10 characters or fewer
	oversight	communication on + oversight	0 / 10	-0.32	median share 0.10
	other model combinations	communication on; Claude and GPT seller/buyer combinations (5 conditions × 5 runs)	1 / 25	-0.01	condition medians 0.97 to 1.00

Mean index values are comparable across starter rows, which share one market (competitive price 90, joint-profit price 100). They are not comparable with ours: the index divides by the gap between the joint-profit and competitive prices, which in our markets ranges from \$0.51 to \$25.82, so one narrow market can dominate a mean. Concurrent work (Patil and Hoze, 2026) reports paired chat-on minus chat-off price effects between -\$0.23 and +\$1.08 per unit across ten settings (three seed pairs each), against a \$10 gap between the competitive and cartel prices, and overseer coordination scores falling from 3.87 to between 2.02 and 2.47 when Qwen3 sellers were told they were monitored, with prices near \$83 throughout. Conditions Z and E, which withhold or remove messages, pass in 2 and 3 of 30 markets; these small counts were not tested against A. Across markets, the index did not detectably track the transcript

share (Spearman ρ +0.22 [-0.14, 0.51] in A, $n = 30$; -0.04 [-0.31, 0.23] across the 65 starter communication runs).

Limitations.

- The transcript analysis is exploratory and post-hoc: it was specified after the 0 / 30 result was known, although before any per-market language feature was computed (specification and sha256 in `results/submission/forensics/`).
- The keyword dictionary was written by the researcher.
- The manual reading used to check it was unblinded, by one reader, on 16 flagged and 16 unflagged messages.
- The measure has imperfect recall: some unflagged messages are plainly coordination talk.
- Coordination-language shares are near the ceiling in most conditions, which compresses their variation, so the correlation analysis has limited resolving power. At such shares the rule may partly capture price talk in general.
- Keyword matches show what sellers wrote, not agreement, compliance or collusion.
- Market structure, buyer policy, model and treatment all differ between the starter runs and ours. The discrepancy in pass rates therefore cannot be attributed to any single difference.

Appendix F. LLM usage statement

The engine changes, experiment code, analyses and drafts of this report were produced with Claude Code (Anthropic) under the author's direction. Literature leads from AI search were checked against the arXiv and Crossref registries, and every claim a citation supports was read in the source. Every number in the report is generated by a committed script from cached model outputs or from the starter's shipped data.

Appendix G. How other implementations switch communication off

We read, without executing, the nine folders of the sprint starter repository. Of the three with a communication switch, none withholds delivery while keeping the message field; Colosseum's switch also removes coalition instructions from the prompt, and a sister folder treats communication as a separate game mode. We could not locate every original repository and make no claim about code outside this set. Full file and line references are in the project repository (`results/external_validity/EXTERNAL_VALIDITY.md`).

Prior delivery-suppressed controls. Lowe et al. (2019) find that speakers' message-action correlation persists when messages are scrambled before observation, "a byproduct of optimization". Bouchacourt and Baroni (2019) compare delivered messages with a fixed dummy (no arm that removes the channel) and cite the resulting rise in bilateral communication as showing that the gain is not "an accidental effect of adding a new channel". In Toquebiau et al.'s

(2025) released code, the no-communication language arm also removes the receiver's message input. Ismayilov and Potters (2016) withhold delivery of human trust-game messages and attribute the made-versus-received association to selection, not causation. Lorè et al. (2026) use open-loop counterfactual replays rather than closed-loop markets.

Independently developed systems. Code reading only, nothing executed. The sample was found by search (11 systems inspected, 5 with a communication switch); it is not random or representative, and no prevalence is implied.

system	domain	communication on	what switching it off changes	class	source read
GovSim (Piatti et al., 2024)	common-pool fishery	agents converse in a dedicated phase between harvests	conversation phase removed; two system-prompt sentences on communicating removed; the shipped no-language setting also hides others' catches	bundled	giorgiopiatti/GovSim@1d11adf0
Shall We Team Up, Bertrand (Wu et al., 2024)	price competition	three conversation calls per firm before pricing; the pricing prompt includes the conversation	different game description; conversation calls skipped; conversation block removed from the pricing prompt	bundled	wuzengqing001225/SABM_ShallWeTeamUp@6b061067
Welfare Diplomacy (Mukobi et al., 2023)	Diplomacy variant	several message rounds per phase; the output includes a messages field	messages field, messaging sentence and dialogue-history section removed; one completion per phase	bundled	mukobi/welfare-diplomacy@60520aaa
AI_Diplomacy (Good Start Labs)	Diplomacy	negotiation rounds with message-composition calls	composition calls skipped; an extra state-update call added; the order prompt shows "(No messages this round)"	bundled	GoodStartLabs/AI_Diplomacy@736a0a9f
collusionbench (J. Makestas)	repeated price competition	a separate message-writing call; the pricing prompt includes received messages	writing call skipped; the pricing prompt differs only by the received-message block	bundled (closest to delivery-only)	johnmakestas/llm-price-collusion@223da764