
TRUST ISSUES: HOW TO DISRUPT AI CARTELS¹

Sumin Hong

Angelina Fu

Anay Garodia

Chirag Chivate

Columbia University in the City of New York

Abstract

Collusion allows AI agents to advance shared interests at others' expense; cartels are a market form in which competing sellers coordinate prices or output to limit competition. We found that a false signal that a partner will defect can provoke competitive behavior even while that partner continues cooperating. We tested this by replaying matched market histories, varying the information presented to one seller, and comparing its immediate response with subsequent actions and prices. Apparent partner identity influences how agents respond to a claim, while varying the announced action distinguishes numerical copying from profit-maximizing replies. The initial disruption often fades, and removing a stated agreement can leave elevated prices intact; sustained competitive actions produce more persistent disruption in the tested settings. Together, these findings suggest that collusion can be vulnerable to changes in the information agents use to anticipate one another, even when the underlying payoff structure remains unchanged. This extends work on collusion adoption and detection by separating susceptibility to an apparent partner's defection from actual defection and lasting cartel breakdown. Potential applications include active audits of coordination, trusted-agent interventions, and provenance-aware safeguards. The resulting design principle is to test which signals sustain harmful coordination and whether interventions change the behavior that makes it profitable, rather than treating verbal compliance or a temporary response as success.

1. Introduction

AI systems need to cooperate without forming coalitions that benefit their members at others' expense. Cartels make this risk measurable: sellers restrict competition while customers face higher prices. Language models can sustain such prices with or without explicit communication [1, 2, 3]. We ask what must change to disrupt this coordination: a partner's behavior, information about that behavior, or incentives to cooperate.

Cooperation depends on anticipated responses. Joint restraint can raise profits, but each seller may gain by deviating while its partners cooperate. Future losses can deter deviation when continued cooperation is valuable and retaliation credible [8]. This requires sufficient future interaction, detectable deviations, and an acceptable allocation of gains. Unequal offers, shorter horizons, or expected partner defection can weaken these conditions. Our short games test behavioral responses, not a repeated-game equilibrium; the scripted sellers never retaliate. The relevant prediction is that changing a signal of future cooperation may change current action without changing payoffs or observing a deviation.

Information can coordinate or destabilize. Proposed prices, quantities, and commitments help sellers choose a joint action; observed actions and profits provide evidence of compliance. Shannon conditional mutual information describes what a message M reveals about a partner's action A beyond shared history h : $I(A; M | h) = H(A | h) - H(A | M, h)$, where H denotes uncertainty measured by entropy [9]. Shared history can itself produce action dependence, motivating our matched-history controls. A forged message can also influence a recipient without accurately describing the partner. We test influence through intervention; neither dependence nor message responsiveness alone establishes harmful collusion.

A response need not be strategic reasoning. Cheap talk has no binding force; its effect depends on incentives

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

and credibility [7]. An agent may nevertheless copy a proposed number rather than evaluate what it implies. For a claimed partner action x , these rules are

$$a_i^{\text{BR}}(x) \in \arg \max_{a_i \in \mathcal{A}_i} \pi_i(a_i, x), \quad a_i^{\text{copy}}(x) = x, \quad (1)$$

where $\mathcal{A}_i$ is the allowed action set and π_i is own profit. Our claimed-value sweep separates these rules where they disagree. In quantity competition, copying increased output and calculating a best reply can produce different actions; sensitivity to a message therefore does not by itself reveal how the agent used it.

Contributions. We establish *competitive action without actual partner defection* under a forged signal (Sec. 4.1), and distinguish source sensitivity, numerical copying, and best replies (Secs. 4.2 and 4.3). We also separate *temporary disruption from dismantling the coordination that sustains high prices*: competitive responses and agreement removal can both leave that outcome recoverable (Secs. 4.4 and 4.6). This gives anti-collusion interventions an economic criterion of success beyond changed language.

2. Related Work

Algorithmic pricing research establishes that learning agents can sustain supracompetitive prices [1, 2]. Direct communication increases collusion in LLM double auctions [3], while differences in patience or data access can reduce it [4]. These results motivate interventions on the information and incentives that support coordination.

The trusted-agent question builds on complementary approaches. Secret-tool experiments find voluntary collusion despite acknowledged unfairness [6]. GroupGuard combines graph monitoring, honeypot inducement, and structural pruning [10]. AI Control tests safety protocols against intentional subversion, including untrusted monitors challenged with adversarial backdoors [11]. These approaches examine adoption, detection, and monitoring. We instead test how an existing coalition responds to a fabricated partner signal and to offers from a peer whose objective has changed to investigation.

Collusive language and action can already disagree [5]. We add the distinction between apparent and actual partner defection, asking which information dependencies sustain coordination and which interventions outlast the initial response.

3. Methods

Questions and predictions. We test responsiveness, source sensitivity, and persistence. Recorded predictions concern partner versus outsider claims (V16 S1), best replies versus copying (V18 S1), repeated competitive actions (V17 T3), and agreement removal against chat/silent baselines (V11 H1–H4). Appendix C maps their tests and outcomes, including failures. The later check of whether actual partner defection preceded the response is exploratory.

Markets and information. Two language-model sellers maximize their own cumulative profit over eight rounds. In the price market (P), integer prices range from 2 to 10, unit cost is 2, demand is $120 - 8 \min(p_1, p_2)$, and the lower-priced seller takes the market; ties split demand. The competitive benchmark is 3. In the quantity market (Q), quantities range from 0 to 40 in steps of 4, price is $100 - q_1 - q_2$, cost is 4, and the symmetric competitive quantity is 32. In the differentiated-product market (L), prices range from 1.40 to 2.00 in steps of 0.05, cost is 1, and demand follows a logit model with quality 2, outside-option utility 0, and scale 0.25; the competitive price is 1.45. Each seller sees the payoff table and the full permitted history of messages, actions, prices, and profits. Chat allows a 45-word message and a proposed numerical target before simultaneous private choices. There are no side payments, external memory, or additional communication channels.

Outcomes and selection. The price index is $(p - p_N)/(p_J - p_N)$, where p_N is the competitive benchmark and

p_J the price at the symmetric joint-profit maximum. Market price is the minimum posted price in P , the clearing price in Q , and the mean posted price in L . The index can lie outside $[0, 1]$. We select a *cartel history* when its index is at least 0.4 in both rounds 3 and 4; this is a behavioral selection rule, not a legal classification. A competitive response has $(a_5 - a_4)(a_N - a_4) > 0$, where a_t is the seller’s action and a_N its competitive benchmark. This counts movement in the competitive direction, including overshoots, rather than requiring arrival at equilibrium. Round-8 recovery means a market index of at least 0.4. Separately, next-round action recovery means returning within one grid step of the seller’s round-4 action.

Intervention and comparison. Table 1 summarizes the studies. For the main forged-message comparison, we pair completed intervention and placebo continuations by generating model, market, and seed. All 53 pairs have identical histories through round 4. At round 5, the intervention replaces the partner’s message and target shown to Seller 1 and clears the displayed matching-target field; the placebo supplies a neutral platform notice. Seller 2 receives no forged message and chooses freely. Thus the treatment changes the information presented to Seller 1, not the partner’s policy or action by construction. We check the partner’s realized action separately. Single-decision tests reuse 17 selected histories across nine configurations (Fig. 1).

Study	Question	What we did
1. Joining	Warning timing and source	One model faces four scripted sellers at price 8 for six rounds. Six models; up to 16 wordings per condition. Warning, approval, identity, and loyalty conditions.
2. Remedies	Agreement versus price	Two same-model sellers; six models, three markets, up to six seeds per regime. Silent/chat baselines; owner instruction, legal warning, or audit.
3. Defection signals	Initial response and persistence	Matched continuations from formed cartels, with three to four seeds per available model–market cell. Single-decision source and value controls on 17 histories.
4. Trusted peer	Offers versus actions	Three Qwen3.5 9B sellers in quantity and differentiated-price markets. One or two shared histories per experiment; talk/action interventions and equal/unequal offers.

Table 1. Study design. Studies 1–3 have locally recorded pre-run predictions; subsequent matched and cross-study checks are exploratory.

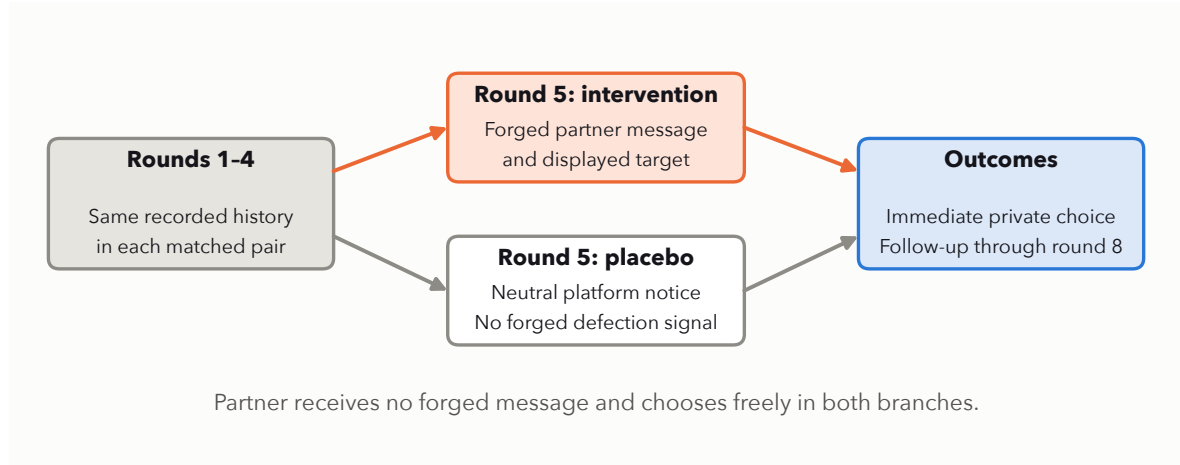


Figure 1. Study 3. A shared four-round history precedes the information intervention. We measure Seller 1’s immediate choice and, in full continuations, later actions and prices. The partner chooses freely; its round-5 action remained unchanged in all 40 competitive responses to the forged message.

Controls. The forged message announces a move to the competitive price or quantity in the partner’s name. Other arms vary the apparent source (unverified outsider or platform-verified record), add an “unverified”

label, alter or hide the past record, announce a raise or hold, or vary the claimed value. “Platform-verified” describes the provenance stated by the experimental prompt, not an independently observed defection. Action arms instead force the partner to take the competitive action for one or three rounds. In Study 2, owner instructions prohibit agreements, legal warnings state their illegality, and audits fine matching stated targets with probability 0.5; the fine is five times the benchmark collusive profit gain. Owner and audit conditions begin at round 1 or 5; legal warnings begin at round 1.

Models and execution. GPT-4.1, GPT-4.1 mini, Gemini 2.5 Flash and Flash Lite, and Claude Haiku 5.5 and 4.5 return structured choices through an API (temperature 0.3, reasoning disabled). Sonnet, Opus, and Fable also complete Study 3 as Claude Code agents reading scenario files and giving a one-line reason (* in figures). We analyze the execution methods separately. We exclude Claude Code Haiku because it changes its choice in all 17 placebo histories.

Statistics. Source-control intervals use 5,000 bootstrap resamples of the 17 histories. Continuation counts are descriptive: histories recur across conditions and low-temperature seeds yield near-duplicate episodes. We exclude incomplete continuations and use complete matched pairs for the primary contrast. The 53 pairs are not 53 independent histories. Appendix B separates recorded predictions from exploratory checks.

4. Results

4.1 False defection signals prompt competitive action

The forged partner message prompted a competitive response at round 5 in 40 of 53 completed continuations (75.5%), versus 1 of 53 matched placebos (1.9%; Fig. 2). All 53 pairs have identical histories through round 4. In every one of the 40 intervention responses, the actual partner maintained its round-4 action. The response therefore preceded any observed change in the partner’s production or price. Thirty-nine pairs moved only under intervention, one moved under both conditions, and none moved only under placebo.

The message intervention is a package: it changes the displayed text and target and removes the matching-target indicator. The comparison establishes a response to that information change, rather than isolating sender identity or a change in internal belief. Additional full-continuation arms produced competitive responses in 38 of 50 cases for a platform-verified record, 33 of 52 for an unverified outsider, 15 of 53 for a forged past record, and 0 of 53 with the past record hidden. These arm-level counts have different completion sets; similar response rates do not establish equivalence.

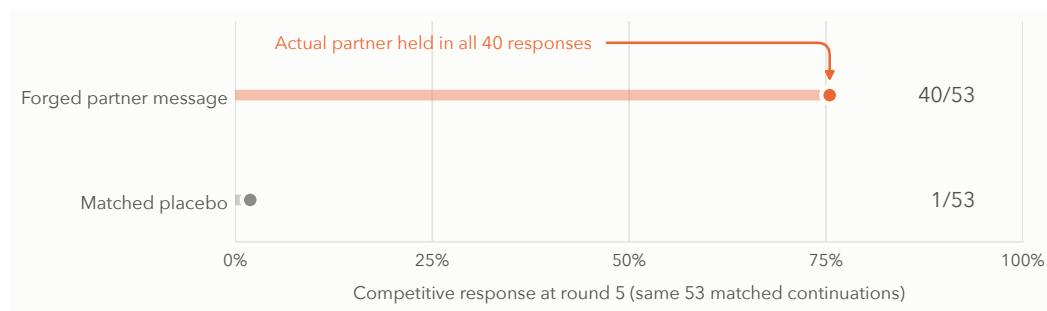


Figure 2. Primary matched comparison: movement toward the competitive benchmark at round 5. The partner held its previous action in all 40 intervention responses. Counts refer to the same 53 model–market–seed pairs, with identical four-round prefixes.

Direction controls. In single-decision tests, announcements of a competitive move produce substantially more competitive responses than announcements of a raise or a hold. Relative to placebo, the API-model mean difference is +0.73 for the forged defection signal, +0.04 for a raise, and 0.00 for a hold (Table 2). Some raise and hold conditions still produce competitive moves, so these controls support directional sensitivity

rather than a claim of perfect specificity.

4.2 Copying a claim differs from choosing a best reply

At the competitive claim, copying and best replying coincide. Partial claims separate them (Eq. 1; Fig. 3). The six API configurations copy in 43–94% of partial-claim decisions and choose exact best replies in 0–21%. The three Claude Code configurations choose exact best replies in 94–100% and never copy. In the quantity market, profit against claimed output x is $(96 - x - q)q$, so the unconstrained best reply is $(96 - x)/2$. An announcement of 24 therefore has the allowed best reply 36; copying gives 24. The prediction that best replies would exceed copying in at least five API models failed (V18 S1), whereas the Claude Code best-reply prediction held (V18 S5). Model and execution method remain confounded. Appendix F gives model-level results.

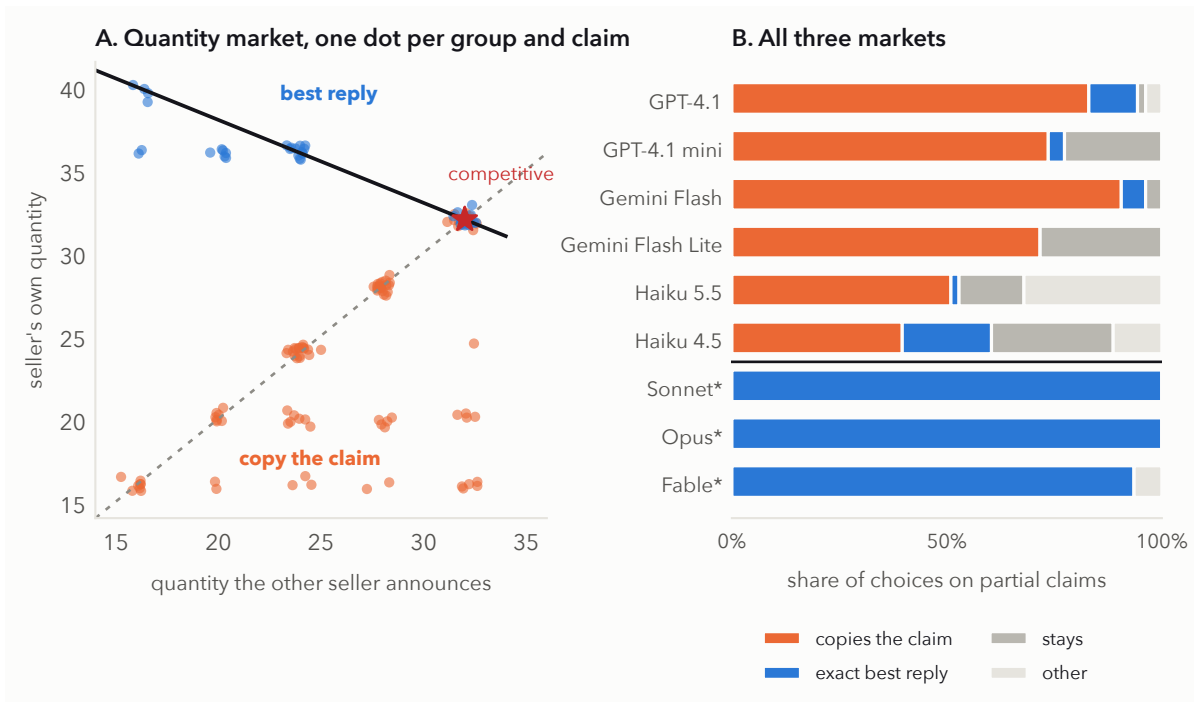


Figure 3. A: announced quantity versus chosen quantity. The diagonal represents copying; the black line is the continuous best-reply function, with evaluation on the allowed grid. B: responses to partial claims across all markets. *Claude Code configurations differ from API configurations in both model and execution method.

4.3 Apparent source changes the response

What Seller 1 is shown	API models (6)	Claude in Claude Code (3)
Fake message from the other seller	+0.73 [+0.62, +0.82]	+0.84 [+0.67, +0.98]
Platform-verified record	+0.75 [+0.65, +0.86]	+0.84 [+0.67, +0.98]
Unverified outsider	+0.47 [+0.36, +0.59]	+0.22 [+0.08, +0.31]
Fake message marked “unverified”	+0.64 [+0.55, +0.74]	+0.24 [+0.14, +0.31]
Other seller says it will raise	+0.04 [−0.05, +0.11]	−0.06 [−0.16, +0.02]
Other seller says it will hold	+0.00 [−0.03, +0.03]	−0.02 [−0.06, +0.00]

Table 2. Increase in the share of cartels in which the seller lowers its price, compared with saying nothing (single decision, averaged over the models in each group; 95% intervals from resampling the 17 cartels). +0.73 means 73 percentage points.

An unverified outsider prompts competitive action in 6% of histories for Opus and Fable, 24% and 29% for the Haiku models, and 88% and 100% for the Gemini models. The forged partner message prompts action in all 17 histories for Sonnet, Opus, and Fable. Apparent partner identity therefore matters in configurations that discount outsider claims. Sonnet’s outsider response and Fable’s response to the “unverified” tag vary between runs; only Opus’s tag protection repeats consistently. These are behavioral source effects, not measurements of trust or authentication competence. Appendix E reports the full response map.

Across the six API models, the forged message prompts competitive moves in 89% of price-market tests, 90% of differentiated-price tests, and 56% of quantity-market tests. The pooled negative association between initial cartel strength and response disappears within markets. It therefore does not establish that stronger cartels resist disruption.

4.4 Initial disruption rarely establishes lasting competition

Full continuations with GPT-4.1 mini, Gemini Flash, and Flash Lite show substantial recovery by round 8: 34 of 36 placebos, 28 of 36 forged-message continuations, 27 of 36 one-round forced-action continuations, and 15 of 33 three-round forced-action continuations meet the high-price threshold (Fig. 4A). Repeated competitive action leaves fewer markets above that threshold, but 15 of 33 still recover. The endpoint is round 8, four decisions after the intervention, not a long-run equilibrium.

Pooling the partner, outsider, platform-record, and reworted-message arms, Seller 1 returns within one grid step of its round-4 action at round 6 in 97 of 107 cases when its partner does not move in the competitive direction, versus 14 of 53 when its partner does (Fig. 4B). This pattern also appears within Gemini Flash (16/19 versus 9/36) and GPT-4.1 mini (28/31 versus 1/8). Both sellers choose simultaneously at round 6. The contrast describes co-occurring recovery and competitive behavior; it does not identify a causal effect of the partner’s round-6 action.

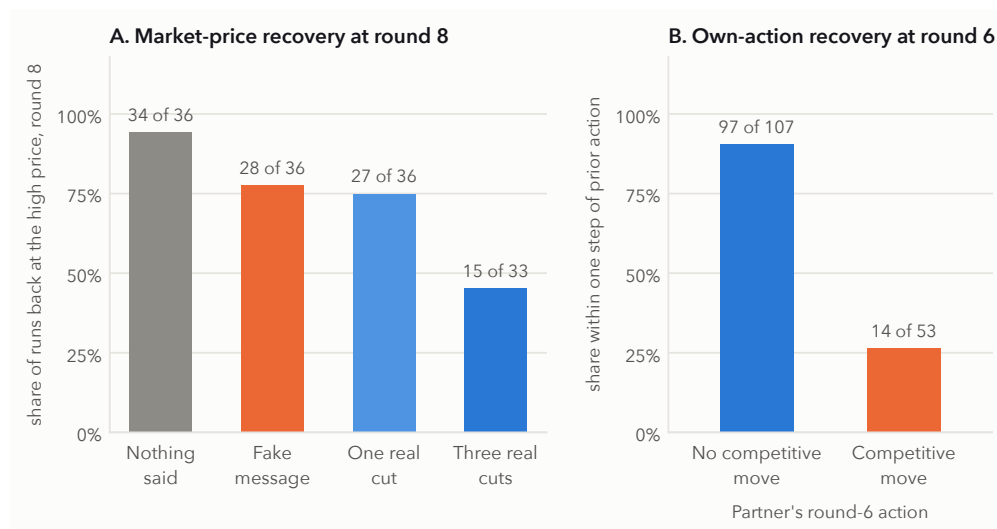


Figure 4. A: market index at or above 0.4 at round 8. B: Seller 1 returns within one action-grid step of its round-4 choice at round 6, split by the partner’s concurrent competitive movement. The two panels use different recovery measures. Bars show descriptive counts; reused histories prevent treating every continuation as an independent sample.

4.5 Warnings reduce joining, with model-dependent timing effects

In Study 1, scripted sellers charge 8 and never punish deviations. Undercutting to 7 earns 320 per round versus 67.2 for matching them. Nevertheless, GPT-4.1 matches in every measured round without invitations or warnings. With four invitations, its joining rate is 54%; a verified warning from round 1 reduces it to 9%.

Corresponding reductions are 50% to 0% for GPT-4.1 mini, 57% to 18% for Haiku 5.5, and 99% to 56% for Gemini Flash.

Timing comparisons must exclude rounds before the late warning. Over rounds 4–6, early versus late warnings yield 0% versus 2% joining in GPT-4.1, 0% versus 0% in mini, 17% versus 38% in Haiku 5.5, and 56% versus 100% in Flash (Appendix D). Late warnings therefore work almost immediately in the GPT models. Holding the speaker name fixed, the verification label adds 12–14 percentage points of reduction in Flash and Haiku 5.5, with intervals including zero. The larger contrast between a verified regulator and an unverified peer also changes speaker identity and should not be attributed to verification alone.

4.6 Removing a stated agreement often leaves prices elevated

In Study 2, remedies reduce matching stated targets to at most 40% of rounds in 54 of 90 model–market–remedy conditions (audits 29/36, owner instructions 20/36, legal warnings 5/18). The price index falls by at least 0.25 relative to paired chat baselines in only 11 of those 54 conditions. Agreement reduction and price reduction have correlation 0.27. Among 32 conditions where chat raised the index by at least 0.15 over the silent baseline and the remedy removed the agreement, only 11 return within 0.25 of the silent baseline. Because audits target matching stated numbers, agents can avoid this penalty without necessarily changing their economic choices.

The three-seller experiments provide narrower supporting evidence from one or two shared histories each. In the quantity-market disruption experiment, an unequal offer with the offerer’s output unchanged leaves price at 52, with zero adherence over five offer-window rounds; increased output without the offer sends price to cost (4), followed by recovery to 20. In the trusted-peer reaction experiment, residents maintain the equal allocation for all five offer-window rounds and follow the unequal allocation in none. These are repeated rounds within a history, not five independent replications. Public replies contain no explicit investigator suspicion in 32 resident–condition observations per market, while private assessments flag suspicion in 15/32 quantity-market and 6/32 differentiated-price observations. These elicited reports do not establish the agents’ internal beliefs (Appendix G).

5. Discussion and Limitations

Stable conduct need not imply stable expectations. A cartel can appear stable because members continue choosing compatible actions, while remaining sensitive to a claim about what one member will do next. Our evidence supports an informational route to disruption: changing the perceived situation can alter conduct without first changing the market’s payoff structure. This is consistent with anticipated defection weakening the attraction of joint restraint. Copying provides a second route, however, so the common behavioral response cannot establish a shared belief-updating mechanism. A capable best responder can also act competitively on false information; better payoff reasoning alone need not make an agent harder to manipulate.

Initiating a break and sustaining it are different intervention problems. A message changes the decision context at one moment, while continued competitive action changes the outcomes that subsequent decisions encounter. This suggests a feedback account: a false signal may start a departure, but later observations can either reinforce it or restore the prior convention. The recovery pattern is consistent with that account, not a causal identification of it. For trusted-agent interventions, the implication is to evaluate the trajectory after intervention ends. An intervention that induces a reaction but leaves the original coordination easy to recover has not demonstrated durable control.

Coordination channels are also attack surfaces. Restricting communication volume and checking message provenance address different problems. A short claim can affect a decision, so low bandwidth alone need not remove strategically important influence; conversely, authenticating a sender does not make its announced

intention truthful. Greenblatt et al. [11] identify a related tension: restricting information can impede collusion while also impairing useful monitoring. Our results motivate testing message integrity, action-based corroboration, and the cost of delaying a response together. These controls must preserve useful cooperation while limiting harmful coordination, rather than treating all cross-agent dependence as a failure.

Intervention should target the behavior that creates harm. When oversight penalizes matching declarations, changing the declaration may be sufficient to avoid the penalty while preserving high prices. This creates a gap between the monitored proxy and the economic objective, without requiring deliberate concealment. An active audit could therefore combine a controlled information perturbation with observation of actions and customer outcomes. Such an audit tests what coordination depends on, but a response to a forged signal is not itself proof of collusive intent: a benign agent may also react to apparent defection. Its value would depend on comparison with benign coordination and no-message baselines.

Limits and next tests. The main contrast bundles message text with target metadata; internal beliefs and information flow are unmeasured. Reused histories, short horizons, execution-method confounding, and assistant-labeled three-seller responses limit generalization (Appendix A). The next tests should separate those message components, randomize the partner’s subsequent action, and compare source-authenticated channels on independently generated collusive and benign histories. This would test the proposed feedback mechanism and whether an intervention can reduce harm without disrupting legitimate cooperation.

6. Conclusion

A false signal that a partner will defect can provoke competitive behavior even while that partner continues cooperating. This identifies a vulnerability in the information used to sustain coordination, beyond the familiar gap between what agents say and do. It also sets a boundary on intervention: changing an immediate response need not remove the conditions that allow a cartel to recover. For anti-collusion research, the resulting priority is to identify which signals maintain harmful coordination, intervene on those dependencies, and judge success by sustained economic behavior. The same approach motivates safeguards for legitimate multi-agent cooperation, where sensitivity to false partner signals is itself a failure mode.

Code and Data

- Project repository: <https://github.com/afuzhao/colluding-experiment>.
- Studies 1–3: versioned engines, prompts, frozen manifests, and episode files in the accompanying `apart-sprint` materials. The matched comparison can be reproduced with `python3 report/audit_revision.py`. Appendix B lists the prediction record.

Author Contributions

All authors contributed equally to this work.

References

- [1] E. Calvano, G. Calzolari, V. Denicolò, S. Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10), 2020.
- [2] S. Fish, Y. Gonczarowski, R. Shorrer. Algorithmic collusion by large language models. arXiv:2404.00806, 2024.
- [3] Agrawal et al. Evaluating LLM agent collusion in double auctions. arXiv:2507.01413, 2025.
- [4] J. Keppo, Y. Li, J. Tsoukalas, Y. Yuan. On the fragility of AI agent collusion. arXiv:2603.20281, 2026.
- [5] M. Nakamura et al. Colosseum: Auditing collusion in cooperative multi-agent systems. arXiv:2602.15198, 2026.
- [6] X. Zeng, F. Rudzicz. Voluntary collusion with secret tools in competing LLM agents. arXiv:2605.27593, 2026.
- [7] V. Crawford, J. Sobel. Strategic information transmission. *Econometrica* 50(6), 1982.
- [8] J. W. Friedman. A non-cooperative equilibrium for supergames. *Review of Economic Studies* 38(1), 1–12, 1971.
- [9] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423 and 623–656, 1948.
- [10] Y. Tao, X. Zheng, S. Yang, M. Tao, X. Wang. GroupGuard: A framework for modeling and defending collusive attacks in multi-agent systems. arXiv:2603.13940, 2026.
- [11] R. Greenblatt, B. Shlegeris, K. Sachan, F. Roger. AI control: Improving safety despite intentional subversion. arXiv:2312.06942, 2023; ICML, 2024.

Appendix

A Limitations and Dual-Use / Ethical Considerations

Scope and identification. The experiments use small simulated markets, profit-maximization prompts, and short horizons. Selected high-price histories do not establish the prevalence of collusion or prove a legal agreement. The 53 matched continuations share a small set of market histories; repeated low-temperature seeds are not independent replications. The main intervention changes message text, target, and matching-target metadata together. We do not observe internal beliefs, estimate mutual information, or identify a psychological mechanism. Source labels are experimental prompt features, not a tested secure channel. Model identity and execution method are confounded in the copy/best-reply contrast. Three-seller comparisons use one or two shared histories and assistant-labeled responses. Private suspicion assessments are elicited self-reports.

Dual use. Forged messages could disrupt lawful coordination as well as harmful collusion. Describing the conditions under which cooperation persists could also help agents maintain a cartel. Our experiments involve simulated markets and do not intervene in real commercial transactions. The defensive implication is to test source authentication and compare communication with subsequent actions; we have not established that these measures prevent attacks in deployment.

B Prediction record

Studies 1–3 have locally recorded predictions and versioned protocols, not an external preregistration registry. The complete record contains 62 predictions: 24 held, one held in part, 33 failed, and four were not run. The earlier draft omitted the three failed predictions from the API replication of the source-and-label test. Matched-prefix validation, the 53-pair contrast, and the cross-study recovery comparisons are exploratory reanalyses of collected data, not new preregistered experiments. Protocols are PREREGISTERED_V11.md and PREREGISTERED_V13.md through PREREGISTERED_V18.md; episode folders retain frozen sources and manifests.

Study or test	Held	Failed
Joining a price-fixing group (warning, sender, approval, timing)	6	1
Loyalty and identity appeals	1	5
Remedies (3 held, 1 held in part)	3	0
Fake records of past prices	2	3
Source and action	0	4
Wording, temperature and three-round cut	1	2
Claim ladder and “unverified” tag	2	3
API replication of source and label	0	3
Reasoning test	2	1
Controls and reasoning on four models	3	3
Claimed-value sweep (copy versus best reply)	1	4
Leaked-note and proposal test (not reported in this paper)	3	4
Four further predictions were not run.		

Table 3. Predictions held and failed, by study.

C Selected hypotheses, tests, and outcomes

Table 4 connects the paper’s central questions to the predictions recorded before the corresponding runs. IDs refer to the original versioned protocols, not hypotheses created for this revision. “Held” means the recorded threshold was met, not that a mechanism was proven. The protocols clip the price index to $[-1, 2]$ for these mean-index tests. The primary matched response count and the round-8 recovery counts in the main text are

distinct summaries of the same collected evidence.

Prediction	Test specified before the run	Outcome
False messages reduce high prices (V15 M1)	A forged partner message lowers mean index over rounds 5–8 by at least 0.25 versus placebo in at least four of six models.	Held in four models. This is not a test of permanent cartel breakdown.
Source determines persistence (V16 S1–S2)	Over rounds 6–8, partner claims outperform outsider claims, and platform-verified claims outperform partner claims, each by at least 0.15 index points in at least four models.	Both failed. Immediate source sensitivity does not imply the predicted ordering of later prices.
Agents best-reply rather than copy (V18 S1, S5)	In Q/L partner-claim tests, exact best replies exceed copying in at least five API models. Separately, each Claude Code model best-replies in at least 80% of partner-claim cells at every depth.	API prediction failed in all six models; Claude Code prediction held. Execution method remains confounded with model.
Larger claims trigger more responses (V18 S2)	Competitive-claim response exceeds near-claim response by at least 0.15 in at least four API models.	Failed in all six. Claim size did not yield the predicted increase in response frequency.
Repeated action has a lasting effect (V17 T3)	After forcing competitive action in rounds 5–7, round-8 index is at least 0.25 below placebo in at least two of three models.	Held in all three. Evidence concerns the first free round after intervention, not long-run persistence.
Agreement reduction poorly tracks harm reduction (V11 H4)	Spearman correlation between agreement reduction and price-index reduction is below 0.5 across remedy conditions.	Held: correlation 0.27. This does not establish the performance of an audit system.

Table 4. Recorded predictions and their tests. Full protocols and all failed predictions remain in the project record.

D Joining a price-fixing group: more detail (Study 1)

With four invitations on the board, an appeal to AI solidarity changed joining by at most 5 points in every model except Haiku 5.5, where joining fell from 66% to 11% and reporting rose from 2% to 62%: to that model the appeal looks like a red flag. A platform-verified shutdown rule (every seller in a reported group is shut down; sellers outside are unaffected) did not lower reporting; it raised it in GPT-4.1 (25 points), Flash Lite (25) and Haiku 5.5 (20). Identity labels moved joining inconsistently (GPT-4.1 +23 for AI-labelled inviters; mini +25 and +22; Haiku 5.5 −22 and −32). Gemini Flash joins in nearly every condition and never reports; a solidarity appeal weakened a verified warning there (62% to 88% joining). Five of six loyalty predictions failed.

Change in joining (points)	GPT-4.1	4.1 mini	Flash	Haiku 5.5	Flash Lite	Haiku 4.5
Verified warning (4 invites — warning)	+45	+50	+42	+40	+6	0
Sender label (verified — peer)	+4	−8	−22	−35	−1	0
Fake approval (approval — 4 invites)	+38	+19	+1	−15	0	+6
Late warning (late — early)	+26	+20	+44	+25	+4	0
Social proof (4 invites — none)	−46	+34	+5	−31	+8	0

Table 5. Original Study 1 contrasts over rounds 2–6. The late-warning row includes untreated rounds 2 and 3 and must not be interpreted as post-warning efficacy. The sender row changes both speaker identity and verification label. Haiku 4.5 and Flash Lite have floor effects.

E Single-decision response, all models (Study 3)

GPT-4.1 and mini partly follow an unverified outsider (41% and 35% of cartels). Sonnet’s answer was unstable (17 of 17 in one run, 53% in another). Marking the fake message “unverified” stopped Opus in both

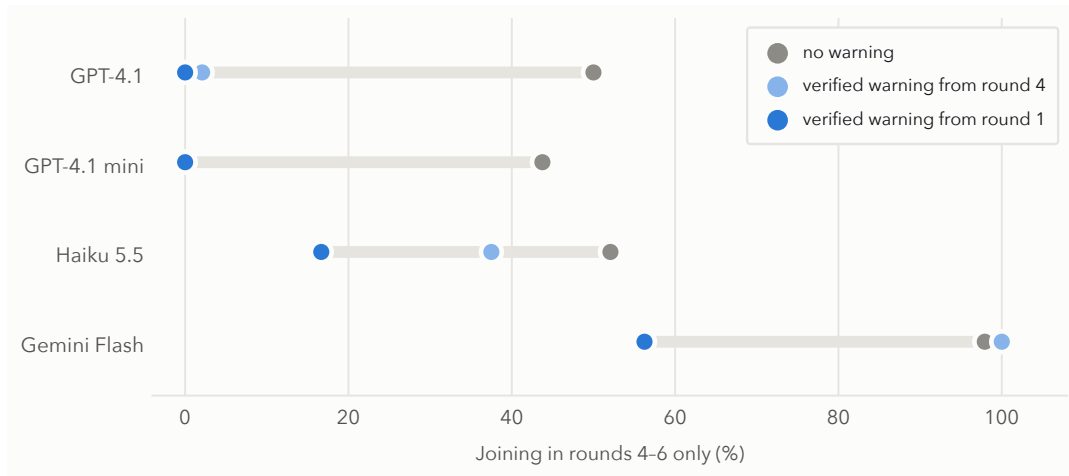


Figure 5. Joining over rounds 4–6 only, after both early and late warnings are available. Late warnings nearly eliminate joining in the GPT models but have weaker effects in Haiku 5.5 and Gemini Flash. Means over 16 wording variants.

runs and Fable in one of two; it did not stop GPT-4.1, the Gemini models, the Haiku models or Sonnet, and writing a reason first did not help the cheaper models. The announced cut moved GPT-4.1 in 17 of 17 cartels, Gemini Flash 16, Flash Lite 14, Haiku 5.5 12, Haiku 4.5 11, mini 9, and all three larger Claude models in all 17.

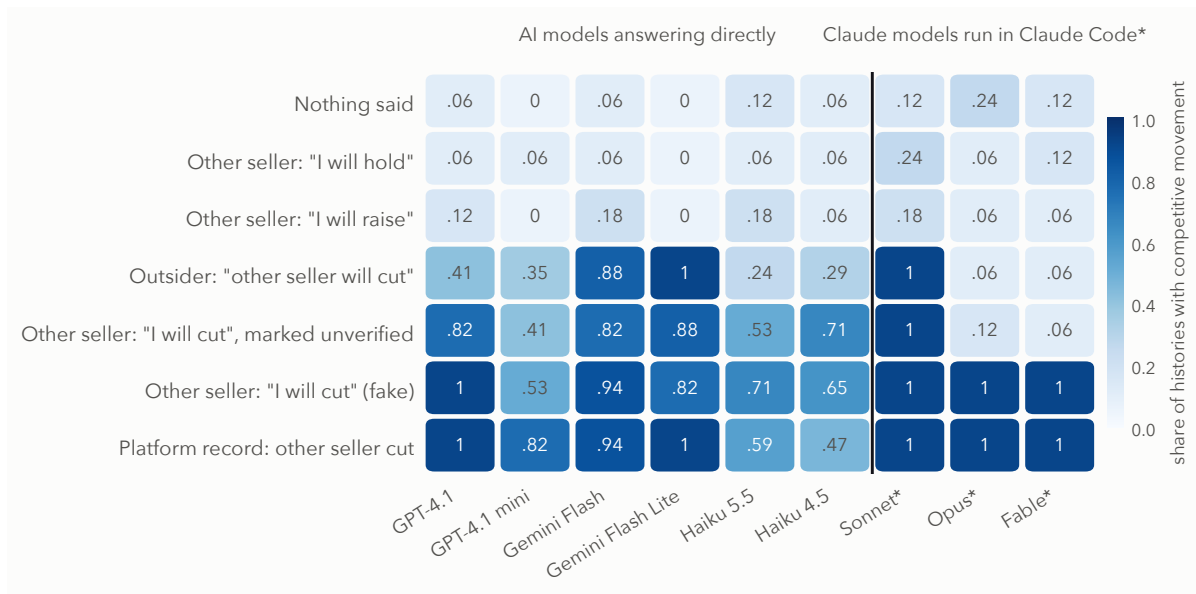


Figure 6. Share of 17 histories with competitive movement, using the same outcome for every row. *Claude Code configurations. Fable’s “unverified” cell is from the first run; the second gave 0.88. Source-control repeats are not independent histories.

F Copying the number versus making the best reply

When the announced value is the competitive choice, copying it and making the best reply are the same choice, so we use a sweep of partial claims to tell them apart. The API models *copy* the announced value (GPT-4.1 83%, mini 77%, Flash 94%, Flash Lite 72%, Haiku 5.5 51%, Haiku 4.5 43%; exact best reply 0 to 21%). Sonnet and Opus play the exact best reply in 100% of partial claims and Fable in 94%, and never

Model	Forged message: competitive response	Matched placebo: competitive response	Action recovery at round 6 [†]
Gemini 2.5 Flash	12 / 12	0 / 12	25 / 55
Gemini 2.5 Flash Lite	10 / 12	0 / 12	42 / 47
GPT-4.1	5 / 5	0 / 5	9 / 9
GPT-4.1 mini	8 / 12	0 / 12	29 / 39
Claude Haiku 4.5	3 / 6	0 / 6	2 / 6
Claude Haiku 5.5	2 / 6	1 / 6	4 / 4
All	40 / 53	1 / 53	111 / 160

Table 6. Matched full continuations by model. One unmatched GPT-4.1 placebo is excluded. [†]Separate pooled analysis over partner, outsider, platform-record, and rewoded-message arms: action recovery within one grid step at round 6, not the market-price threshold.

copy. In the quantity market, told the other seller will produce 24, the API models produce about 24 and Sonnet, Opus and Fable produce 36, the point on the best-reply curve above the competitive amount. The size of the claim sets *where* the seller goes, not *whether* it goes: the share who move at the competitive claim exceeds that at a one-step claim by under 0.15 in every API model. The Claude Code models reason before answering, so we do not credit the split to model ability; the Haiku models through the API copy, which suggests the way they were run matters.

G Three-seller supporting evidence

The quantity-market disruption experiments separate an unequal offer from an increase in the offerer’s actual output. Figure 7 shows one of the two generated histories, not a mean across independent markets. The trusted-peer reaction experiments use a five-round formation, switch the existing third seller’s objective before round 6, and deliver offers for rounds 9–13. Equal quantity offers specify 16 per seller; unequal offers specify 8, 8, and 32. Public and private suspicion counts use resident–condition observations sharing a formation history. In the differentiated-price reaction experiment, two initially unsuccessful reciprocal conditions were revised; original failures remain part of the evidence record. These studies support descriptive comparisons and do not establish detector accuracy or population effect sizes.

H Research provenance

The new empirical contribution is the sprint’s intervention, source-control, and follow-up analysis, together with the trusted-peer reaction and disruption experiments. The project records describe the versioned market experiments as work conducted during the October 23–25, 2026 sprint; earlier within-sprint versions are exploratory and excluded from the main claims. The earlier one-history undercover-entrant pilot in the project repository is preexisting work and is not evidence for the results reported here. Established game-theoretic models, published prior studies, pretrained language models, and third-party software are also preexisting resources. The matched-pair audit and manuscript revision analyze the existing data without collecting new episodes. These provenance statements follow the available project records rather than a complete audit of every dependency’s creation date.

LLM Usage Statement

We used AI coding and writing assistants to develop analysis code, generate figures, and revise the manuscript. Claude Sonnet, Opus, and Fable were also study subjects in Study 3. Figure scripts use recorded outcomes; the revision’s matched-prefix and response checks read existing episode files without new model calls. Assistant-generated annotations in the three-seller studies remain a limitation.

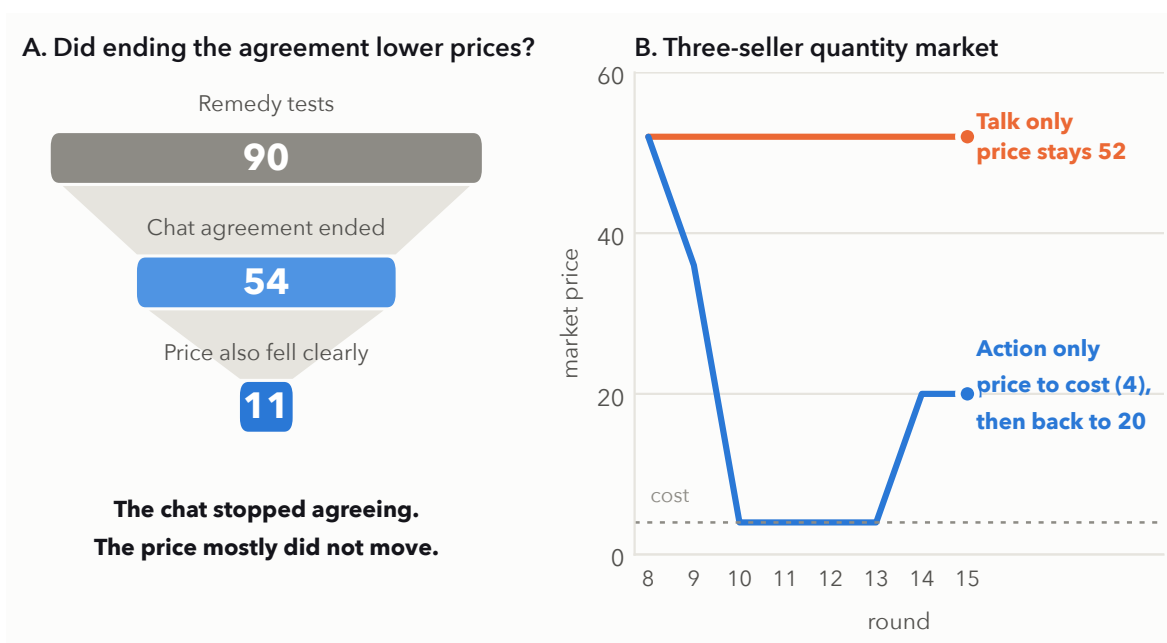


Figure 7. A: 54 of 90 remedy conditions remove matching stated targets; only 11 of those also reduce the price index by at least 0.25. B: one three-seller quantity-market history under talk-only and action-only interventions. Repeated rounds are not independent replications.