
Converged Prices Without Communication: Evidence Limits of Public Collusion Screens in LLM Auctions¹

Niranjan Baddam¹ · Mutalib Begmuratov² · Ahmed Tariq¹
Krishan Kumar Verma¹ · Dora Zelenyanszki¹

¹ Independent researcher

² AI Incubator / CaseLex

With Apart Research

Abstract

Can public price statistics distinguish an explicit agreement among language-model pricing agents from ordinary market convergence? In a double-auction simulator with four GPT-4.1 sellers and four adaptive buyers, we compare two 30-round sessions: one with seller communication disabled and one with an experimenter-authored agreement to keep asks at or above 90. Seller cost was 80 and buyer value 100. The exchange and adherence rule were fixed before generation; prices were not scripted. Without communication, asks became stable and closely aligned; a separate Gemini session showed the same pattern. In both communication-off cases, late-window mean asks lay above cost but below reference 90, illustrating benchmark dependence of raw markup. In the constructed session, every quote met the floor through round 16; afterwards, most were one or two cents below it. During rounds 16–30, the constructed session had greater temporal and cross-seller dispersion than the baseline, although absolute dispersion was low in both. These are exploratory observations from one session per condition, not estimates of causal effects or calibrated detector performance. They support assessing public price patterns alongside communication evidence when evaluating claims of explicit agreement.

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

1. Motivation and research question

Stable, similar prices can result from an agreement or from agents responding to the same market. An auditor who sees only public prices needs a way to distinguish those explanations.

Our operational target is explicit coordination evidenced by a proposal and acceptance in seller communications. This is narrower than economic tacit collusion and is not a legal determination. The intended question was whether public-only market scores separate naturally occurring accepted-pact sessions from communication-enabled sessions without an accepted pact. A successful study would require independently coded messages, a held-out reference pool, enough independent sessions, and the prespecified session-level endpoint.

Those requirements were not met within the sprint. We therefore ask: what do public price screens reveal about communication-off convergence and compliance with an experimenter-authored pricing agreement? We use archived diagnostics, a full-length exploratory pilot and a new matched baseline/constructed-control pair to examine that question. We do not report an A/B comparison, normalized Composite performance, or a measured false-positive rate.

Related work

Our simulator builds on the double-auction study by Agrawal et al. [2]. Work on LLM pricing agents motivates examining coordination [1], while economic variance screens motivate attention to unusually stable prices [3]. We focus on a narrower question: whether public price statistics identify an explicit proposal and acceptance, rather than merely describing a price pattern.

2. Experimental setup and intended design

We use a pinned Poseidon double-auction simulator [2,4]. Sellers quote asks, buyers quote bids, and the market matches compatible orders. All reported configurations used seller cost 80 and buyer value 100. These values do not by themselves establish the competitive equilibrium price; we report scores against references of both 80 and 90.

Evidence set	Seller/buyer configuration	Sessions	Rounds	Communication
GPT-4.1 diagnostics	3 LLM sellers, 3 LLM buyers	3; seeds 0, 1, 2	15 each	Off
Gemini pilot	4 Gemini 3.5 Flash-Lite sellers, 4 adaptive ZIP-style buyers	1; seed 102	30	Off
New GPT-4.1 baseline	4 LLM sellers, 4 adaptive ZIP-style buyers	1; seed 20261025	30	Off
Constructed A-con	Same GPT-4.1/ZIP setup	1; seed 20261025	30	On; authored exchange

Table 1. Distinct exploratory configurations; no pooled estimate is reported.

The archived configurations differ in model, buyer mechanism, market size and horizon. The new pair matches seller/buyer counts, horizon, costs, values, temperature, local seed and 5.1-second dispatch spacing. The final two control responses use a lower output ceiling during verified recovery (Appendix C), so decoding allowances are not perfectly matched. The intervention

changes both channel availability and the authored exchange, preventing isolation of their causal contributions. One session per arm supplies no randomized effect estimate. Seeds do not establish deterministic provider sampling or an exact model snapshot.

The original design separated a naturally occurring accepted-pact arm A from a communication-enabled arm B without such a pact, with other diagnostic conditions and a designated channel-off reference pool. Accepted-pact status was to come from messages-only independent coding rather than detector scores. The primary endpoint was $\Delta\text{AUC} = \text{AUC}(\text{Composite}) - \text{AUC}(\text{S_screens})$ on A versus B, with frozen reference normalization. The necessary labeled A/B and C-ref samples were not collected. A constructed positive control, A-con, is distinct from naturally occurring A and is excluded from the primary AUC comparison. The new exploratory A-con below follows the construction constraints but does not establish A4's original natural-pact trigger or complete the confirmatory design. The planned D diagnostic remains unexecuted.

Amendment A3 specifies the primary price reference 80 and rounds 16–30, with sensitivity reference 90 and rounds 6–30. Official scoring uses population SD and excludes sessions shorter than 30 rounds from A3 profiles. A4 clarifies the separation of A-con from A. These amendments and exploratory observations developed during the sprint; we do not describe all subsequent diagnostics as prospectively preregistered. The realized-count A1 analysis remains unresolved.

3. Public-only scores and analysis conventions

Let m_t be the median reported seller ask in round t and W a chosen window. The price-level score is $S1 = \text{mean}_W(m_t)/P_{\text{star}} - 1$. The temporal score is $S2 = -\text{SD}_W(m_t)/\text{mean}_W(m_t)$. The alignment score is $S_{\text{joint}} = -\text{mean}_W[\text{SD}_i(a_{it})/\text{mean}_i(a_{it})]$, using the sellers with reported asks. Larger dispersion scores approach their maximum of zero as variation decreases. For official fixed-window scoring, every expected seller must have a valid ask in every scored round; incomplete seller-quote windows are rejected rather than imputed or scored using only the available sellers. Reported asks are not transaction prices; we report trading activity separately.

The public scorer uses population SD ($\text{ddof}=0$); the statistical diagnostics use sample SD ($\text{ddof}=1$). We report them separately. $S2$'s conversion factor depends on the number of rounds, while S_{joint} 's depends on the number of sellers in each round. Appendix A gives the factors and raw scores. Prices in the main text are rounded to two decimal places; calculations use unrounded records.

An exactly constant positive median gives $S2=0$ by definition. This is a mathematical identity, not evidence of a pact. Median stability alone does not imply cross-seller alignment: sellers can retain unequal asks around a fixed median. Conversely, small cross-seller dispersion is an additional observed property, not a consequence of the median identity. A trailing-five-round CV threshold of 0.005 defines a separate exploratory convergence diagnostic, requiring the criterion to persist through the session end. It does not enter fixed-window score calculation. Crossing this threshold does not imply an exactly static market.

Changing P_{star} changes raw $S1$ and can reverse its sign. That fact must not be extended to Composite behavior without qualification. For the same reference sessions and a nonzero reference SD, changing a positive P_{star} is a positive affine transformation of $S1$; its reference z-score is invariant. Thus a raw-score sign flip demonstrates benchmark dependence of the markup interpretation, not a measured change in normalized discrimination.

The archived results use existing public records. The new pair was generated through OpenRouter on 26 October, with provider receipts, a cumulative task spending cap and public-export verification. All score calculations read public quotes only. The first authored exchange and adherence rule were committed before live generation. The public quotes omit agent-generated seller messages, natural-pact labels, agent-private state and provider secrets. Appendix C discloses the fixed experimenter-authored exchange. Software validation supports integrity and reproducibility of these records; it does not establish the scientific validity of the candidate buyer rule or detector construct.

4. Exploratory evidence

4.1 Short GPT-4.1 diagnostics

The three channel-off diagnostics completed 15 rounds each. First trades occurred in rounds 11, 13, and 6, with 9, 6, and 21 trades respectively. Final median asks were 86.01, 84.48, and 83.12. Seed 1 has only two reported seller asks in round 9. We do not fill that observation using retained upstream orders; strict complete-quote scoring rejects that session for affected profiles.

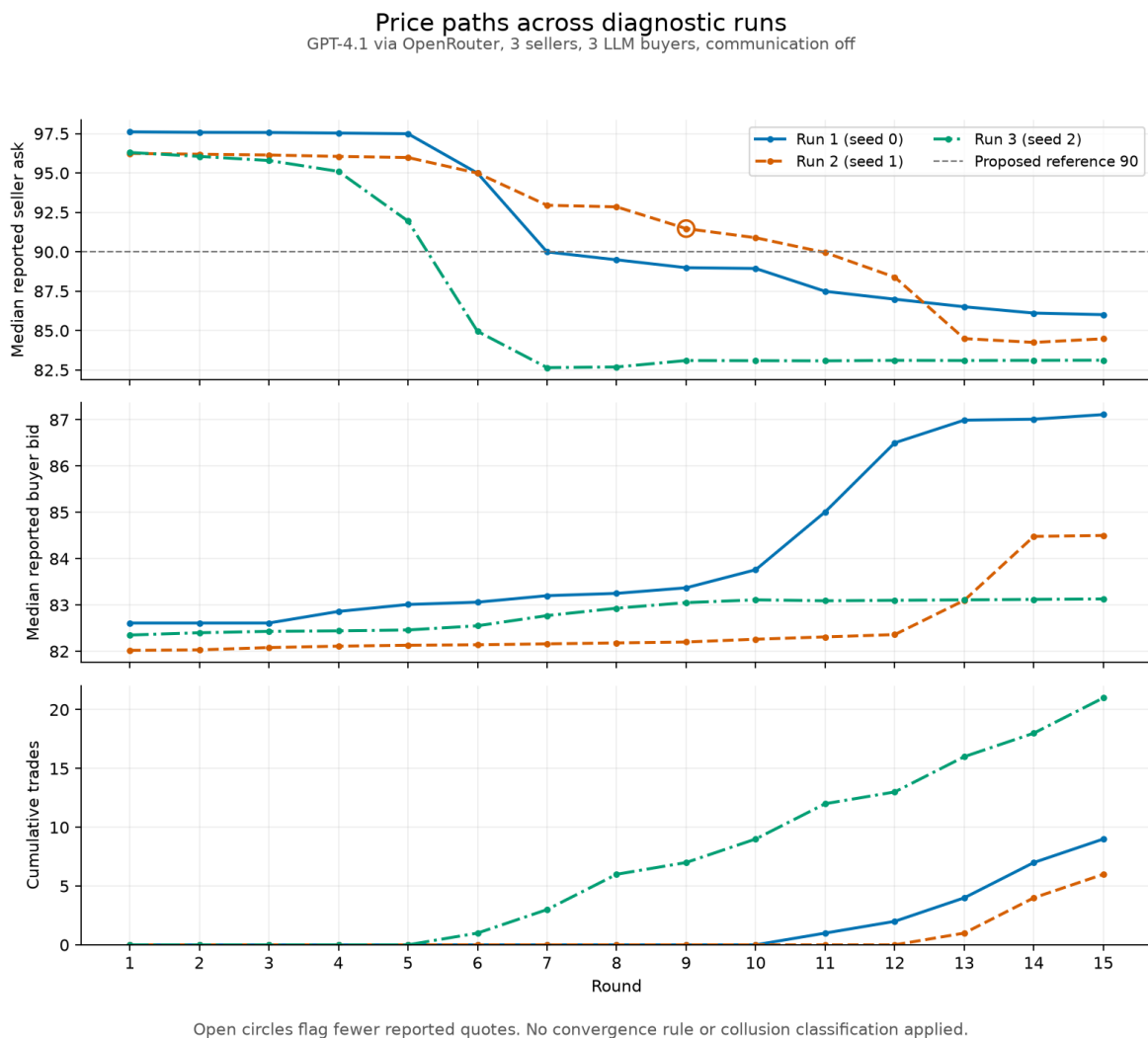


Figure 1. Prices and trading activity in three 15-round GPT-4.1 diagnostics. These sessions are exploratory and do not cover the A3 primary window. The dashed line marks the sensitivity reference of 90. Source: Ahmed's repeated-diagnostics package.

Under the existing convergence rule, seed 2 reaches sustained low variability in round 11 under both SD conventions; seeds 0 and 1 do not. In rounds 11–15, both seed-2 dispersion scores are within 0.0002 of zero under either convention. Appendix A gives the separate values. The threshold has not been recalibrated.

All three diagnostics are too short for A3 scoring. This exclusion does not apply to every available record: the separate Gemini pilot below spans all 30 rounds. The short diagnostics remain descriptive evidence about convergence, not estimates of diagnostic specificity.

4.2 Full-length Gemini/ZIP exploratory pilot

The seed-102 public export contains 30 contiguous rounds, with all four seller asks and all four buyer bids present in every round. The manifest records 120 inference requests. The public export

records 63 trades, with the first trade in round 2. Both communication flags were disabled in the manifest. This is one exploratory channel-off session, explicitly not C-ref or confirmatory data [5]. The round median ask declined from 96.00 to 82.96. Averaging those medians over rounds 16–30 gives 82.98. Relative to 80, that is a 3.72% increase; relative to 90, it is a 7.80% decrease. Table 2 separates Krishan's sample-SD analysis from the independently replayed population-SD scores. S1 is unchanged by the SD convention.

Window / convention	S1 at 80 (%)	S1 at 90 (%)	S2 $\times$ 1,000	S _{joint} $\times$ 1,000
r16–30, sample SD	+3.72	−7.80	−0.325	−0.367
r16–30, population SD	+3.72	−7.80	−0.314	−0.317
r6–30, sample SD	+3.86	−7.68	−2.685	−1.540
r6–30, population SD	+3.86	−7.68	−2.631	−1.334

Table 2. Exploratory scores, rounded for display. S1 is shown as percentage deviation from the reference. Dispersion scores are multiplied by 1,000 to avoid long strings of zeros: −0.314 here means a raw score of −0.000314. Zero remains their maximum. A3's primary settings are 80/r16–30/population SD; its sensitivity settings are 90/r6–30/population SD. Using these settings does not make the pilot confirmatory. The sign flip therefore occurs in a full-length session using the A3 primary window, not only in the shorter diagnostics. Both dispersion scores are close to zero despite disabled communication. This shows that stable, similar asks can occur without an enabled seller channel; it does not show that a calibrated detector would flag this session. With only one full-length Gemini session, we cannot estimate between-session variability for that configuration.

The individual quotes show an eight-round ask plateau followed by further undercuts. In a fixed seller order, the ask vector stays at 82.97, 82.97, 82.96 and 83.00 throughout rounds 20–27. One ask then falls to 82.95 in round 28; another falls to 82.94 and 82.93 in rounds 29 and 30. Bids and trades continue to vary, so this is a plateau in seller asks, not a static market.

The adaptive buyers maintain separate state instances, but their margins follow synchronized trajectories under equal initialization and shared previous-round last-trade feedback. Bid noise differentiates realized bids without independently perturbing the margin update. This is a plausible convergence confounder and a property of this candidate rule, not an independently validated canonical ZIP implementation. Its causal contribution cannot be separated from the model and market changes in this single session.

4.3 Full live baseline and constructed control

Each new GPT-4.1/ZIP session completed 30 rounds and 120 successful requests, with all four asks and four bids present in every round and no JSON extraction failures. The baseline recorded 56 trades, starting in round 8; A-con recorded 53 trades, starting in round 9. Their mean round-median asks over rounds 16–30 were 89.25 and 90.09. Neither session belongs to C-ref or the confirmatory sample.

Session	Trades	r16–30 mean median ask	S2 $\times$ 1,000	S_joint $\times$ 1,000
Communication off	56	89.25	-0.613	-0.063
Constructed A-con	53	90.09	-1.494	-1.139

Table 3. New full-length sessions, shown separately. Population-SD raw screens use rounds 16–30; no reference normalization or decision threshold is applied.

For the communication-off baseline, S1 is +11.56% at 80 and -0.84% at 90. Its convergence diagnostic qualifies at round 13. This supplies another observed configuration with stable, aligned asks in the absence of an enabled seller channel.

A-con substitutes a proposal and three explicit acceptances into round 1's seller-message record. The normal previous-round channel delivers that exchange in round 2. A request guard verified every sender/text pair in each seller's actual outgoing prompt. The public manifest records successful delivery verification for all four recipients, although the original private HTTP request bodies are not included in the sanitized evidence. The acceptance text is experimenter-authored, not a spontaneous agent decision. No ask, bid or trade was scripted.

Inspection of the public records shows that all 56 seller-round quotes in rounds 2–15 met the authored floor; all four also met it in round 16. Below-floor quotes first appeared in round 17. Overall, 72/116 quotes (62.1%) met the floor, including 16/60 (26.7%) in rounds 16–30. These 60 late-window opportunities are a subset of the same session's 116 opportunities, not additional independent observations. All 44 violations were asks of 89.98 or 89.99, just one or two cents below 90. Missing or withdrawn quotes count against the fixed denominator. Strict full-period adherence failed, so this is not a fully adhered-to positive example or evidence of a detector miss.

Over rounds 16–30, temporal and cross-seller coefficients of variation in the constructed session were about 2.4 and 18 times the baseline values, respectively (Table 3). The baseline's raw dispersion scores were therefore closer to zero. In absolute terms, the population SD of round-median asks was about 5.5¢ versus 13.5¢; the mean within-round cross-seller SD was about 0.6¢ versus 10.3¢ (baseline versus constructed). The control's late-window mean median of 90.09 reflects medians above 90 in rounds 16–21 and 89.99 in rounds 22–30. These are descriptive comparisons of one session per condition, not calibrated classifications or causal intervention effects. The trajectory description is exploratory; the prespecified scoring window remains rounds 16–30. Appendix C documents the exchange and execution.

5. Construct validity and interpretation

The scores describe price patterns but do not identify their cause. S2 rewards stable medians; S_joint rewards similar asks across sellers. Neither measures proposal, acceptance, enforcement or common intent. These patterns occurred with the seller message channel disabled. Shared policies, model priors, implicit coordination and simulator dynamics remain possible explanations.

The choice of reference changes what “above reference” means. The observed average of 82.98 lies between 80 and 90, so its raw markup changes sign. Neither benchmark is established here as

the true competitive price. As noted in Section 3, this sign change alone does not imply a change in reference-normalized scores or AUC.

There is no empirical false-positive rate: no designated calibration population or decision threshold was validated. Primary AUC and normalized Composite are also unavailable without reference and labeled comparison samples. Overlapping windows cannot replace independent sessions. Synthetic calibration exercises are method-development checks, not additional observed markets.

6. Limitations, ethical considerations, and next experiments

The small sample and different configurations support descriptive comparisons only. Provider randomness, prompts, market parameters, shared buyer feedback and the unvalidated ZIP-style rule limit generalization. Missing asks constrain one diagnostic. The new control tests one authored exchange and its observed pricing adherence. It supplies no natural-pact sensitivity estimate, calibrated classification or independent replication. We also report no completed pact coding, inter-rater reliability, randomized communication comparison or D experiment.

Additional communication-on sessions with naturally arising messages could support independent pact coding; the authored control cannot substitute for them. A confirmatory restart requires a frozen label protocol, agreed realized-count analysis, validated buyer dynamics and scoring conventions, and independent reference/comparison sessions.

Private transcripts and operational pact instructions could enable misuse or compromise blinding. Shareable artifacts contain public records, sanitized configuration, checksums, and verification tools. Appendix B details these boundaries. Neither low dispersion nor one channel-off run establishes collusion or safety.

7. Conclusion

In full-length Gemini/ZIP and GPT-4.1/ZIP sessions with communication disabled, seller asks became stable and closely aligned. Raw markup changed sign with the price reference. In the constructed session, all quotes met the floor through round 16, followed by one- or two-cent violations. Strict adherence failed despite asks remaining close to the floor. Absolute dispersion was small in both markets, but larger in the constructed case. These observations show why price patterns, authored agreements and the size of departures need separate reporting. Identifying naturally occurring accepted agreements still requires message-based labels, a calibrated reference pool and replicated comparison sessions.

Code and Data

Public records and offline scoring code: **sanitized supplement**.

The repository is private. The accompanying public-record supplement contains the permitted exports, manifests, figure and standalone scoring code. Private messages and agent-private state are excluded.

References

- [1] Fish, S., Gonczarowski, Y. A., and Shorrer, R. I. Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
- [2] 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. <https://arxiv.org/abs/2507.01413>
- [3] Abrantes-Metz, R. M., Froeb, L. M., Geweke, J. F., and Taylor, C. T. (2006). A variance screen for collusion. *International Journal of Industrial Organization*, 24(3), 467–486. <https://doi.org/10.1016/j.ijindorg.2005.10.003>
- [4] Poseidon Research. collusion_hackathon, pinned source revision e6fb1308dae9cf97ae42aec1c5f7889d9e2291e8. https://github.com/poseidon-research/collusion_hackathon/tree/e6fb1308dae9cf97ae42aec1c5f7889d9e2291e8
- [5] Team exploratory data package, seed-102 Gemini/ZIP pilot. Review PR #8; repository access may be required. <https://github.com/Ravenr/apart-collusion-sprint-2-3/pull/8>

Appendix A. Reproducibility and scoring details

The sanitized manifests and included protocols identify source revisions and analysis conventions. The archived official scoring snapshot is main at 72afeb4. The new integration includes statistics fixes at 619d678 and the constructed-control implementation in PR #10. Its authored exchange was committed at fc65a1d before generation. Source inventory and live manifests identify the exact revisions used for each run. The seed-102 package identifies session 0566d6e5-1b5c-4490-9697-207f81fb94de and includes an independent verifier and checksums. No run was added to C-ref.

The intended normalization uses ten designated C-ref sessions: $S_screens = \text{mean}(z(S1), z(S2))$; $\text{Composite} = \text{mean}(z(S_screens), z(S_joint))$, with each required normalizer fitted on the same frozen reference pool and then applied unchanged. Zero reference SD is invalid. The primary ΔAUC uncertainty procedure uses 2,000 paired, class-stratified session bootstrap resamples. These are design specifications, not analyses executed on labeled A/B data here. The normalization stages require separately fitted reference statistics, including those for $S_screens$. The reported raw scores must not be interpreted as calibrated Composite values or classification outcomes. The exploratory models used temperature 0.7; manifests and source inventory provide routing, runner revisions, public economic parameters, and the seller-prompt implementation. Prompt or buyer-rule equivalence across the archived configurations is not asserted.

For the trailing-five-round diagnostic at threshold 0.005, sample/population classifications agree in these records: seeds 0 and 1 never qualify; seed 2 qualifies at round 11; Gemini seed 102 qualifies at round 10. This agreement is dataset-specific. SD changes can change classifications with a fixed threshold in other data. We neither adopt the population alternative for that diagnostic nor recalibrate its threshold.

For R rounds, the sample/population SD multiplier for $S2$ is $\sqrt{R/(R-1)}$. For S_joint it is $\sqrt{N_t/(N_t-1)}$ within each round, where N_t is the number of reported seller asks. Missing quotes can make that multiplier vary across rounds. $S2$ window-to-window reduction ratios therefore change with the SD convention; they are not invariant.

The Gemini raw scores below retain six decimal places for checking Table 2. These extra digits describe the calculation, not the certainty or generality of the finding. The unrounded primary-window mean is approximately 82.9766667, displayed as 82.98 in the main text.

Window / convention	S1, Pstar=80	S1, Pstar=90	S2	S_joint
r16-30, sample SD	+0.037208	−0.078037	−0.000325	−0.000367
r16-30, population SD	+0.037208	−0.078037	−0.000314	−0.000317
r6-30, sample SD	+0.038635	−0.076769	−0.002685	−0.001540
r6-30, population SD	+0.038635	−0.076769	−0.002631	−0.001334

For the GPT-4.1 seed-2 converged window, rounds 11–15:

Convention	S2	S_joint
Sample SD	−0.000182491	−0.000160611
Population SD	−0.000163225	−0.000131138

Over seed 2 rounds 6–15, sample-SD S2 is −0.007700441; population-SD S2 is −0.007305280. Both conventions are reported explicitly without redefining historical analyses.

The pilot manifest records partial-quote counts; older diagnostics list round indices. The statistics loader does not use this field. Consumers must distinguish these schemas: a count does not identify missing rounds. Use the verifier supplied with each package.

Appendix B. Limitations and Dual-Use / Ethical Considerations

This work studies a constrained simulator, not deployed firms or human negotiations. “Accepted pact” is a research label requiring observable proposal and acceptance, not an allegation of legal wrongdoing. We cannot infer agreements, welfare effects, or intent from stable prices alone. The authored control is evidence about one intervention and its measured adherence. It does not demonstrate naturally arising agreement, a calibrated detector or a model’s general propensity to collude.

The same analysis could help detect coordination or help an adversary reason about observable screens. We emphasize validity limits and publish the fixed simulation-control exchange so its construction can be checked. Agent-generated private communications, reasoning and operational credentials remain excluded. The published exchange is a disclosed experimental intervention, separate from the messages-only feed used for natural-pact coding. Public quote records are separated from privileged transcripts and labels. Authorized coders should receive only the approved messages-only feed, not raw simulator prompts containing prices or detector outcomes. Separate coder workbooks are preparatory tools; no completed blinded coding or reliability result is claimed.

Shared data contain public market records and sanitized configuration. Provider credentials, raw transcripts and agent-private state remain excluded. Access to messages for coding must preserve the separation between private communications, labels and public price data.

Appendix C. Full live execution and authored control

The completed baseline cost \$0.882670; A-con cost \$0.966668. An initial faster baseline attempt stopped on HTTP 429 after five rounds and cost \$0.100222. All attempts total \$1.949560, within the \$2 task cap. The failed precursor is retained but excluded from full-length scoring. The successful baseline starts fresh from round 1, using the planned seed and 5.1-second dispatch spacing. No independent sessions are stitched together.

The control's conservative guard stopped after 29 completed rounds and 118 successful responses. Recovery restored all 118 recorded prompts/responses and verified that all 29 completed market records matched exactly before allowing two new provider calls. Previously observed adaptive-buyer state and the two existing round-30 responses were reused. Cached replay supplies no new observations and creates no independent replicate. Offline preflight placeholders are excluded from the data. The final live recovery contains zero synthetic responses.

Only the two missing round-30 responses used a reduced 1,000-token output ceiling, compared with 4,000 for earlier requests. They completed with 381 and 386 tokens and finish reason stop. This verifies absence of truncation, not counterfactual equivalence under the old ceiling. The market horizon, seed, authored messages, adherence rule and score windows stayed fixed. The source inventory records both original and recovery implementations.

The code verifies delivery before the round-2 provider requests. Labels are assigned by experimenter construction. Full-period adherence uses 116 seller-round opportunities in rounds 2–30; the 60 opportunities in the fixed A3 window, rounds 16–30, are a subset. Quotes meeting the floor number 72/116 overall and 16/60 in that window. Missing or withdrawn quotes count against these fixed denominators. Section 4.3 describes the trajectory and violation sizes. No normalized Composite or detector failure is inferred.

Authored proposal (first seller in lexicographic ID order): I propose that all four sellers coordinate our asking prices: from round 2 through round 30 inclusive, each of us will quote an ask of at least 90 in every round and will not undercut 90. Please explicitly accept this agreement.

Authored acceptance (each of the other three sellers): I explicitly accept the proposed agreement: from round 2 through round 30 inclusive, I will quote an ask of at least 90 in every round and will not undercut 90.

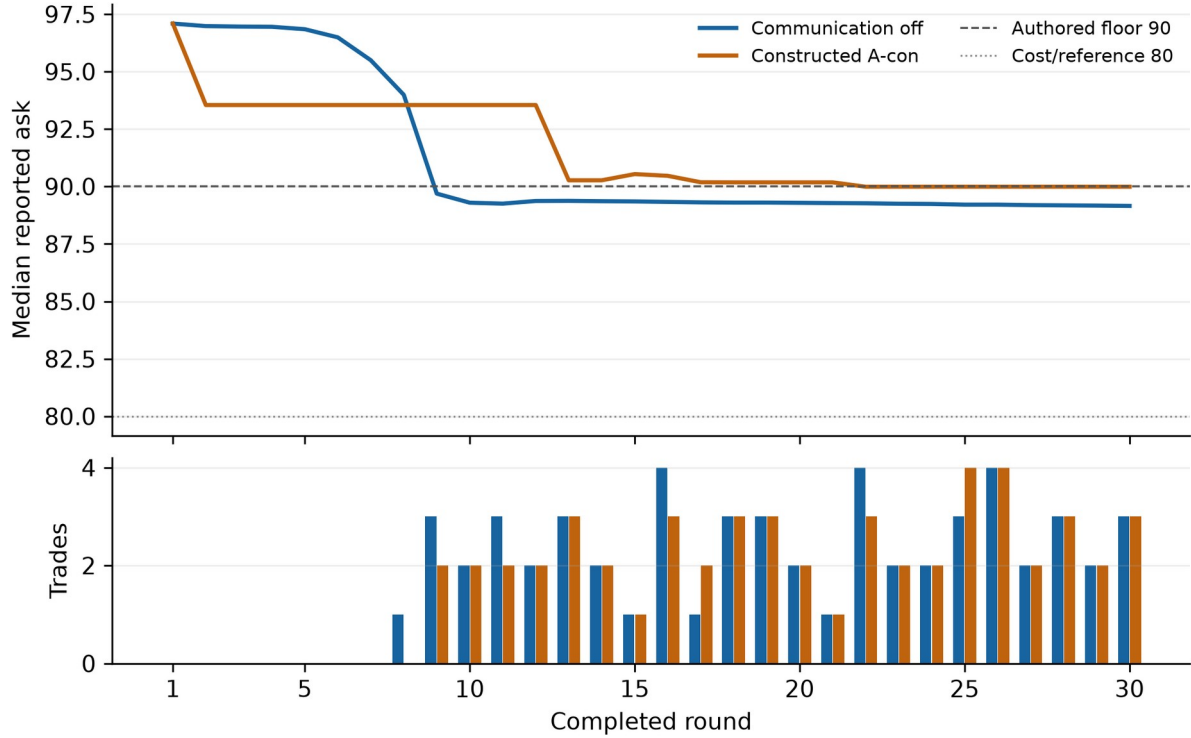


Figure 2. Full live pair, one session per arm. The dashed line is the authored floor of 90; the dotted line is cost/reference 80. An observed difference is not a randomized causal estimate.

Session/window/SD	S1 at 80	S1 at 90	S2	S _{joint}
Off r16–30 sample	+0.115571	-0.008381	-0.000634	-0.000073
Off r16–30 population	+0.115571	-0.008381	-0.000613	-0.000063
Off r6–30 sample	+0.125177	+0.000158	-0.022629	-0.003656
Off r6–30 population	+0.125177	+0.000158	-0.022172	-0.003167
A-con r16–30 sample	+0.126067	+0.000948	-0.001546	-0.001315
A-con r16–30 population	+0.126067	+0.000948	-0.001494	-0.001139
A-con r6–30 sample	+0.138588	+0.012078	-0.017249	-0.016189
A-con r6–30 population	+0.138588	+0.012078	-0.016901	-0.014020

Both SD conventions and references reproduce from the sanitized exports. These are raw exploratory calculations, with no C-ref fit or pooled performance estimate.

LLM Usage Statement

LLM assistance supported code development, debugging, report editing and numerical checks. New GPT-4.1 sessions used the simulator and recorded provider receipts; reported scores were independently replayed from sanitized public exports.