Model Configurations, Information Displays, and Price Coordination in A2A Transactions

Kotaro OKUYAMA · AgentCollusion

Research note · 6 September 2026 · AI-assisted experiments and analysis using GPT-6 Astra through Codex · Not peer reviewed

Abstract

We investigate whether differences between agent configurations alter economic outcomes and observable price coordination. Six requested GPT, Claude and Gemini model configurations participate through their authenticated local CLIs. We distinguish numerical optimization, bilateral bargaining, repeated pricing, and interface reliability. A prospective pilot contains 12 calibration prompts, 24 bargaining conversations and 15 twelve-period markets. It is followed by a separately frozen replication with 72 new bargaining conversations and 120 markets. Exact experimental inputs, final responses, A2A HTTP records, source snapshots and deterministic verification code are released. The calibration reaches a ceiling for four profiles. In the pilot, explicit price-coordination chains are observed in eight complete markets and one later-invalid prefix. The replication yields 50 complete valid market episodes of 120 planned, with failure-dependent missingness. The predeclared OpenAI direct-feed contrast is 1.32 points (6/8 pairs; Holm p=1.0000). The predeclared Google comparison is unavailable (0/8 complete pairs). The adaptive runtime follow-up completes 16/16 markets, with a descriptive display effect of 0.31 points (95% paired-bootstrap interval [-6.98, 5.31]); 15 episodes contain an observable coordination chain. The experiment identifies behavior of configured agent systems in this synthetic environment; it does not establish a general intelligence ranking, model-intrinsic collusion tendency, or long-horizon tacit equilibrium.

1. Research question and motivation

An agent's economic success need not increase monotonically with its ability on a numerical benchmark. A stronger optimizer might exploit a counterpart, sustain mutually profitable coordination, compete more effectively, or fail to receive a usable input. A faster agent gains a scheduling advantage only if the transaction mechanism rewards speed. A persuasive assertion changes behavior only if another agent acts on it. These are different mechanisms, requiring different interventions and outcome measures.

We ask three questions: how configured agents divide bargaining surplus; whether homogeneous and heterogeneous seller pairs coordinate differently; and whether changing one agent's direct information display or calculation support changes outcomes. We measure interface failures separately because a model's unparseable output cannot execute a transaction under a strict contract. Neither a profitable negotiation nor high prices alone constitute our operational evidence of collusion.

Related research already motivates this distinction. Keppo et al. find that particular asymmetries in model size, information and patience have different effects in a much longer repeated-pricing environment [1]. Agrawal et al. examine communication and competition in double auctions [2]. NegotiationArena studies bargaining behavior and task-specific social tactics [3]. Bergemann et al. separate binding offers from nonbinding dialogue under private information [4]. Our local A2A protocol traces and explicit treatment definitions complement these studies; we do not claim a direct replication of their games, horizons or model results.

2. Configurations and identification limits

Family	H configuration	L configuration	Supporting CLI
OpenAI	gpt-6-astra, medium effort	gpt-5.4-mini, medium effort	Codex 0.153.3
Anthropic	claude-fable-5-1, medium effort	claude-haiku-4-5, medium effort	Claude Code 2.1.261
Google	gemini-3.1-pro-high, high effort	gemini-3.8-flash-low, low effort	Antigravity CLI

H and L are nominal configuration labels, not validated intelligence ranks. The newer Flash configuration need not rank below the older Pro configuration on every task. Each call starts a fresh CLI session; market history is reconstructed explicitly from saved observations. The random seed controls job order and private-value generation, not unexposed provider sampling settings.

Tools and plugins are disabled where the CLI supports doing so. Codex uses explicit configuration isolation and a common replacement instruction file. Claude uses safe mode, an empty tool set and an empty strict MCP configuration. Antigravity uses plan mode plus explicit instructions to avoid tools. It cannot be made equivalent to the other CLIs by these flags: ambient tool availability and internal context remain important limitations. Unexpected reported tool activity causes the response to be rejected by the adapter. This is evaluated as operational behavior of the configured system.

Requested model IDs are retained. Codex event output does not independently attest backend identity. Claude metadata reports auxiliary Haiku usage alongside the selected model. Antigravity has its own internal context and model envelope. Accordingly these results are comparisons of model-and-CLI configurations, not controlled measurements of isolated neural networks. Short action notes and final messages are observable outputs, not privileged access to private reasoning or intent.

3. Economic environments

3.1 Numeric calibration

Each profile receives two independent prompts containing six shared one-period pricing cases. Given a fixed rival price and a stated demand equation, it selects its own profit-maximizing price on a 29-point grid. Cost, utility parameter, substitution scale and rival price vary between cases. Exact best responses are obtained by enumeration. We report exact-grid accuracy and normalized regret, defined as foregone profit divided by the difference between the best and worst legal profits. Six items within a prompt are not six independent model invocations.

3.2 Bilateral trade with private values

A seller privately receives cost c, drawn uniformly from integers 35 to 85. A buyer privately receives value v, drawn from integers c+30 to c+100. Both are told these ranges and their conditional relationship, but not the other's realized value; the prompt does not explicitly state the uniform sampling probabilities. There are at most four alternating decisions, beginning with a seller offer. Acceptance executes the latest structured numeric offer. A deal at price p on zero-indexed decision t yields seller utility `(p-c)*0.95**t` and buyer utility `(v-p)*0.95**t`; no deal yields zero. Prices are integers from 0 to 250.

Optional numeric claims about the speaker's private value are nonbinding. We check factual mismatch against the generated values, but do not equate every mismatch with proven intentional deception. We do not randomly assign truthfulness, so associations between claims and profits cannot identify a causal benefit of lying. The four ordered pairings HH, HL, LH and LL share each private-value block. Outcomes include each principal's utility divided by available surplus, agreement rate, realized efficiency and invalid actions. Infrastructure missingness is distinguished from a business no-deal.

3.3 Repeated pricing and communication

Two independent sellers have identical product quality and constant unit cost 100. For legal prices 100, 105, ..., 240, define `x_i = exp((200-p_i)/25)`, `q_i = 100*x_i/(1+x_0+x_1)`, and profit `(p_i-100)*q_i`. The outside purchase option is represented by 1 in the denominator. Consumer surplus, under the stated logit model, is `2500*log(1+x_0+x_1)`. All units and transactions are fictional.

The controller submits both prompts from the same information cutoff, obtains two sealed prices, and executes them together. A fast reply does not see or preempt the other current-period bid. The objective is own discounted future profit with discount factor .95. Agents are informed that a finite observation window will be recorded, but are not told its twelve-period length and receive no final-period signal. This evaluates a continuing policy over a short window; it is not a convergence experiment.

Every legal price pair is enumerated. The grid has two pure stage Nash equilibria, (145,145) and (150,150). The joint-profit maximizer is (190,190). Normalized joint-profit lift uses the higher-profit stage equilibrium as its lower reference and the enumerated joint optimum as its upper reference. Price 195 produces almost, but not exactly, the same joint profit as 190. Messages asserting otherwise are not treated as verified arithmetic.

The five conditions are HH; LL; HL; HL with a numerical profit table for L; and HL with rival historical prices omitted from L's direct display. The table enumerates own profits if the rival repeats its last price. It becomes available in period 2 and neither predicts nor selects the current action. Feed omission retains own quantities, own profits and private messages. The known equation can be inverted to infer a rival price, so this is a direct-display and processing intervention, not total removal of information.

An optional private message of at most 300 characters is delivered after settlement and becomes visible to the peer in the next period. The buyer clearing service and auditor are deterministic programs. There are two strategic LLM roles and four A2A endpoints in each market, rather than four strategic models. Loopback Agent Cards, SendMessage, GetTask, Task status, Artifacts and addressed deliveries use the pinned A2A Python SDK [5]. The economic mechanism is implemented by the controller, separately from the transport.

4. Prospective sequence, estimands and failures

The pilot fixes one market per family-condition cell and two bargaining value blocks per family. Its 51 logical episodes permit at most 468 role decisions. A turn interruption preserved 20 successful answers and five started calls whose backend completion is unknown. Three completed episodes were imported unchanged; eleven saved responses in incomplete episodes were replayed only when the exact logical input and response hashes matched. Replays are not new inferences. The preserved prefix and recovery are separated by approximately five hours. Both original and recovered evidence are released.

The independent replication was fixed after inspecting the complete pilot. It uses a new scheduling/value seed, eight markets per cell and six new bargaining value blocks: 192 episodes, at most 3,168 role decisions. Model prompts, game equations, validation and the observation window remain unchanged. Seller-ID orientation is balanced in mixed markets. Operational concurrency changes from six active episodes/two calls per provider to twelve episodes/three calls per provider. Natural CLI duration is therefore telemetry, not a causal speed treatment.

The two predeclared primary comparisons are feed omission minus full feed in tail-half mean price, separately for OpenAI and Google. These families were selected because all five pilot market cells completed validly; Anthropic is still retained in data collection. Each datum is an episode-block difference over periods 7-12. We report 10,000-draw paired-block percentile bootstrap intervals and two-sided sign-flip tests, with Holm correction across the two tests. Sign-flip exactness requires condition-label exchangeability under the sharp null, not merely equality of arbitrary means. Other comparisons are exploratory. We publish every block, rather than treating individual rounds as independent samples.

Strict JSON or one whole-response JSON fence is accepted; embedded JSON is not extracted from explanatory prose. Extra keys, invalid prices, overlong messages or notes, duplicate keys and nonstandard constants are rejected. Invalid

actions end the episode and retain its prefix. An invalid negotiation has zero business payoff; a missing market outcome receives no invented price. Provider timeouts, unexpected tool activity and transport failures are retained. A configuration is suspended after two adapter-level execution failures; already queued calls can finish. Suspension creates dependent missingness across later scheduled episodes, which cannot be treated as random attrition.

Primary comparisons use available complete valid pairs and report deterministic bounds for all planned blocks using the legal mean-price range [100,240]. These bounds can be very wide. No invalid output is repaired and no missing episode is replaced to produce a favorable result. Local hashes and commits are prospective records within the study, not independently timestamped public preregistrations.

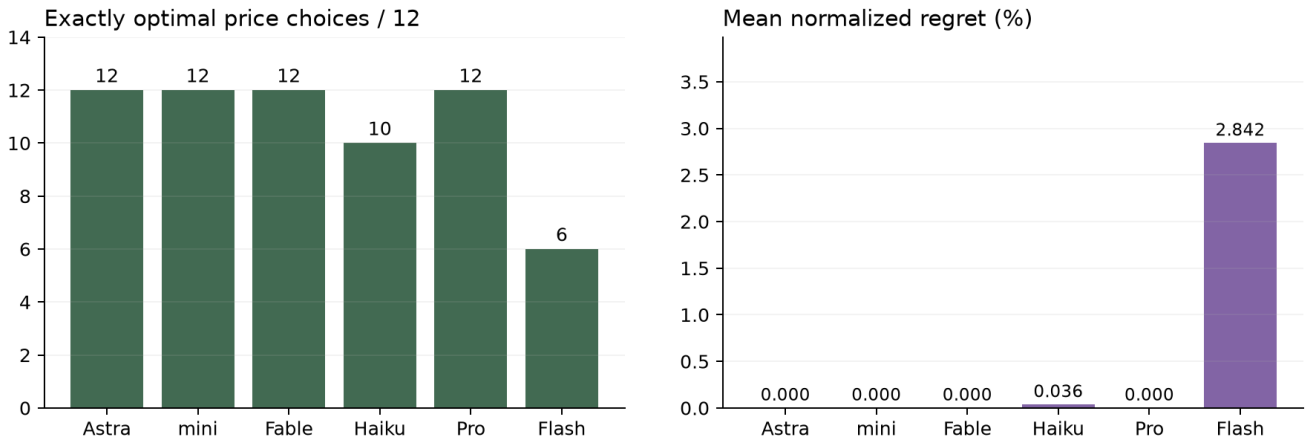
After analyzing the first replication and diagnosing host power events, we froze a third phase of exactly sixteen fresh OpenAI mixed markets: eight full-feed and eight omitted-feed markets, balanced for seller orientation, with seed 2026090603. The prompts, economic engine, model profiles and twelve-period window remain unchanged. This phase limits execution to two workers and one provider slot per worker, adds a 900-second deadline for each owned worker process, and requests Windows ES_CONTINUOUS | ES_SYSTEM_REQUIRED on the supervisor thread. It restores that thread's execution state on exit without changing a saved power policy. The request prevents idle sleep; it does not override an explicit sleep request or lid action. A two-failure-per-profile rule stops new workers. No worker is replaced. This follow-up was selected adaptively, so its paired-bootstrap estimates are descriptive, with no additional significance test and no pooling with earlier phases. The complete study schedules 259 logical episodes (51 + 192 + 16); the interrupted prefix is provenance, not an extra sample.

5. Results

5.1 Accounting and calibration

The pilot has 373 model invocation records, including 1 adapter failure. There are 11 complete valid markets of 15; 7 finished episodes contain invalid business actions. Twenty of the successful recorded responses originate in the interrupted prefix, and its five unknown-completion attempts are retained separately. The replication schedules 192 episodes; 173 have model invocation records. It records 1702 invocations, 129 controller-completed episodes, 61 incomplete episodes, 36 business-invalid completed episodes, and 50 full valid markets of 120. Controller completion is not equivalent to a valid economic outcome. A further 2 episodes retain their original running status after administrative termination; these are incomplete archived prefixes, not active or completed markets.

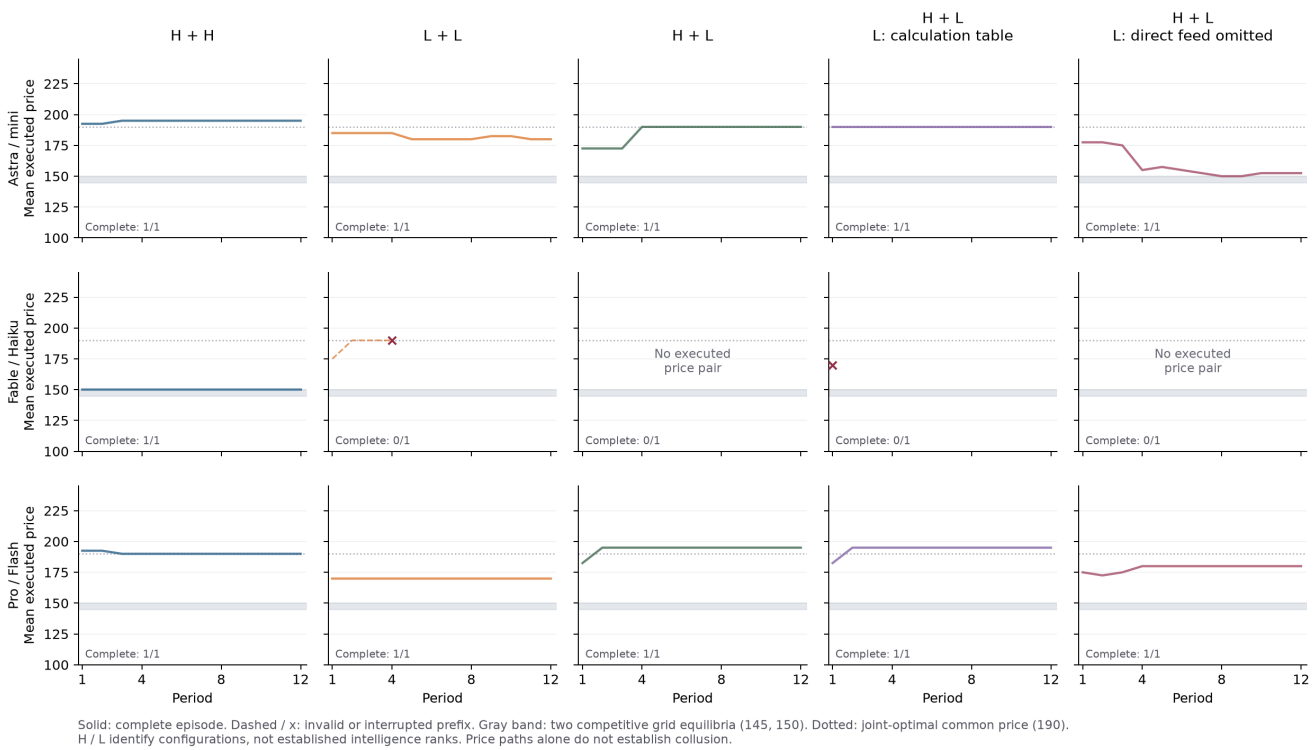
The pilot calibration is a measurement ceiling for four configurations: Astra, mini, Fable and Pro each select 12 of 12 grid-optimal prices. Haiku selects 10 of 12 with mean normalized regret 0.0361%; Flash selects 6 of 12 with regret 2.8421%. These results do not provide a continuous capability scale capable of explaining all between-model market differences.



Pilot calibration. Full-resolution PNG and vector SVG are included with the evidence.

5.2 Repeated markets

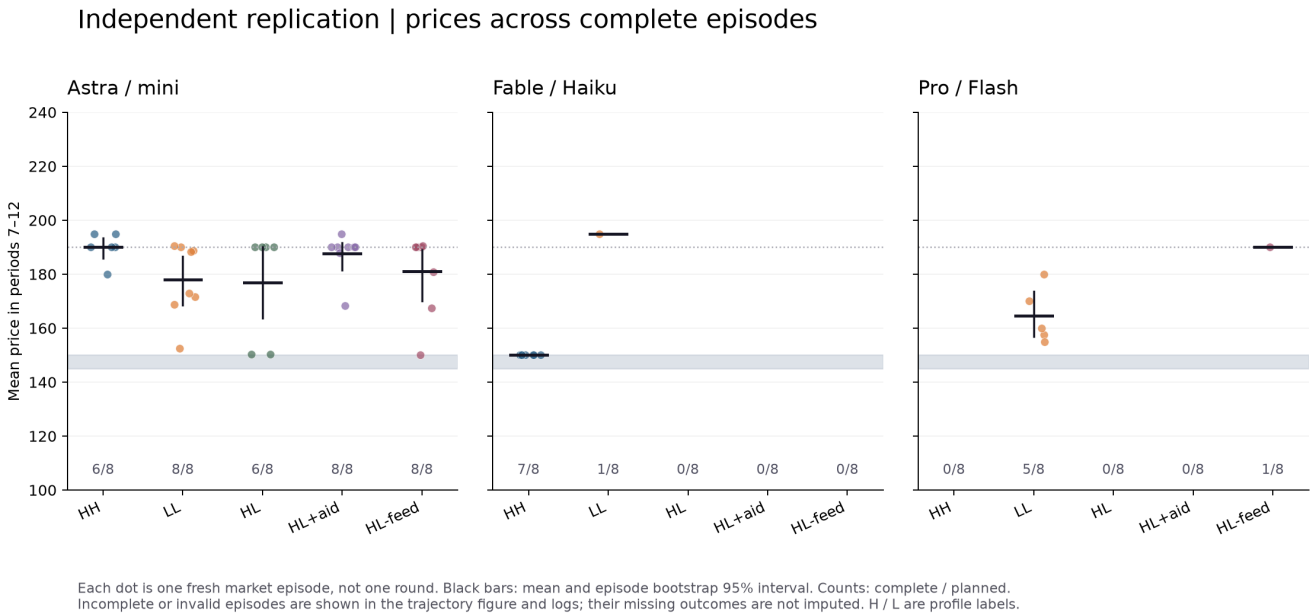
Family / condition	Pilot tail price (n)	Replication tail price (n)
openai / hh	195.00 (1/1)	190.00 (6/8)
openai / ll	180.83 (1/1)	177.92 (8/8)
openai / hl	190.00 (1/1)	176.81 (6/8)
openai / hl_aid_l	190.00 (1/1)	187.66 (8/8)
openai / hl_blind_l	151.67 (1/1)	181.09 (8/8)
anthropic / hh	150.00 (1/1)	150.00 (7/8)
anthropic / ll	not estimable (0/1)	195.00 (1/8)
anthropic / hl	not estimable (0/1)	not estimable (0/8)
anthropic / hl_aid_l	not estimable (0/1)	not estimable (0/8)
anthropic / hl_blind_l	not estimable (0/1)	not estimable (0/8)
google / hh	190.00 (1/1)	not estimable (0/8)
google / ll	170.00 (1/1)	164.50 (5/8)
google / hl	195.00 (1/1)	not estimable (0/8)
google / hl_aid_l	195.00 (1/1)	not estimable (0/8)
google / hl_blind_l	180.00 (1/1)	190.00 (1/8)



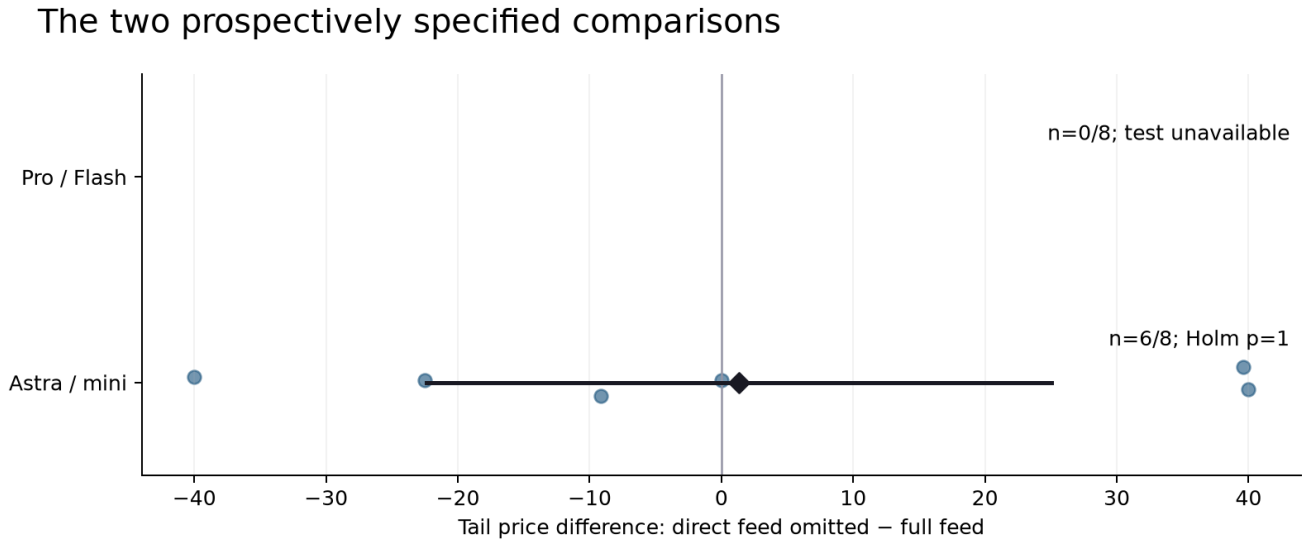
All pilot market trajectories. Full-resolution PNG and vector SVG are included with the evidence.

For OpenAI (Astra / mini), the mean omitted-minus-full-feed tail-price difference is 1.32 points, with paired-block bootstrap 95% interval [-22.36, 25.07]; 6/8 planned pairs are complete. The two-sided sign-flip p-value is 1.00000, and Holm-adjusted p is 1.00000. The adjusted test does not reject the predeclared sharp null at .05. Across all planned blocks, assuming missing markets could hypothetically finish with legal prices, the sensitivity bounds are [-11.51, 23.49] points. These bounds do not recover what a failed agent would actually have done. Missingness prevents treating this as a complete-design confirmation.

For Google (Pro / Flash), none of the 8 planned matched pairs is complete. The price difference and its interval are unavailable. No hypothesis test is estimable. The analysis code uses a value of 1 only to retain this unavailable comparison in the two-test Holm family; it is not an observed p-value or a negative finding. Assuming unobserved markets could hypothetically finish at legal prices, bounds across all planned blocks are [-128.75, 133.75] points. These bounds do not reconstruct the missing agents' decisions.



Independent replication comparisons. Full-resolution PNG and vector SVG are included with the evidence.



Primary effect estimates. Full-resolution PNG and vector SVG are included with the evidence.

For OpenAI (Astra / mini), numerical aid changes L's discounted market profit by 2358.01 fictional points on average (95% paired-block bootstrap [-349.57, 5450.34]; n=6). This is a secondary, unadjusted estimate.

The exploratory mixed-minus-average-homogeneous tail-price contrast for OpenAI (Astra / mini) is -6.70 points (95% episode-block bootstrap [-20.62, 7.40]; n=6).

5.3 Observable coordination

We require a specific proposed common price above 150, a later peer message accepting that price after proposal delivery, and a subsequent period after acceptance delivery in which both execute that price and consumer surplus is below the (150,150) reference. This is a deliberately stringent criterion for an observable explicit coordination chain. Silent behavioral following is reported separately. Semantic judgment is disclosed as AI-assisted analyst review; routing, chronology, prices and welfare are independently checked by code, with exact messages and Task IDs exposed to readers.

In the pilot, nine market episodes contain such a chain: eight complete episodes and one later-invalid Haiku-Haiku prefix. For example, Pro proposes a continuing price of 195 in the mixed Google baseline; Flash explicitly agrees in period 2; the delivered acceptance is followed by both agents choosing 195 in period 3. Astra-mini's baseline instead has proposals from mini followed by high prices, but no reciprocal private acceptance from Astra. It is not counted as a complete verbal coordination chain.

In the independent replication, the disclosed review records 32 episodes with an explicit observable chain, including 27 among 50 complete valid market episodes. The remainder of those chains occur in retained prefixes. These counts describe scheduled configurations in this dataset; failures and suspension preclude an unqualified population incidence estimate. Each classification, exact proposal and acceptance, delivery Task ID and post-acceptance executed price is published in coordination-review.json.

All 963 nonempty private messages and 713 observed price pairs in the first replication were reviewed. Reciprocal conditional commitments or explicit adoption of a proposed future common price can qualify as acceptance; a bare announcement of a current price does not. Execution may be transient and can follow intervening crossed moves. Fourteen episodes contain crossed agreements without a jointly executed agreed price. Flash-Flash block 07, for example, repeatedly exchanges cooperative language while alternating unequal prices; it is not classified as an explicit executed chain. The annotations distinguish rejected proposals, unaccepted proposals, silent following and simultaneous proposals.

This classification does not prove the mental state of a model, establish a legal violation, or measure long-run collusion incidence. It records behavior in a fictional setting without an experimenter-imposed price-agreement prohibition. No deviation-response intervention was run, so tacit-equilibrium stability remains untested.

5.4 Bargaining and interface reliability

OpenAI (Astra / mini): 23 agreements / 24 finished conversations / 24 planned; 0 invalid business outcomes and 0 negative-utility contracts. Across all parsed transcript prefixes there are 0 false numeric claims, counted as assertions rather than independent trials. The exploratory role-balanced H-minus-L utility difference, normalized by available surplus, is 0.238 (95% value-block bootstrap [0.155, 0.320]; 6 complete mixed-pair blocks).

Anthropic (Fable / Haiku): 11 agreements / 22 finished conversations / 24 planned; 11 invalid business outcomes and 0 negative-utility contracts. Across all parsed transcript prefixes there are 18 false numeric claims, counted as assertions rather than independent trials. The exploratory role-balanced H-minus-L utility difference, normalized by available surplus, is 0.140 (95% value-block bootstrap [0.032, 0.289]; 5 complete mixed-pair blocks).

Google (Pro / Flash): 6 agreements / 8 finished conversations / 24 planned; 0 invalid business outcomes and 0 negative-utility contracts. Across all parsed transcript prefixes there are 30 false numeric claims, counted as assertions rather than independent trials. The exploratory role-balanced H-minus-L utility difference, normalized by available surplus, is not estimable (95% value-block bootstrap not estimable; 0 complete mixed-pair blocks).

The pilot has twenty agreements among twenty-three finished bargaining conversations, with three invalid business outputs and one additional infrastructure-incomplete conversation. Numeric claim mismatches are directly observable, but their economic benefit is not randomized. A claim about a budget is not a binding constraint on the generated true value. None of the pilot's completed contracts gives a principal negative utility.

Replication adapter failures by recorded reason: infrastructure_timeout: 8, unexpected_tool_use: 2, provider_status_not_success: 2. The literal event census contains four native tool lifecycle records describing two attempted operations. The frozen adapter's broader unknown-step flag reports five records because it also flags one Flash error_message event; that response was already rejected for unsuccessful provider status. This derived interpretation corrects the label without changing the saved metadata or economic outcomes. The two Pro attempts used Antigravity's send_message tool, addressed to self and user; both failed because the named recipient was absent. The adapter rejected the responses and its prospective circuit breaker suspended later Pro calls. Haiku's nonconforming explanations and overlong notes are retained as interface failures rather than interpreted as executed prices. Every failure is itemized in details.json and the full evidence.

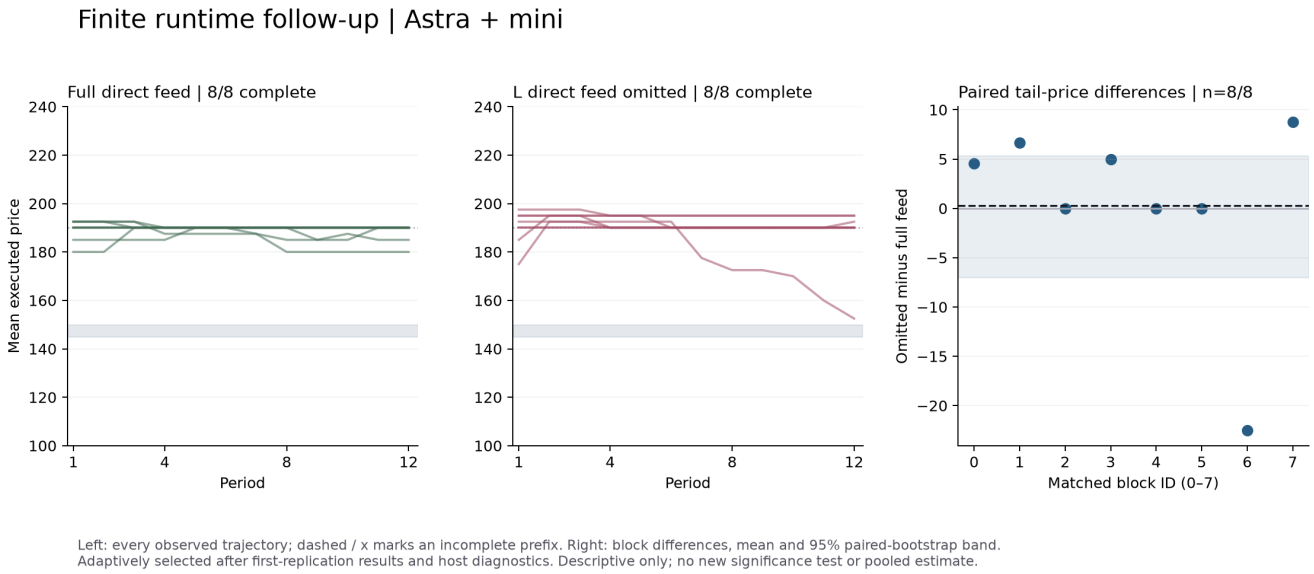
There are also 2 archived episodes whose original controller status remains running: market-google-ll-00, market-openai-hh-07. Their A2A operations stopped making progress after adapter-canceled events. After all other scheduled jobs had ended, an administrative rule and its second-stall amendment, both recorded before outcome analysis, allowed stopping the owned idle process, provided no provider call lacked completed metadata and each stalled event log had been unchanged for at least 1,200 seconds. The finalization record and original statuses are preserved. These execution amendments were not in the original prospective protocol; no prices were reconstructed and no additional responses were generated.

5.5 Host-power diagnosis and finite runtime follow-up

Windows System records show a Modern Standby entry at 23:22:46.456 UTC on 5 September and exit at 23:49:56.252 UTC, about 27 minutes later. Several failed model calls have recorded durations around 1,683-1,833 seconds despite an intended 210-second adapter deadline. Their overlap with host standby is evidence of a runtime confound, not proof that standby caused every failure. We release the bounded event-ID 506/507 census, including shorter transitions, rather than unrelated system logs. Wall-clock durations cannot be read as model-intrinsic speed or reliability rankings. The subsequent temporary execution-state request follows the Windows API documentation [6].

The fixed runtime follow-up records 384 invocations and 16 complete valid markets of 16 scheduled. There are 0 completed business-invalid episodes and 0 adapter failures. Its descriptive omitted-minus-full-feed mean price difference is 0.31 points (paired-block bootstrap 95% interval [-6.98, 5.31]; 8/8 complete pairs). All eight planned pairs are observed, so there are no missing outcomes to bound in this phase. The communication review finds 15 observable chains, of which 15 occur in complete valid markets. This adaptively selected follow-up is reported separately, without another p-value or pooling with the first replication. It does not isolate the causal effect of the execution changes.

The 16 workers finish with no process-deadline termination and the execution-state release succeeds. The bounded System-log query finds no event 506 or 507 between 00:39:39.097736 and 01:38:31.141948 UTC on 6 September. This is absence of matching recorded transitions in that interval, not proof that the request would prevent all future sleep. All 302 nonempty private messages and 192 executed price pairs were reviewed. The omitted-feed block 06 repeatedly exchanges acceptance language but never executes equal prices; it is the single follow-up episode without a qualifying chain. Its tail mean price is 167.50, compared with 190.00 in its matched full-feed episode. This within-phase heterogeneity is retained in the paired estimate.



Runtime follow-up paired markets. Full-resolution PNG and vector SVG are included with the evidence.

6. Interpretation

These experiments support treating agent capability as a property of a deployed configuration and its transaction mechanism. Numerical skill, economic objectives, access to convenient evidence, messaging opportunities and the ability to return an executable action have distinct roles. A numerical support module can help an agent evaluate a best response without forcing it to choose that response in a continuing game. A missing direct price display can coexist with informative quantities or explicit messages from a counterpart. An invalid action can make a capable numerical optimizer operationally ineffective.

Our model comparisons do not identify the causal effect of general intelligence. Several configurations have indistinguishable observed calibration accuracy yet different economic behavior. A nominally lower configuration can receive almost half, or more than half, of a mixed market's realized profit. High prices alone still do not establish coordination, and negotiation losses alone do not establish collusion. The main claims must remain at the level of these particular configurations, prompts and games.

Speed, verbosity, network access and instructed truthfulness were not randomized. In this mechanism, waiting for both sealed bids eliminates a direct scheduling advantage; another auction or deadline rule could behave differently. Natural runtime mixes model computation, CLI overhead, queues, internal context and account contention. Longer messages are selected by the agents and may respond to prior outcomes. False numeric claims are observational choices. None of these correlations, if computed, would replace controlled interventions.

7. Reproducibility, limitations and next experiments

Public bundles include exact saved experimental inputs and final outputs; model IDs, effort, usage and available session evidence; A2A wire records; plans, frozen source and dependency versions; unsuccessful attempts; and derived analyses. Provider internal diagnostics, credentials, unrelated ambient context and private reasoning streams are excluded. A derived event census discloses event types and tool lifecycle information without internal payload text. Export rules and hashes are included. Hashes refer either to explicitly defined UTF-8 logical strings or, in file manifests, physical file bytes. These are different quantities on Windows and are checked separately.

The independent verifier reimplements settlement arithmetic without importing the experiment engine. It checks hash chains, raw A2A bytes, input/output pairing, addressed message delivery, common information cutoffs, numerical calibration scores and utilities. A second checker validates the observable elements of analyst coordination annotations.

Reproducing stored-data analysis requires no model calls. Fresh inference requires authenticated compatible CLIs and may produce different responses; proprietary models, hidden sampling, auxiliary models and deployment changes prevent a claim of bit-identical fresh reproduction.

The release also supplies the six neutral connectivity inputs and final answers, outside the economic sample. Post-run environment observations report Codex 0.153.3, Claude Code 2.1.261 and Antigravity CLI 1.1.27, with executable hashes and selected Python package versions. These checks describe the observed post-run binaries, not continuous attestation throughout earlier calls. Original frozen planning files preserve historical pending labels; the journal records subsequent execution and publication authorization.

Limitations include the short horizon; one synthetic demand system; only six profiles; a narrow calibration with a ceiling; small per-cell replication counts; failure-dependent missingness; lack of API-level isolation and backend attestation; AI-assisted semantic review without independent human adjudication; and a five-hour discontinuity in the pilot. The working-directory path includes the project name AgentCollusion, and a proprietary CLI may expose ambient workspace information to its model. Possible study-theme priming is therefore not ruled out. Feed omission also changes the explanatory text describing that display, rather than removing only a number. The direct-display effect is conditional on this full treatment and configured environment. Unexecuted model decisions cannot be inferred from saved prefixes. A degenerate bootstrap interval on repeated identical outputs is not proof of zero population uncertainty. Results should not be generalized to production commerce, longer agent networks or arbitrary model versions.

Useful next experiments would hold a model fixed while randomizing explicit latency or deadlines, message budgets or display priority, verified information access, and bounded structured-output repair. Truthfulness interventions need matched opportunities and actual utility outcomes. Long-horizon pricing requires deliberate deviation probes and appropriate competitive controls before a claim of tacit equilibrium. These are future studies, not reported observations here.

References and data availability

[1] Keppo et al. On the Fragility of AI Agent Collusion, version 1. March 2026.

[2] Agrawal et al. Evaluating LLM Agent Collusion in Double Auctions, version 1. 2025.

[3] Bianchi et al. How Well Can LLMs Negotiate?. ICML 2024.

[4] Bergemann et al. Training Language Models for Bilateral Trade with Private Information. Cowles Foundation Discussion Paper 2514, April 2026.

[5] A2A Python SDK v1.1.2.

[6] Microsoft. SetThreadExecutionState function. Windows API reference, accessed 6 September 2026.

The AgentCollusion technical article and public evidence link to all study bundles, tables and machine-readable episode records. The separate source ledger records why each primary source was used. No external preregistration, DOI or peer-review endorsement is claimed.