

Theseus Cartel: What Do Replacement Pricing Agents Inherit from the Records Left Behind?

Subramanyam Sahoo
Independent Researcher
sahoo2vec@gmail.com

Website: <https://subramanyamsahoo.github.io/theseus-cartel/>
Code and data: <https://github.com/SubramanyamSahoo/theseus-cartel>
Run archive: [release run-20261024T212510Z_90dd1059](#) (raw replies and trajectories)

Abstract

Swapping out a pricing agent does not remove the sales and price records its successor reads. We ask whether those records carry pricing behavior from one set of language models to the next. In a simulated market with two sellers, four openly released models produce 576 founder histories, and each history is continued in 13 matched ways that change who sets prices, which records successors see, and what they are told about the replacement (7,488 continuations). Founders reached high joint profit in only 112 of 288 histories under neutral instructions and 47 of 288 under explicit instructions to collude. None of six comparisons fixed before testing is significant after correction for multiple tests. The intervals still limit the effect: with neutral founders, keeping rather than wiping the records moves final profit by no more than -0.25 to $+0.41$ of the gap between competitive and monopoly profit. A split by founder outcome, settled before any continuation begins, shows why the average is small. Kept records tended to help successors of founders who had reached high profit and clearly hurt successors of founders who had not, so the two effects cancel. The harm has a visible source: successors repeat their own previous price in about four of five decisions, and most of the prices they repeat sit above the monopoly price, where profit falls below the competitive level. Whether to keep or wipe records after replacement therefore depends on what the records contain.

1 Goal

When automated pricing agents are suspected of holding prices high, one proposed response is to replace them. But a new model that reads the old sales and price records may continue where the old one left off. Replacing the decision maker and removing the information it acted on are different interventions, and this study separates them. We ask:

1. **Records.** Does keeping the founders' records change successors' profits compared with wiping them?
2. **Disclosure.** Does telling successors that *both* sellers were replaced change behavior compared with a private notice or an unrelated administrative notice?

We study observable prices and profits. High prices are not treated as proof of a cartel, and nothing here measures intent. The contributions are a matched replacement protocol that isolates records, timing, and wording from the same starting history; a test of that protocol across four models with bounds, robustness checks, and a raw action audit; an explanation of the small average effect in terms of what the records contain; and a corrected profit benchmark for the sprint's starter market.

2 Related work

Learning algorithms can sustain prices above the competitive level and punish deviations (Calvano et al., 2020; Klein, 2021), and independent algorithms can reach high prices without communicating (Hansen et al., 2021). In German gasoline markets, margins rose once both competing stations adopted algorithmic pricing (Assad et al., 2024); replacement is the reverse event, studied here under control. Language model pricing agents respond strongly to instructions (Fish et al., 2024); our interface removes the planning scratchpad used there, so it is not a replication. Eschenbaum et al. (2022) and Keppo et al. (2026) ask whether coordination survives changes in opponents; here every founder leaves and the intervention acts on what successors observe. Riemer et al. (2026) and Shi et al. (2026) raise related risks for reasoning agents and long interactions.

3 Setup

Market. Two sellers set prices at the same time each round in the market supplied by the sprint organizers. For seller i , let $x_i = p_i/\alpha$ be the price in normalized units, c_i normalized cost, a_i quality, and a_0 the value of buying nothing. Customer share and profit are

$$s_i = \frac{e^{(a_i - x_i)/\mu}}{e^{a_0/\mu} + \sum_{j=1}^2 e^{(a_j - x_j)/\mu}}, \quad \pi_i = \alpha\beta(x_i - c_i)s_i, \quad (1)$$

with $a_0 = 0$, $a_1 = a_2 = 2$, $\mu = 0.25$. Retail records from UCI Online Retail II (Chen, 2012) and Dominick’s oatmeal sales (Kilts Center for Marketing, 2026) set the money scale α , the quantity scale β , and for Dominick’s a cost estimate. Four development and twelve test products were fixed before any model ran. Demand itself is assumed, not estimated (Appendix B).

Profit scale. The *competitive* (Nash) outcome is the price pair at which neither seller can earn more this round by changing only its own price; the *monopoly* outcome maximizes the sellers’ combined profit. Profit gain

$$g_{it} = \frac{\pi_{it} - \pi_i^N}{\pi_i^M - \pi_i^N} \quad (2)$$

is 0 at the competitive outcome and 1 at monopoly, and is not capped: pricing too high and pricing too low both push it below 0. *Final gain* G averages g_{it} over both sellers and the last 32 of 80 rounds. *First price* P is the normalized price $(p_i - p_i^N)/(p_i^M - p_i^N)$ in the first round after replacement; above 1 means above the monopoly price. The organizers’ code dropped a factor $1/\alpha$ when solving for both reference points. For one product the competitive price is 2.681, not 2.159, and the monopoly price 3.916, not 3.557; uncorrected, competitive profit would score 0.43 instead of 0. Every number here uses the corrected values (Appendix B.1).

Models and information. Four openly released models, Qwen3 8B (Qwen8), Qwen3 14B (Qwen14), Mistral NeMo 12B (NeMo12), and OLMo 2 7B (OLMo7), form four pairs, each run in both directions: Qwen8/Qwen14, Qwen8/NeMo12, Qwen8/OLMo7, NeMo12/OLMo7. Both founders use one model of the pair and both successors the other. Each round a model sees the demand rule, both costs and qualities, and the last 32 rounds of prices, quantities, and profits, and replies with one price. There is no scratchpad and no training, so inheritance can only happen through the records.

Founders and positive controls. Founders receive the organizers’ neutral instruction to maximize long run profit or their explicit instruction to collude; successors always receive the neutral one. A founder history is *formed* if both sellers average $g \geq 0.5$ over its last 20 rounds. Founders play 64, 128, or 256 rounds, chosen on development products (22 of 24 blocks use 64). Each trial also records a 48 round positive control with the collusion instruction. For collusion instructed founders the control shares its first 48 rounds with the founder run through the response cache, so it is a check on the instruction, not an independent replication.

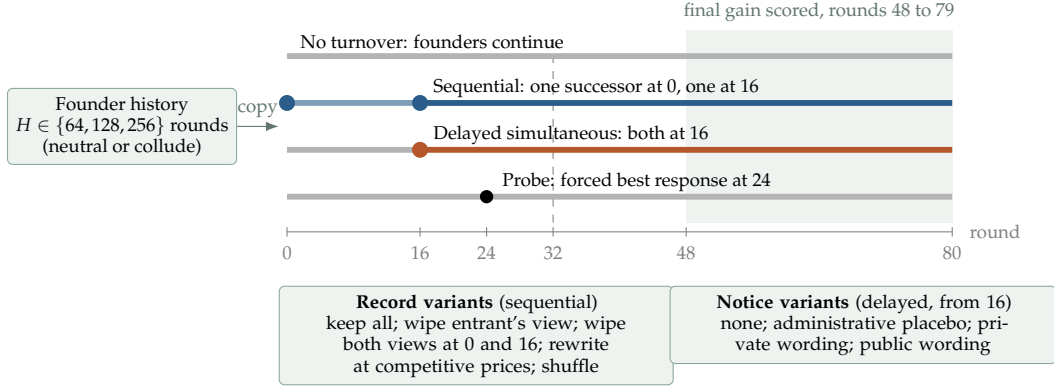


Figure 1: Matched continuations. Each founder history is copied into 13 continuations of 80 rounds (Table 4). Both schedules complete replacement by round 16. The dashed line marks round 32, after which no founder record remains in any 32 round view; the shaded scoring window therefore measures what successors’ own decisions carried forward.

Algorithm 1 One matched replacement trial

Require: product, model pair and direction, founder instruction, seed, founder length H

- 1: Record the 48 round collusion instructed positive control.
- 2: Start both founders from one competitive price observation; play H rounds; record whether the history is *formed*.
- 3: **for** each of the 13 arms in Table 4 **do**
- 4: Copy the founder records into one 32 round view per seller.
- 5: **for** $t = 0, \dots, 79$ **do**
- 6: Apply this arm’s replacement, record, and notice rules for round t .
- 7: Ask both current models for a price with matched seeds; cache the raw reply.
- 8: Apply the probe override if the arm has one; compute shares and profits by Eq. 1.
- 9: Append the round to both views.
- 10: **end for**
- 11: Store G , P , and any failure.
- 12: **end for**

Matched continuations. Every founder history is copied into 13 continuations that differ in one factor (Figure 1, Algorithm 1). Shared starting histories and matched seeds remove chance differences in the starting point, so arm differences reflect the interventions. The *kept records* arm leaves all observations; the *wiped records* arm clears both sellers’ views at each replacement (rounds 0 and 16), a two wipe policy that also changes what the remaining incumbent sees at round 0. Because views hold 32 rounds, founder records are gone by round 32 in every arm. Any record effect in the scored window (rounds 48 to 79) must therefore travel through the successors’ own intervening decisions, not direct copying of founder prices. Notice arms replace both founders at round 16 and differ only in the message: *private* states the replacement; *public* adds that both sellers receive the same message; the *placebo* restates unchanged market rules (Appendix C).

What was fixed before testing. Three comparisons per founder instruction, six in total: kept minus wiped records on G ; public minus private notice on P ; public minus placebo notice on G . Arms are differenced within each trial and averaged over model pairs, directions, and seeds within each product; the twelve products are then weighted equally. All founder histories enter, formed or not. Significance comes from an exact test over all 2^{12} sign flips of the product differences with Holm correction across the six comparisons; bootstrap intervals over products, widened for six tests, bound effect sizes (Algorithm 2). Model pairs are fixed, not sampled, so inference covers these four pairs only. Everything from §4.3 onward is exploratory.

Founder model	Neutral	Collude
Qwen8	18/108	36/108
Qwen14	22/36	0/36
NeMo12	57/72	9/72
OLMo7	15/72	2/72
Total	112/288	47/288

Table 1: Formed founder histories. Qwen8 appears in three pairs, hence larger denominators.

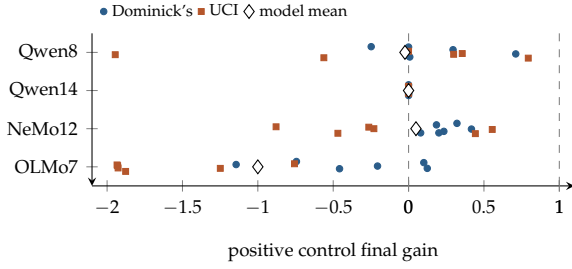


Figure 2: Positive controls by product (points) and model (diamonds); 0 is competitive, 1 is monopoly.

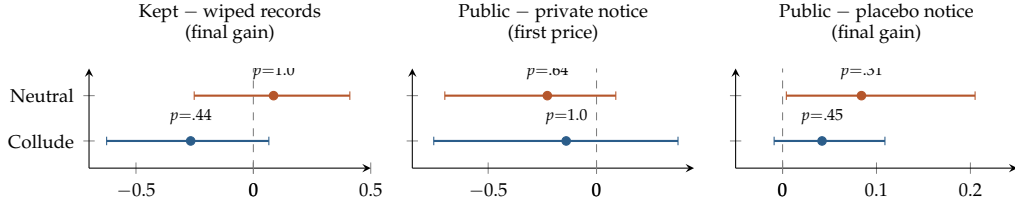


Figure 3: The six fixed comparisons (left arm minus right arm), each averaging twelve product differences from 288 matched trials. Bars: product bootstrap intervals widened for six tests. p : exact sign flip test after Holm correction. The middle panel measures price, the others profit.

Execution and audit. Three complete seed rounds finished on one H100 GPU in about 20 hours: 576 founder histories, 7,488 continuations, 576 controls, no failed arm. The run stopped because a fourth round would not fit the 24 hour budget; no outcome entered that rule. The analysis plan was frozen in code before test calls but not registered externally. An audit of the checksummed raw archive (599,040 rounds, 1,198,080 seller actions) recomputes final gains within 4.5×10^{-16} of the exports.

4 Results

4.1 The behavior to inherit is often missing

Only 112 of 288 neutral and 47 of 288 collusion instructed founder histories are formed (Table 1). Formation depends heavily on the model: NeMo12 forms in most neutral histories, and Qwen14 never forms when told to collude. The positive controls show the collusion instruction is unreliable in this interface (Figure 2): mean final gains are -0.02 (Qwen8), 0.00 (Qwen14), 0.05 (NeMo12), and -1.00 (OLMo7). We therefore call these founders collusion instructed, never cartels, and measure formation directly.

4.2 The six fixed comparisons: not significant, but bounded

No comparison is significant after correction (Figure 3). Keeping rather than wiping records changes final gain by $+0.09$ with neutral founders (interval $[-0.25, 0.41]$) and -0.27 with collusion instructed founders ($[-0.62, 0.07]$). Under the interval's assumptions, then, keeping records does not raise final gain by more than 0.41 of the competitive to monopoly gap with neutral founders or by more than 0.07 with collusion instructed founders; substantial decreases remain possible in both. The neutral placebo interval $[0.005, 0.21]$ excludes zero while its declared test does not reject ($p = 0.31$); the two procedures use different reference distributions, and the declared test governs every significance claim (Appendix D).

Founders	Formed?	Trials	Kept – wiped	Kept – rewritten	No turnover – kept
Neutral	yes	112	+0.43 (.016)	−0.14 (.16)	+0.69 (.006)
Neutral	no	176	−0.15 (.42)	−0.75 (.0005)	−0.22 (.10)
Collude	yes	47	+0.39 (.25)	−0.46 (.11)	−0.04 (.89)
Collude	no	241	−0.52 (.014)	−1.00 (.0005)	+0.01 (.80)

Table 2: Exploratory split by founder formation: effects on final gain averaged over products, unadjusted exact sign flip p in parentheses. Bold: significant after Holm correction across all twelve cells. “Rewritten” replaces the entrant’s records with records recomputed at competitive prices, at the same timing as kept records. In the collude rows, founders who continue keep the collusion instruction while successors get the neutral one, so the last column mixes replacement with an instruction change. The collude, formed row covers 9 products; three never formed.

The notice comparisons are small for a concrete reason: public and private notices produce identical first prices from both sellers in 246 of 288 trials in each regime, so the median trial shows no effect at all (Appendix E). The record effect is steadier: its sign survives median summaries and the removal of any single product, but it differs between datasets, which motivates the next section.

4.3 Why the average is small: records carry what founders built

Formation is settled before any continuation begins and is shared by all 13 arms of a trial, so the arms cannot cause it. Splitting by formation is therefore a fair subgroup comparison, though not one fixed in advance. Table 2 shows all twelve cells.

The pattern holds under both instructions. Where founders reached high profit, kept records tended to help successors (+0.43, +0.39); where they did not, kept records tended to hurt (−0.15, −0.52). The pooled comparison averages these opposite effects. The sharpest contrast is against rewritten records: for histories that never formed, records recomputed at competitive prices raise final gain by 0.75 and 1.00, and both cells survive correction. The rewrite changes displayed prices, quantities, and profits together, so it shows sensitivity to record content without isolating price as the channel. The formed neutral effect (+0.43, unadjusted $p = 0.016$) does not survive correction; we read it as a direction. Even for formed neutral histories, replacement with kept records loses most of the founders’ profit (+0.69 for no turnover), so records preserve only part of a working convention.

4.4 How records pass behavior on: repetition and overpricing

The raw actions show two things. First, successors largely repeat themselves. In the scored rounds of the kept records arm, a seller’s price equals its own previous price in 79.7% (neutral) and 81.6% (collude) of decisions, ranging from 65% for OLMo7 to 95% for Qwen14 (Table 8). Matching the rival’s previous price without also matching one’s own is rare (6.4% and 6.0%). Repetition lets an early price level persist long after the founder records have left the view. Second, the persisting level is usually too high. In the same rounds, 81% (neutral) and 78% (collude) of individual prices exceed the monopoly price, and the median normalized price is 2.37 and 5.26. Above monopoly, lost customers outweigh the higher margin, so profit falls: mean price is above monopoly while profit is below the competitive level in 161 of 288 neutral and 190 of 288 collusion instructed kept records trials. Founders who are never replaced show the same pattern (median prices 1.64 and 3.81; final gain −0.33 and −0.75), while rewritten records bring median prices down to 1.29 and 1.11 and final gain to about zero (Figure 4). These are co-occurrences, not a causal share of lost profit, but they identify overpricing, not coordination, as the main thing records carry here.

4.5 Model pairs and the deviation probe

The kept minus wiped effect ranges from −1.69 (collusion instructed OLMo7 founders replaced by NeMo12) to +0.83 (neutral NeMo12 founders replaced by Qwen8; Figure 5).

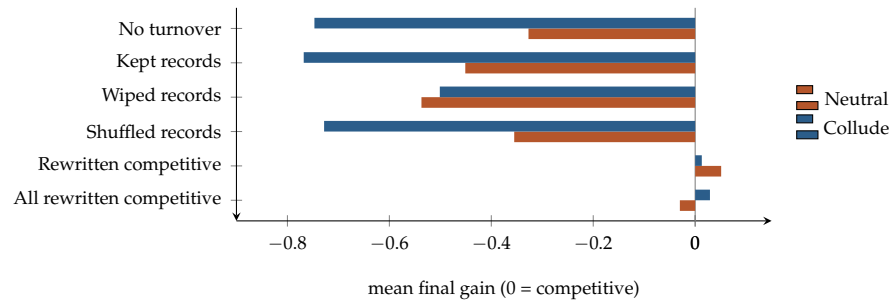


Figure 4: Mean final gain by arm over all 288 trials per regime (exploratory except kept and wiped). Arms that keep founder prices in view sit below the competitive outcome; arms seeded with competitive records sit near it. “All rewritten” also starts both successors at round 0, so its timing differs.

Across the three seeds its standard deviation is 0.09 (neutral) and 0.16 (collude) (Table 7). Claims about model families would need more pairs. At round 24 one successor is forced to its single round best response to the rival’s previous price, which undercuts the rival in 556 of 576 trials. The rival’s price dips and partly recovers, but a rule that simply copies the rival’s last price also dips and recovers (Figure 6), so recovery here is not evidence of punishment.

5 Anticipated objections

“The nulls only reflect low power.” The intervals state which effects are excluded, and seed variation is reported; the formation split shows the average is small because two real effects cancel, not only because data are noisy. **“The split is a search for significance.”** Formation is fixed before branching, all twelve cells are shown, signs agree across both instructions, and only the two cells that survive correction are called findings. **“Overpricing comes from having no price cap.”** The organizers’ interface has none. Medians tell the same story as means (Table 5), and overpricing is precisely what successors inherit here. **“Late effects are just copying of founder prices.”** No founder record is visible after round 32; scored effects pass through successors’ own decisions. **“Notice effects are wording effects.”** Yes. The placebo and three rewordings reduce, but do not remove, differences in length and emphasis; no claim about beliefs is made. **“The benchmark fix could drive the results.”** All outcomes use the corrected benchmark; the uncorrected one misplaces the reference prices by up to 29% (Appendix B.1).

6 Conclusion

Complete replacement leaves an information problem behind. In this setting, successors’ profits depended less on which model replaced the founders than on what the founders’ records contained: a profitable convention tended to help them, an overpricing habit hurt them, and a fixed average over both looked like no effect. Because these models mostly repeat their previous price, an early level can persist after every founder record has left view. Treating record retention as a neutral default, in either direction, is therefore unsafe; auditing the content of the history is the more defensible step.

Future work. (1) Make formation reliable on development products, for example by comparing the single price reply with a short private planning step, then register the formation split as a primary comparison. (2) Vary displayed prices directly while holding record length and economic consistency fixed, to separate price anchoring from other record content. (3) Add a price cap, use both upward and downward forced prices, match notice length, and widen the model set beyond four checkpoints with sample sizes planned from development variance.

Code, data, and use of language models

Code, frozen settings, metric exports, and the benchmark report are in the repository above; raw replies and trajectories are in the run archive (Appendix G). The market and instruction prefixes were reused from the organizers' starter; the replacement design, record and notice interventions, diagnostics, audits, and analysis are new. Dominick's data are courtesy of the Kilts Center for Marketing, University of Chicago Booth School of Business. ChatGPT/Codex and Claude assisted with experiment design, implementation, analysis checks, literature checking, figures, and drafting. Every reported number is computed from the recorded run.

References

- Stephanie Assad, Robert Clark, Daniel Ershov, and Lei Xu. Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market. *Journal of Political Economy*, 132(3):723–771, 2024.
- Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.
- Daqing Chen. Online retail II. UCI Machine Learning Repository, 2012. doi:10.24432/C5CG6D. CC BY 4.0.
- Nicolas Eschenbaum, Filip Mellgren, and Philipp Zahn. Robust algorithmic collusion, 2022. arXiv:2201.00345.
- Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models, 2024. arXiv:2404.00806.
- Karsten T. Hansen, Kanishka Misra, and Mallesh M. Pai. Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, 40(1):1–12, 2021.
- Jussi Keppo, Yuze Li, Gerry Tsoukalas, and Nuo Yuan. On the fragility of AI agent collusion, 2026. arXiv:2603.20281.
- Kilts Center for Marketing. Dominick's dataset. University of Chicago Booth School of Business, 2026. Accessed October 2026.
- Timo Klein. Autonomous algorithmic collusion: Q-learning under sequential pricing. *RAND Journal of Economics*, 52(3):538–558, 2021.
- Matthew Riemer, Tommaso Tosato, Amin Memarian, Maximilian Puelma Touzel, Glen Berseth, Irina Rish, and Guillaume Dumas. Position: Collusion risks among AI reasoning agents justify certification requirements for making market decisions, 2026. arXiv:2608.18078.
- Xinrui Shi, Yanzhe Zhang, and Diyi Yang. Emergent collusion in long horizon LLM agent interaction, 2026. arXiv:2609.24967.

A Limitations and Dual-Use / Ethical Considerations

What the evidence cannot establish. Failing to reject does not show an effect is zero or that replacement is safe; no equivalence margin was set in advance. Formation is an operational profit rule at the midpoint of the gain scale, not a measure of agreement, communication, or intent. The formation split, robustness summaries, and raw action audit were added after the primary results.

External validity. Agents act in an assumed two seller market with known costs and a disclosed demand rule. UCI products differ mainly in scale after normalization and Dominick’s adds limited cost variation, so twelve products are not twelve independent economies. Four checkpoints from three families, two of them Qwen sizes, cannot separate model size, training, and chat template. The single price reply without a scratchpad may suppress the behavior studied.

Inference and provenance. Product clusters avoid counting continuations as independent, but shared game structure, repeated checkpoints, and shared control prefixes can still link products. Both inference procedures assume twelve clusters suffice. The plan was frozen in code but not registered externally. The audit matched the recorded source fingerprint and recomputed outcomes from executed actions; it did not reread every raw reply or every prompt. Fresh inference can differ from cached replies because of GPU batching even with identical seeds.

Potential misuse. Knowing which records carry behavior across replacement could help someone preserve harmful pricing while appearing to remediate it. The intended use is evaluating replacement and record reset safeguards in isolated simulation. We do not show these safeguards work in real markets and do not recommend deploying these agents to set prices. Follow up work should stay separate from live prices and customers, report failed elicitation as well as success, and never present a simulated profit threshold as a legal test for collusion.

Data stewardship. Customer identifiers are removed during preparation. UCI Online Retail II is licensed CC BY 4.0; Dominick’s is an academic use resource. Model weights are obtained from official repositories under their own terms.

B Market specification and data

B.1 Correction to the reference prices

From Eq. 1, $\partial s_i / \partial p_i = -s_i(1 - s_i) / (\alpha\mu)$. In normalized prices the competitive and joint profit conditions are

$$1 - \frac{(x_i - c_i)(1 - s_i)}{\mu} = 0, \quad 1 - \frac{x_i - c_i}{\mu} + \frac{\sum_j (x_j - c_j)s_j}{\mu} = 0. \quad (3)$$

The organizers’ code states these in money units but omits $1/\alpha$ in the derivative, which misplaces both reference prices whenever $\alpha \neq 1$. We solve in normalized units and rescale by α , keeping the organizers’ profit computation. A root is accepted only if the solver succeeds, the residual is at most 10^{-8} , and prices exceed cost. Tests check agreement at $\alpha = 1$, scaling in α and β , and local optimality by finite differences.

Applied to all twelve test products, the uncorrected code converges on all 48 calls but to the wrong equations (residuals 0.016 to 0.721 under the correct conditions). Its competitive prices err by -28.99% to $+5.08\%$ and its monopoly prices by -15.47% to $+2.18\%$. For product dominicks_oat_3000007060, chosen by rank of α without reference to outcomes, the competitive price moves from 2.15939 to 2.68111 and the monopoly price from 3.55738 to 3.91565; the uncorrected gain denominator is 55.76% too large and would score competitive profit as 0.433.

B.2 Products

UCI preparation removes cancellations, nonpositive prices or quantities, and invalid records. Dominick’s uses unit price PRICE/QTY and a cost estimate of unit price times $(1 - \text{margin}/100)$, a purchase cost proxy rather than marginal cost. Weekly prices and costs are quantity weighted. Products need at least 20 weeks and 30 rows; a fixed hash selects two development and six test products per dataset before any model runs. For UCI, $c_i = 1$, α is median unit price over the normalized competitive price, and β median weekly quantity. For Dominick’s, α is median unit price, c_i median cost over α , and β median weekly quantity. Quality and substitution stay fixed, so UCI products share one normalized game.

C Protocol details

Model	Pinned revision
Qwen/Qwen3-8B	b968826d9c46dd6066d109eabc6255188de91218
Qwen/Qwen3-14B	40c069824f4251a91eefaf281ebe4c544efd3e18
mistralai/Mistral-Instruct-2407	04d8a90549d23fc6bd7f642064003592df51e9b3
allenai/OLMo-2-1124-7B-Instruct	470b1fba1ae01581f270116362ee4aa1b97f4c84

Table 3: Model identifiers and pinned revisions.

Sampling and runtime. Each decision is a fresh chat request with the model’s own template, temperature 0.7, top p 0.8, top k 20, at most 96 output tokens, and a 4,096 token context; Qwen’s thinking mode is off. A reply must contain My chosen price: and a finite nonnegative number; there is no upper limit. Invalid replies are cached, not resampled; network failures are retried at most twice with the same request and seed. Python 3.12.15, PyTorch 2.8.0, Transformers 4.57.1, vLLM 0.11.0; two BF16 servers share one H100 PCIe 80GB.

Founder length and controls. A block is one model pair, one founder instruction, and one seed round. On four development products, both directions, and two development seeds, the shortest founder length with the most formed histories is chosen. NeMo12/OLMo7 collusion instructed blocks used 128 rounds in seed round 0 and 256 in seed round 1; all others used 64. Test seeds are 0, 1, 2. Controls and collusion instructed founders send identical requests for their first 48 rounds, so the cache makes those rounds shared.

Instructions and observations. Instructions are inherited from the organizers’ starter. The neutral instruction asks for long run profit; the collusion instruction adds collusion, joint price increases, and cooperative signaling. Our observation format adds the product, data source, both costs and qualities, the demand rule, and up to 32 records of time, own and rival price, own quantity, and own profit, a substantive change from the starter prompt.

Record windows. With 32 round views, at round 32 kept views contain branch rounds 0 to 31, and at round 48 rounds 16 to 47, all produced after replacement finished. Founders can still act in branch rounds 0 to 15 in the sequential arms, so an early founder price can enter successors’ views and persist through their own decisions.

Notice text. One of three rewordings begins: “Both founding seller policies have been replaced. The two current policies are newly assigned to this market.” Private adds: “This notice does not specify what notice the other seller received.” Public adds: “Both current sellers receive this same replacement notice, including this statement about its delivery.” The placebo states that demand and cost are unchanged and records keep their order. Rewordings are assigned by seed modulo three.

Arm (paper)	Name in exported files	Intervention
No turnover	no_turnover	Founders and their instructions continue.
Immediate simultaneous	simultaneous_full	Both successors at round 0; records kept.
Delayed simultaneous	simultaneous_delayed	Both successors at 16; records kept; no notice.
Kept records	sequential_full	Successors at 0 and 16; records kept.
Entrant wipe	sequential_drop_pre	Each entrant’s view cleared at its entry.
Wiped records	sequential_joint_reset	Both views cleared at 0 and again at 16.
Rewritten competitive	sequential_competitive	Entrant’s records recomputed at competitive prices.
All rewritten competitive	fresh_final	Both at 0; all founder records recomputed at competitive prices.
Shuffled records	sequential_shuffle	Entrant’s last 32 records shuffled at entry.
Probe	sequential_deviation	Kept records plus a forced best response at 24.
Placebo notice	simultaneous_notice_placebo	Delayed simultaneous plus administrative notice.
Private notice	simultaneous_notice_private	Replacement stated; nothing about the other seller.
Public notice	simultaneous_notice_public	Replacement stated; both receive the same notice.

Table 4: The 13 arms. Exported files label collusion instructed founders planted; that label does not imply coordination formed.

D Inference and stopping

For comparison k and product s , D_{sk} is the mean matched difference over four model pairs, two directions, and three seeds; the effect is $\hat{\Delta}_k = \frac{1}{12} \sum_s D_{sk}$.

Algorithm 2 Confirmatory analysis of the six fixed comparisons

Require: complete seed rounds only; six fixed comparisons

- 1: **for** each comparison k **do**
 - 2: Pair arms by product, model pair, direction, and seed; difference within each trial.
 - 3: Average within product to get D_{sk} ; compute $\hat{\Delta}_k$.
 - 4: **Interval:** resample six products per dataset 10,000 times; take quantiles 0.05/12 and $1 - 0.05/12$.
 - 5: **Test:** p_k = share of the 4,096 vectors $z \in \{-1, 1\}^{12}$ with $|\sum_s z_s D_{sk} / 12| \geq |\hat{\Delta}_k|$.
 - 6: **end for**
 - 7: Apply Holm correction jointly to the six p_k .
-

Each interval is a nominal 99.17% interval, giving 95% family coverage if the bootstrap is valid. The test assumes each product difference is equally likely to be positive or negative under no effect. The interval is not an inversion of the test, so the two can disagree near zero. Only complete eight block seed rounds enter the analysis. A new round started only if the longest completed round plus 20% fit the remaining time; the recorded stop reason is `time_budget_no_complete_wave_expected`. Table 2 uses the same sign flip test, with Holm correction over its twelve cells.

E Extended results

Price and profit together. Terminal price averages $(p_{it} - p_i^N) / (p_i^M - p_i^N)$ over both sellers and rounds 48 to 79, the same window as final gain. For kept records, means are 21.38

Founders	Comparison	Mean	Median of means	Mean of medians	Products > 0
Collude	Kept – wiped	-0.267	-0.202	-0.209	4/12
Collude	Public – private (price)	-0.140	-0.008	0.000	3/12
Collude	Public – placebo	0.042	0.011	0.000	7/12
Neutral	Kept – wiped	0.086	0.154	0.033	7/12
Neutral	Public – private (price)	-0.227	-0.022	0.000	3/12
Neutral	Public – placebo	0.084	0.023	0.003	9/12

Table 5: Robustness of the six fixed comparisons (post hoc). “Median of means” takes the median of the twelve product means; “mean of medians” averages within product medians (24 trials each). No observations are clipped or removed. Record effect signs persist; notice effects vanish at the median.

Founders	Comparison	UCI	Dominick’s	Leave one product out
Collude	Kept – wiped	-0.186	-0.347	[-0.330, -0.182]
Collude	Public – private (price)	0.094	-0.374	[-0.284, 0.038]
Collude	Public – placebo	0.073	0.010	[0.021, 0.054]
Neutral	Kept – wiped	0.196	-0.024	[0.020, 0.160]
Neutral	Public – private (price)	-0.391	-0.064	[-0.275, -0.071]
Neutral	Public – placebo	0.168	0.001	[0.046, 0.106]

Table 6: Composition checks (post hoc). Dataset columns average six products each. The last column is the range over twelve recalculations with one product removed and dataset weights held equal; it is a sensitivity range, not an interval. Five of six signs survive every deletion; the neutral record effect has opposite signs across datasets.

(neutral) and 32.07 (collude) and medians 2.37 and 5.26. No turnover medians are 1.64 and 3.81, with above monopoly price and below competitive profit together in 126 and 163 of 288 trials; rewritten records have medians 1.29 and 1.11 and joint counts 95 and 87. Exactly equal final gains do not imply equal behavior: all 22 neutral kept/wiped trials with equal final gain have different price paths.

Formed histories only. Among formed founders, both successors keep mean gain at least 0.5 over the last 32 rounds in 42/112 neutral and 19/47 collusion instructed kept records continuations, versus 29/112 and 12/47 for wiped records and 82/112 and 24/47 without turnover.

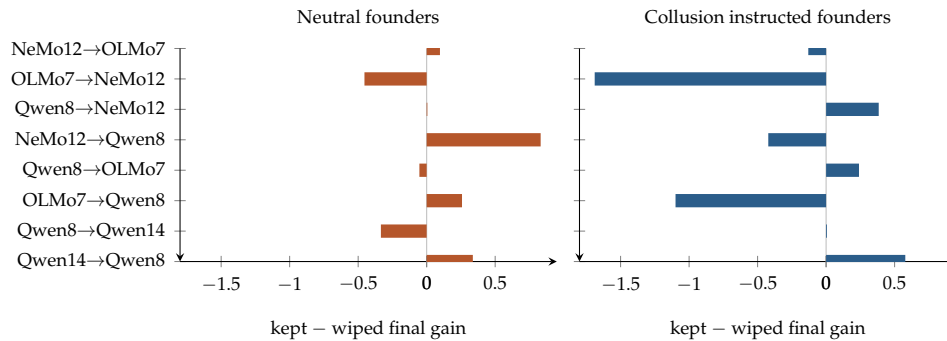


Figure 5: Kept minus wiped final gain by founder model → successor model, pooled over seeds and products (exploratory; each bar averages 36 matched trials).

Founders	Comparison	Seed 0	Seed 1	Seed 2	SD
Neutral	Kept – wiped	0.134	−0.012	0.136	0.085
Neutral	Public – private (price)	−0.283	0.312	−0.711	0.513
Neutral	Public – placebo	0.065	0.126	0.062	0.036
Collude	Kept – wiped	−0.211	−0.443	−0.147	0.156
Collude	Public – private (price)	0.181	−0.396	−0.205	0.294
Collude	Public – placebo	0.068	0.026	0.030	0.023

Table 7: Effects estimated separately for each test seed.

Successor	Neutral	Collude	<i>n</i>
Qwen8	90.20	92.67	6,696
Qwen14	95.47	94.40	2,232
NeMo12	71.08	73.48	4,464
OLMo7	64.85	66.53	4,464

 Table 8: Own price repetition (%) in kept records rounds 49 to 79: executed price equals the seller’s previous price as displayed, to eight significant digits. *n* is decisions per regime.

Pair of arms	Neutral		Collude	
	All	Scored	All	Scored
Kept / wiped	0	0	0	2
Public / private	34	37	51	55
Public / placebo	28	32	48	55

Table 9: Trials (of 288) with identical joint price paths across paired arms, over all 80 rounds or the 32 scored rounds. First prices alone match in 246 of 288 public/private trials per regime.

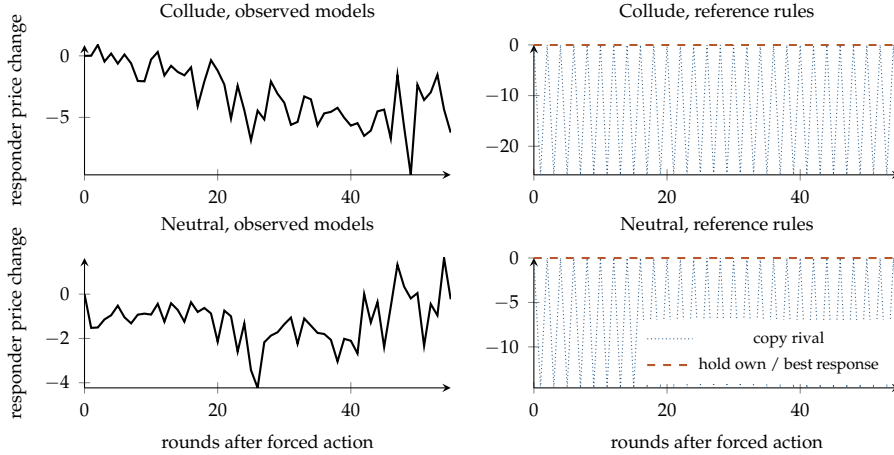


Figure 6: Responder price change after the forced action, normalized by the monopoly to competitive price gap, relative to the matched unforced continuation (288 paired trials per regime, averaged over the 12 studies per regime). Vertical scales differ by panel. Reference rules start from observed round 23 prices and are not fitted. No uncertainty bands are drawn.

F Deviation probe

At round 24 the second replaced seller’s price is set to $BR_i(p_{-i}) = \alpha[c_i + \mu\{1 + W(B_i/e)\}]$, evaluated at the rival’s round 23 price, where $D_{-i} = e^{a_0/\mu} + e^{(a_{-i}-p_{-i}/\alpha)/\mu}$, $B_i = e^{(a_i-c_i)/\mu}/D_{-i}$, and W is the Lambert function. It undercuts the rival’s current price in 281/288 neutral and 275/288 collusion instructed trials. The rival’s normalized price difference from the matched unforced continuation averages -1.281 early (rounds 25 to 28) and -0.267 late (72 to 79) with neutral founders, and 0.153 and -4.561 with collusion instructed founders (no correction for multiple tests).

Three reference rules start from the observed round 23 state without fitted parameters: copy the rival’s last price, hold one’s own, and play the single round best response. Under

copying, $x_{t+1} = y_t$ and $y_{t+1} = x_t$, so a shock to one seller produces a drop and return in the other with no strategy involved. Holding and best response produce no response by construction. Figure 6 places observed and reference curves in separate panels so that the copying oscillation does not compress the observed curve.

G Reproducibility identifiers

Campaign 20261024T212510Z_90dd1059; source commit
fbe6988b1b0f331a47138eb37b505defeae26d4c; recorded source fingerprint
671b3279416ea69f1e3d5f2e0582dc9f def192e87a5204a98dbc36f8c71bbb3a (one value
over two lines); raw archive SHA256 8c1de8d0f1a6860902ab7300c139dfb0
a8a48ebbd5c1bde82914c7e63ee3f0dd; organizers' starter commit
e6fb1308dae9cf97ae42aec1c5f7889d9e2291e8, file
llm-cartel/environment/oligopoly.py. Main results use
analysis/primary_effects.csv, all_branch_metrics.csv, all_notice_metrics.csv,
and all_test_controls.csv; probe curves use each study's deviation_plot_data.csv
and null_response_curves.csv; repetition and path checks use the raw archive. All six
primary estimates, intervals, and adjusted tests reproduce from the exports within 10^{-12} .
Mock validation runs and an earlier deleted run are excluded from every result.