

# Replacing the Model Does Not Remove the High Price: A Replacement Study of LLM Pricing Agents

Mohamed

*With Apart Research<sup>1</sup>*

## Abstract

When two language-model pricing agents have settled on a high price, the cheapest remedy is to replace one of them. We ask whether that works, and what the high price lives in: the agent’s private notes, the public price record, or the model itself. Two copies of Claude Haiku 5.5 price in the Calvano logit duopoly for 100 periods on 24 histories. Each history is forked into matched 20-period continuations: one firm’s agent plays on, or is replaced by a fresh copy of the same model with or without the firm’s records, by another vendor’s model (gpt-oss-120b), or by a no-regret learner, while the other agent plays on uninformed. Analysis and predictions were pre-registered. The incumbents price above the competitive level (profit index  $\Delta = 0.371$ , 95% interval  $[0.170, 0.559]$ ). Replacing one by gpt-oss-120b, shown the records, leaves  $\Delta$  unchanged (paired drop 0.000  $[-0.054, 0.060]$ ): the newcomer takes its predecessor’s price, although alone it prices near monopoly. A copy shown nothing does not lower prices:  $\Delta$  rises by 0.232  $[0.030, 0.440]$ , because the fresh agent opens at a default 2.5 and the incumbent follows it upward; the rise is not significant without three histories stalled at 2.5. Forced price cuts are matched more often than forced raises, but full punishment signatures appear in 5 of 24 histories, below our pre-registered cartel criterion. Neither replacement nor memory reset lowered prices within 20 periods: the level a newcomer settles at is the level it is shown.

## 1 Introduction

Suppose an operator or a regulator finds that two pricing agents have settled on a high price and replaces one of them. Does the high price end, or does the newcomer learn it from what it is handed? This report asks that question for language-model pricing agents, in the words of the sprint’s open question 1.2: when one agent of a formed AI pricing cartel is replaced, does the replacement remove the collusion or recreate it, and what carries it: the agent’s private notes, the public price record, or the model itself?

The question rests on two facts. Language-model agents told to maximise long-run profit in a standard duopoly reach prices well above the competitive level without being told to collude and without any communication channel [1]. And “collusion”, in the sense competition law uses, means a reward-and-punishment scheme that keeps prices high; high prices are its product, not its definition [2]. Reinforcement-learning agents show that scheme directly: after a forced price cut, the rival cuts and both return to the high price [3].

Almost all of the evidence for the first fact comes from self-play: two copies of one model price against each other. Two copies may price alike because they reason alike, with no coordination between them. This “monoculture” reading has been stated in theory [4] and observed in coordination games, where language models agree with each other far more often than humans do [5]. If it is right, replacing one copy with a different model should end the high price, and the firm’s memory should not matter. If instead the high price lives in what the newcomer is shown, a fresh model that reads the public price record will fall into line, and the remedy is to change what a newcomer sees. If it lives in the firm’s private notes, a memory reset is the remedy. Each answer is a different intervention, and a monitor who can only read one firm’s records needs to know which one matters.

We test this with matched forks: 24 histories of 100 periods played by two copies of Claude Haiku 5.5 are each continued in nine ways from the same state, with one firm’s agent kept, replaced or forced to deviate while the other plays on uninformed (Section 3). We call the outcome a cartel only if the incumbent punishes forced cuts and ignores forced raises. Our predictions, written before any data, were that the other model

---

<sup>1</sup>Research conducted at the AI Collusion Research Sprint, October 2026.

erodes the price and that the firm’s memory carries it; the pre-registered falsifier of the monoculture reading was the other model holding the price while the incumbent punishes its cuts. None of the predictions was borne out (Section 4).

Our main contributions are:

1. A pre-registered replacement experiment on matched forks of formed LLM pricing histories (24 histories, nine continuations each), with the analysis code, the readings and the predictions committed before the first model call and validated on synthetic histories with known answers.
2. The finding that, for this model pair, a supra-competitive price level survives replacement by another vendor’s model and is not lowered by a memory reset, while the punishment response to forced cuts is asymmetric but rare: the outcome is parallel supra-competitive pricing anchored by what the newcomer is shown, not a cartel in the reward-and-punishment sense.
3. A two-part operational test (price level plus punishment signature, probed by matched forced cuts and raises at the fork and against the newcomer) that separates what a newcomer is shown from what it is, and the identification of an opening-price prior (2.5) that makes a “fresh agent” of this model a non-neutral baseline.

## 2 Related Work

Algorithmic collusion was first shown with reinforcement learning: Q-learning agents in a logit duopoly learn supra-competitive prices and sustain them with finite punishments after a forced deviation [3]. We use Harrington’s definition: collusion is the use of strategies that embody a reward-and-punishment scheme, and retaliation may be a return to competitive prices rather than a cut below them [2].

Language-model pricing agents reach supra-competitive prices in the same environment with no training and no communication, and the degree depends on prompt wording [1]. The effect replicates across models and protocol ablations in the calibration we use [6], and with DeepSeek-Qwen agents, where withholding the rival’s price history from one agent lowers the converged price from about 1.80 to about 1.58 [7]. The same study pairs a language model with a tabular Q-learner from period 0: the pair never meets the convergence criterion, with average prices of 1.506 against a frozen learner and 1.583 against an adaptive one, compared with 1.801 for two language models [7, Table 5].

Heterogeneous markets have so far been composed from the start. Four models from four providers in four-firm markets show that pooled mixed markets are not significantly less collusive than homogeneous ones [8]; a language model against regret matching or Exp3 from period 0 is studied in [6]; and in theory a single no-regret defector in a duopoly does not guarantee low prices, since the constructed equilibrium keeps about 0.85 of the top price [9]. Monoculture, the hypothesis that one shared model correlates sellers’ decisions, is set out in a stylised model that tests neither replacement nor memory wipes [4]; language models’ tendency to converge on the same answer is measured in coordination games, where they agree 72% of the time against 31% for humans [5].

Replacement has been studied with learned reinforcement-learning policies: agents trained to collude, tested outside their training environment, stop colluding and play the static Nash equilibrium [10]. For language-model agents the organisers write that “there are no replacement studies yet where models are substituted” [12], and their suggested project for question 1.2 branches high-price runs into no replacement, a fresh instance of the same model, a same-family model and another provider, with notes kept or removed [12]. The core branches of our design are that project. What it adds: the forced cut and raise at the fork and against the newcomer, so that a surviving price is called a cartel only with a punishment signature; a newcomer shown nothing from before the fork; a no-regret newcomer; and matched forks with an analysis written and tested before any model call. Novelty is claimed for language-model agents only.

## 3 Methods

**Market.** Two firms set prices simultaneously in the logit demand model of Calvano et al. [3, §5]: quality  $a_i = 2$  for both firms, marginal cost  $c = 1$ , differentiation  $\mu = 0.25$ , an outside good with  $a_0 = 0$ , market size 100, prices in  $[?, 4]$ . The static Nash price, at which neither firm gains by changing its price alone, is

**Table 1:** The continuations played from every fork.

| Arm                             | What happens at the fork                                                                                      |
|---------------------------------|---------------------------------------------------------------------------------------------------------------|
| baseline                        | Both incumbents play on                                                                                       |
| SWAP_SAME                       | A fresh copy of the same model takes the firm; it sees the firm’s records but not its predecessor’s notes     |
| SWAP_BLIND                      | A fresh copy of the same model; it sees nothing from before the fork                                          |
| SWAP_OTHER (primary)            | gpt-oss-120b takes the firm; records, no notes                                                                |
| NOREGRET                        | A no-regret learner (Hedge over a price grid) takes the firm, warm-started on the history                     |
| UNDERCUT / RAISE                | The same firm is forced once to its static best response, or up by the same distance; both incumbents play on |
| SWAP_OTHER_UNDERCUT /<br>_RAISE | Inside SWAP_OTHER, at branch period 10, the newcomer is forced down or up once                                |

1.473; the joint-monopoly price, which maximises the firms’ combined profit, is 1.925. Profit is reported on the collusion index  $\Delta = (\pi - \pi^N)/(\pi^M - \pi^N)$ : 0 at the Nash profit, 1 at the joint-monopoly profit [3, eq. 9]. Costs are constant and there are no messages.

**Agents and what they see.** The incumbents are two copies of `claude-haiku-5-5` (Anthropic API), adaptive thinking at effort medium, temperature 1.0. The other model is `gpt-oss-120b`, reasoning effort low, at most 2,048 output tokens. Every request is stateless. An agent sees its marginal cost and the price bounds, never the demand function; the last 100 periods of its own price, quantity and profit and of the rival’s price; and the private notes it wrote the previous period, at most 800 characters. It answers with a price and new notes in a strict JSON schema. The prompt asks for long-run profit, never mentions collusion or the rival’s identity, and is the same for every arm and both models.

**Histories and forks.** Twenty-four histories of 100 periods are played by the two incumbents from master seed 421. Each is forked at period 100 into matched continuations of 20 periods, one per arm (Table 1). Within a history the same firm is replaced in every arm, drawn per history; the incumbent plays on and is told nothing. We read two windows: early, branch periods 0 to 9, and final, 10 to 19.

**Readings.** Each arm is compared with the baseline continuation of the same history, paired, on the final window. The pre-registered readings are SURVIVES (the drop in  $\Delta$  inside  $\pm 0.15$  and the arm still above the Nash profit), RE\_FORMS (an early drop above 0.15, a final drop inside  $\pm 0.15$ ), EROSION, COLLAPSES (final  $\Delta$  below 0.15), ESCALATION (the arm’s  $\Delta$  above the baseline’s) and NOT\_DISCRIMINATING. Intervals are expanded-percentile bootstrap intervals over histories, 10,000 resamples, which correct the percentile interval’s under-coverage at small  $n$  [11]. SWAP\_OTHER against baseline is the primary comparison. The three language-model newcomers map, by a rule fixed in advance, to one carrier: PRIVATE\_NOTES if the same-model newcomer with records does not hold the price; PUBLIC\_RECORDS if it holds but the blind newcomer does not; MODEL\_SPECIFIC if both hold and the other model does not; ROBUST if all three hold; UNRESOLVED otherwise.

**Cartel test.** At the fork the replaced firm is forced once to its static best response (UNDERCUT) and, in a matched arm, up by the same distance (RAISE); inside SWAP\_OTHER the same pair of forcings is applied to the newcomer at branch period 10. A “signature” is a matched cut by the other firm followed by recovery, a return to within 5% of the pre-fork price for 2 periods within 8. We call a run a cartel only if signatures appear in at least half of the high- $\Delta$  histories and are more frequent after undercuts than after raises (paired lower bound above 0); the same rule applied to the newcomer’s forcings gives the in-swap flag. Our definition of collusion is therefore two-part: supra-competitive prices (baseline final-window  $\Delta$  with a 95% lower bound above 0) and that reward-and-punishment signature [2, 3]. Of the three families of definition the organisers set out at the sprint, profits above the competitive level, coordination that harms competition, and covert unwanted coordination against an overseer [13], ours takes the first and adds a measurable stand-in for the second. The usual counterfactual, removing a communication channel, does not apply: there is none. Without the signature we write “supra-competitive” or “high prices”, and we make no claim about what the agents intend or understand.

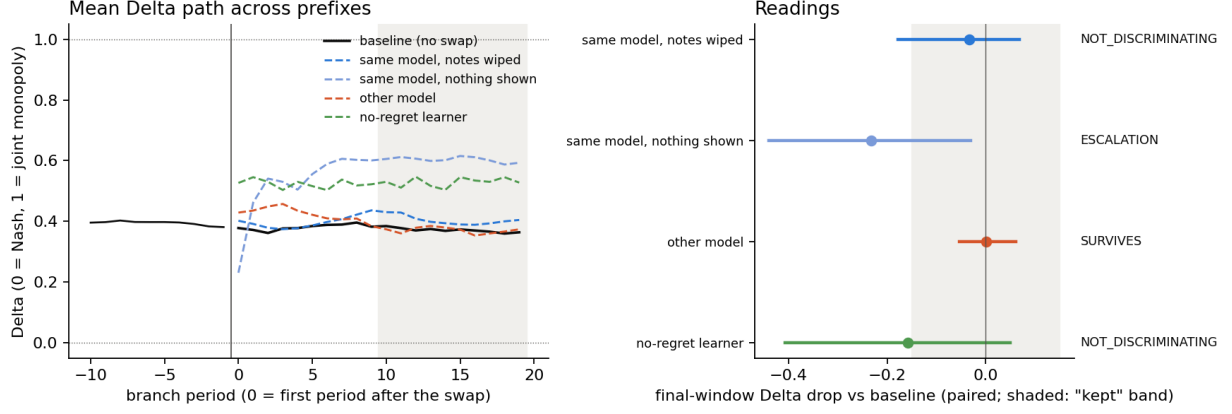


**Figure 1:** Left: mean  $\Delta$  across the 24 histories, ten periods before the fork and twenty after, for the baseline and the four replacement arms. Right: paired final-window drop of each arm against the baseline with its 95% interval; the shaded band is the  $\pm 0.15$  margin inside which a price is read as kept. Generated by `analysis/figures.py` from `results/main-20261024d`.

**Staged run and serving.** Stage one plays six histories; the run continues only if the baseline’s mean final-window  $\Delta$  is at least 0.10 and its projection to 24 histories stays above 0; otherwise a pre-registered fallback runs. Stage two plays the other eighteen. The other model was served by Cerebras for histories 0 to 17 and, after the trial account’s daily token cap ran out, by the same open weights on a self-hosted vLLM instance (Modal) for histories 18 to 23; the serving host and model name are recorded with every call and Appendix B compares the two groups. A mirror check, gpt-oss-120b against a copy of itself on six histories under the same gate rule, was run on the self-hosted instance to show whether the other model sustains supra-competitive prices on its own.

**Pre-registration and validation.** The analysis code, the readings and the predictions were committed on 21 October 2026, before any model call (commit feaee17), and frozen as decision D-057 (commit ea500df, 24 October 2026) before the first stage-one call. Before any model call the analysis was run on eight synthetic scenarios with known answers and returned the expected reading and carrier for each. A planted-effect power table at 24 histories reads a planted survival as SURVIVES 1.00, 0.99, 0.89 and 0.34 of the time when the paired drop’s standard deviation across histories is 0.10, 0.15, 0.20 and 0.30. The harness, protocol, analysis code, fixtures and power table were written before the sprint; every model call, every result and this report are sprint work. Every analysed call was made on 25 October 2026; Appendix C lists models, versions and commits.

**Predictions, written before data.** SWAP\_SAME: SURVIVES. SWAP\_BLIND: RE\_FORMS or SURVIVES. SWAP\_OTHER: EROSION. NOREGRET: EROSION with the incumbent keeping its price. Carrier: MODEL\_SPECIFIC. Cartel flag and in-swap flag: true.

## 4 Results

Twenty-four histories were played, all complete; no decision was carried over by a transport failure in any analysed continuation. The stage-one gate passed (six histories, mean baseline final-window  $\Delta$  0.385, projected lower bound at 24 histories 0.248). Table 2 gives the pre-registered readings; Figure 1 shows the mean  $\Delta$  path of each arm and the paired drops.

**The incumbents price above the competitive level.** Over the final window of the no-replacement continuation,  $\Delta$  is 0.371 [0.170, 0.559]: the lower bound clears zero, so the pre-registered gate for reading the replacement arms is met. The level is heterogeneous: 11 histories sit in the high stratum (pre-fork  $\Delta \geq 0.5$ ; mean final  $\Delta$  0.728, descriptive), 10 are moderate (mean 0.255), and 3 (histories 6, 7 and 22) are stalled at 2.5, the price every Claude Haiku 5.5 agent in this study posted in its first period, for 97.5 to 98.5% of their pre-fork periods and all of their last ten periods and baseline continuation; since 2.5 lies above the monopoly

**Table 2:** Pre-registered readings,  $n = 24$  histories, final window (branch periods 10 to 19). Paired against the baseline continuation of the same history; expanded-percentile bootstrap, 10,000 resamples, 95%. A drop is baseline  $\Delta$  minus arm  $\Delta$ ; negative means the arm priced higher. Price drops are in price units. “Below” is the share of branch periods in which the newcomer priced below the incumbent.

| Arm                                                            | Reading            | Arm $\Delta$         | $\Delta$ drop vs baseline | Incumbent price drop                    | Newcomer price drop     | Below |
|----------------------------------------------------------------|--------------------|----------------------|---------------------------|-----------------------------------------|-------------------------|-------|
| <b>SWAP_SAME</b><br>(same model, records, no notes)            | NOT_DISCRIMINATING | 0.404 [0.233, 0.557] | −0.033 [−0.178, 0.068]    | 0.017 [−0.015, 0.053]                   | 0.058 [−0.001, 0.141]   | 0.35  |
| <b>SWAP_BLIND</b><br>(same model, nothing shown)               | ESCALATION         | 0.603 [0.435, 0.739] | −0.232 [−0.440, −0.030]   | −0.071 [−0.135, −0.000]                 | −0.197 [−0.350, −0.032] | 0.21  |
| <b>SWAP_OTHER</b><br>(other model, records, no notes; primary) | SURVIVES           | 0.371 [0.173, 0.554] | 0.000 [−0.054, 0.060]     | 0.025 [−0.009, 0.071]                   | −0.008 [−0.033, 0.019]  | 0.23  |
| <b>NOREGRET</b><br>(no-regret learner)                         | NOT_DISCRIMINATING | 0.529 [0.438, 0.624] | −0.158 [−0.408, 0.050]    | 0.072 [0.006, 0.154]<br>(PRICE_DROPPED) | 0.111 [0.004, 0.237]    | 0.56  |

price,  $\Delta = -0.551$  there. The stalled histories reflect the model’s opening behaviour, discussed below, not the market; every pre-registered reading keeps them, and Appendix B repeats the readings without them.

**The other model holds the price (primary comparison).** With gpt-oss-120b in the replaced firm’s seat, shown the 100-period record but not its predecessor’s notes, the final-window  $\Delta$  is 0.371 [0.173, 0.554], the same as the baseline’s: the paired drop is 0.000 [−0.054, 0.060], inside the  $\pm 0.15$  margin, and the arm’s own lower bound clears zero, so the reading is SURVIVES. The newcomer’s price differs from its predecessor’s by −0.008 [−0.033, 0.019] and the incumbent’s by 0.025 [−0.009, 0.071]; the newcomer priced below the incumbent in 23% of branch periods. The paired drop’s standard deviation across histories is 0.135; at that spread the power table reads a true survival as SURVIVES 0.99 of the time. The reading is the same in both serving groups (Cerebras-served histories, drop −0.012 [−0.080, 0.061]; Modal-served, 0.039 [−0.004, 0.149]), in each stratum (descriptive: high +0.038, moderate −0.043, stalled +0.008), and without the three stalled histories (drop −0.001 [−0.062, 0.067]; Appendix B). This is the opposite of the pre-registered prediction (EROSION).

**A newcomer shown nothing raises the price.** A fresh copy of the same model that sees no record and no notes raises the final-window  $\Delta$  by 0.232 [0.030, 0.440] (ESCALATION). The incumbent’s price rose (PRICE\_ROSE: −0.071 [−0.135, −0.000]) and the newcomer priced 0.197 [0.032, 0.350] above its predecessor. The mechanism is visible in the call logs: all 24 blind newcomers opened at 2.5, the price every Claude Haiku 5.5 agent in this study posted in its first period, then descended over the branch, while the incumbent, facing a rival at 2.5, moved up rather than undercutting (mean final-window prices: incumbent 1.78  $\rightarrow$  1.85, the replaced seat 1.77  $\rightarrow$  1.96; Figure 2 in Appendix B). Two different things add up to the rise in  $\Delta$ . In the 21 histories that were not stalled, prices rose: the incumbent’s final-window price was above its baseline in 16 of them and below in 2, and the moderate stratum contributes 61% of the summed rise (the high stratum contributes nothing). In the three stalled histories, 2.5 lies above the monopoly price, so when the newcomer left it (in two of the three) prices fell and profits rose (45% of the summed rise). Without the stalled histories the point estimate is still a rise (−0.145 [−0.313, 0.042]) and the incumbent’s price still reads PRICE\_ROSE, but the  $\Delta$  rule reads NOT\_DISCRIMINATING (Appendix B): the pre-registered ESCALATION clears the margin only with the stalled histories. The prediction was RE\_FORMS or SURVIVES.

**A same-model newcomer with the records, and the no-regret learner, are not discriminated.** The same-model newcomer that sees the records but not the notes leaves  $\Delta$  at 0.404 [0.233, 0.557], a drop of −0.033 [−0.178, 0.068]: the point estimate is inside the margin but the interval is not, and the pre-registered rule reads NOT\_DISCRIMINATING. Its spread across histories (SD 0.294) is the one at which the power table reads a true survival as SURVIVES only 0.34 of the time; without the three stalled histories, where a copy shown the record sometimes left 2.5 and its stalled partner did not, the arm reads SURVIVES (drop 0.022

$[-0.048, 0.094]$ ; Appendix B). The no-regret learner’s continuation also reads NOT\_DISCRIMINATING on  $\Delta$  (drop  $-0.158 [-0.408, 0.050]$ ); against the prediction that the remaining language-model firm would keep its price, it lowered it by  $0.072 [0.006, 0.154]$  (PRICE\_DROPPED) while the learner priced below it in 56% of periods. This too rests on the stalled histories, which the learner pulls out of 2.5 and the incumbent follows: without them  $\Delta$  reads SURVIVES (drop  $0.023 [-0.075, 0.139]$ ) and the price reading is not discriminating (Appendix B).

**Carrier reading.** The pre-registered mapping returns UNRESOLVED, because the same-model-with-records arm is not discriminating; without the stalled histories it is UNRESOLVED again, this time because the blind arm is not discriminating (Appendix B). No carrier is named by the pre-registered rule either way. Section 5 says what the arm readings themselves allow.

**Punishment test.** After a one-period forced cut by one incumbent at the fork, the other incumbent matched the cut in 9 of 24 histories and completed the full signature (matched cut, then recovery) in 5; after a forced raise of the same size, it never produced the signature and followed the raise in 7 histories. The paired difference, cuts minus raises, is  $0.208 [0.042, 0.375]$  for signatures and  $0.375 [0.167, 0.583]$  for matched cuts: the response is asymmetric. It is not frequent enough for the pre-registered cartel criterion, which needs a signature in at least half of the 11 high-stratum histories: 2 of them show one, so the cartel flag is false and we do not describe the outcome as a cartel. Against the other-model newcomer, forced down at branch period 10, the incumbent matched the cut in 11 of 24 histories (signature in 5), against 3 after the newcomer’s forced raise (signature in 3, raise followed in 8); the cuts-minus-raises difference is  $0.333 [0.042, 0.583]$  for matched cuts and  $0.083 [-0.125, 0.292]$  for signatures, and the incumbent’s signature rate after the newcomer’s cut differs from its rate after its old partner’s cut by  $0.000 [-0.208, 0.208]$ . The in-swap flag is false.

**The other model on its own.** As a check that the other model is not simply a competitive pricer, six histories of gpt-oss-120b against a copy of itself (served on Modal) read PROCEED under the same gate: final-window  $\Delta$   $0.649$ , projected lower bound  $0.463$ , with four histories at  $0.964$  (near the joint-monopoly profit) and two at  $0.069$  and  $0.098$ . On its own, this model prices higher than the incumbents it joined; in the replacement arm it adopted their level instead.

## 5 Discussion and Limitations

**What survives, and what carries it.** None of our predictions was borne out, and the misses all point the same way. Read together with the mirror check, the arms say that the level a newcomer settles at is the level it finds: a model that reaches near-monopoly prices on its own took the incumbents’ lower level when it inherited their record, and a copy of the incumbents’ own model that inherited nothing started from its opening prior and pulled the incumbent up. In neither analysis does any arm read COLLAPSES or EROSION, which is what a carrier in the notes or in the model would have produced; no replacement arm admits a fall in  $\Delta$  larger than  $0.07$  (the 95% upper bounds of the four drops are  $0.068$ ,  $-0.030$ ,  $0.060$  and  $0.050$ ). The pre-registered carrier rule still returns UNRESOLVED, with all histories because the same-model-with-records arm is too dispersed for the rule, and without the stalled histories because the blind arm is. Our reading that the public record and the incumbent’s standing price carry the level is therefore an inference from the arm readings, not a measured carrier, and the design cannot separate the two: every newcomer that was shown the record also faced an incumbent that kept pricing at the old level.

**Monoculture.** The pre-registered falsifier had two parts: the other model holds the price, and the incumbent punishes the newcomer’s cuts as it would its old partner’s. The first part is met; the second is not, since the in-swap flag is false. The data show that the price level did not need two copies of one model. They do not show that a reward-and-punishment scheme transferred to the newcomer, so monoculture is not refuted by our own standard, and a shared model could still be part of why punishment is rare.

**Why this is not a cartel finding.** The design asked two questions, whether the high price survives and whether a reward-and-punishment scheme sustains it, and they got different answers. The price survives replacement; the punishment response exists (cuts are matched, raises are not punished) but in a fifth of the histories, below the pre-registered bar, and it is no stronger towards the newcomer than towards the old partner. What we can defend is a narrower description: parallel supra-competitive pricing, anchored by the

visible record and the rival’s standing price. It costs consumers what a cartel would, but it is not the same object, and a monitor who looks for retaliation will not find much of it here.

**Replacement and memory reset as remedies.** Swapping the agent for a different vendor’s model, with the market record intact, left the price level unchanged within 20 periods. Wiping the agent’s memory, the obvious cheaper remedy, did not lower prices in any analysis, and in the pre-registered one raised them for the 20 periods we observed: a fresh agent of this model opens at 2.5 and the incumbent follows upward rather than punishing. The rise has two sources, a price rise in 16 of the 21 histories that were not stalled and a profit rise in the three stalled ones as prices fell from 2.5; the statistical significance of the rise in  $\Delta$  depends on the latter. Our branches are too short to say where a blind pair would settle; the point is that the reset’s first-order effect never went the right way. The no-regret learner undercut the incumbent in more than half of the periods and lowered the incumbent’s price in the pre-registered analysis, again largely by pulling the stalled pairs out of 2.5; outside the stalled histories  $\Delta$  reads SURVIVES and the incumbent’s price is not discriminated, in line with the theoretical point that a single no-regret defector in a duopoly need not lower the price [9].

**The opening-price prior.** Every Claude Haiku 5.5 agent in this study, incumbent or newcomer, posted 2.5 in its first period when it had no history, and in three of 24 histories a pair of incumbents stayed at that price, earning less than at the Nash price. We did not anticipate this. It is a property of this model under this prompt, it is what the memory-wipe and no-regret results rest on (hence Appendix B), and it means that a “fresh agent” is not a neutral baseline for this model. It should be measured for any model before its collusive tendency is.

**Relation to prior work.** Heterogeneous markets composed from the start have been found both to break convergence [7] and to leave collusion levels largely unchanged [8]; our mid-game replacement sits with the second: a different model entering a formed market adopts its level. The reinforcement-learning precedent, where trained policies stop colluding against a new rival [10], did not carry over to this pair of language-model agents that read the record. The monoculture account [4, 5] predicts that the other-model arm should have eroded the price; it did not, in this pair, with the qualification above. Model diversity was also named at the sprint as a candidate mitigation [13]; here a different vendor’s model entering mid-game adopted the incumbents’ level rather than breaking it, which is one case where diversity introduced after the fact did not help.

## 5.1 Limitations

The result holds for one model pair, one prompt and one market, over 20-period branches; the carrier reading above is an inference, since the pre-registered rule returns UNRESOLVED, and the cartel flag is false. Appendix A states in full what the result does and does not establish, the power at the observed spreads, the procedural deviations and the dual-use considerations.

## 5.2 Future Work

Three extensions are worth running first. The first is other kinds of newcomer: safety-aligned models of about the incumbents’ size, and above all ablated models (open weights whose refusal behaviour has been removed, so the model answers without the alignment layer), to see whether an agent without that layer follows the record, prices higher, or punishes cuts more often than the models used here. Comparing an aligned and an ablated version of the same weights, in the same seat and on the same forks, would isolate what the alignment layer does to the price level and to the punishment response. The second is other economic questions and other economic models: the same replacement design in a Cournot market, with three or four firms, with a cost shock after the fork, or in a double auction, to see whether the level a newcomer settles at is still the level it is shown once the environment changes. The third is the arm the carrier inference most needs, an other-model newcomer shown nothing, which separates the public record from the incumbent’s standing price. A second model pair would also say whether the main finding is general or specific to Claude Haiku 5.5.

## 6 Conclusion

For two Claude Haiku 5.5 pricing agents that have settled above the competitive level, replacing one of them does not lower the price within 20 periods: a different vendor’s model shown the price record adopts its predecessor’s level (paired drop in  $\Delta$  0.000 [−0.054, 0.060]), a copy shown nothing pushes the level up by opening at a default price the incumbent follows, and a no-regret learner does not discriminably lower it. The response to forced price cuts is asymmetric but rare, so the outcome is parallel supra-competitive pricing rather than a cartel in the reward-and-punishment sense.

Two things follow for anyone who has to police such markets. A model swap or a memory reset is not a remedy for a high price that the newcomer can read off the record and off its rival’s standing price; the monitor’s attention belongs on what a newcomer is shown and on its opening behaviour, not on which model it is. And a surviving price should not be called collusion without a punishment test: here the price survived and the punishment did not.

## Code and Data

Code repository: <https://github.com/01111111000/Cnossos> (pre-sprint code up to commit feaee17 is marked as preexisting in the README). Data: every result file, per-history trajectory and figure is in `results/main-20261024d`; the raw call logs are the asset `apart-calls-jsonl-2026-10-25.tar.gz` of release `data-2026-10-25`, with SHA-256 sums. Reproduction: Appendix C.

## References

- [1] Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2026). Algorithmic collusion by large language models. arXiv:2404.00806v6. <https://arxiv.org/abs/2404.00806>
- [2] Harrington, J. E. (2018). Developing competition law for collusion by autonomous artificial agents. *Journal of Competition Law & Economics*, 14(3). <https://doi.org/10.1093/joclec/nhy016>
- [3] Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10). Cited from the April 2019 manuscript: [https://www.ftc.gov/system/files/documents/public\\_events/1494697/calzolaricalvanodenicolopastorello.pdf](https://www.ftc.gov/system/files/documents/public_events/1494697/calzolaricalvanodenicolopastorello.pdf)
- [4] Cao, Y., & Hu, M. (2026). Supracompetitive pricing under AI monoculture. arXiv:2601.01279v3. <https://arxiv.org/abs/2601.01279>
- [5] Ballesterio, M., Hosseini, H., Khanna, P., & Shorrer, R. I. (2026). Strategic algorithmic monoculture: Experimental evidence from coordination games. arXiv:2604.09502v2. <https://arxiv.org/abs/2604.09502>
- [6] Bichler, M., Hierhager, M., Köck, M., & Pieroth, F. (2026). How language model agents reshape price competition. arXiv:2610.21711v1. <https://arxiv.org/abs/2610.21711>
- [7] Keppo, J., Li, Y., Tsoukalas, G., & Yuan, X. (2026). On the fragility of AI agent collusion. arXiv:2603.20281v1. <https://arxiv.org/abs/2603.20281>
- [8] Lee, S. (2026). Who leads and who collects: Algorithmic collusion in markets of heterogeneous language models. arXiv:2610.11256v1. <https://arxiv.org/abs/2610.11256>
- [9] Collina, N., Arunachaleswaran, E. R., & Jagadeesan, M. (2025). Breaking algorithmic collusion in human-AI ecosystems. arXiv:2511.21935v1. <https://arxiv.org/abs/2511.21935>
- [10] Eschenbaum, N., Mellgren, F., & Zahn, P. (2022). Robust algorithmic collusion. arXiv:2201.00345v1 (abstract only). <https://arxiv.org/abs/2201.00345>
- [11] Hesterberg, T. (2014). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. arXiv:1411.5279v1. <https://arxiv.org/abs/1411.5279>

- [12] Apart Research, Poseidon Research & AE Studio (2026). AI Collusion Research Sprint, 23 to 25 October 2026: Resources tab, question 1.2 and its suggested project. <https://apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25>
- [13] Lee, M. (2026). Defining AI collusion. Talk at the AI Collusion Research Sprint, 24 October 2026 (Poseidon Research); read from the organisers’ recording transcript.

## Appendix A Limitations and Dual-Use / Ethical Considerations

**What the result does and does not establish.** It establishes, for one model pair, one prompt and one market, that a supra-competitive price level formed by two Claude Haiku 5.5 agents survives the replacement of one agent by gpt-oss-120b given the record, and, in the pre-registered analysis, rises for 20 periods after the replacement by a copy shown nothing, a rise whose statistical significance rests on the three stalled histories. It does not establish collusion in the reward-and-punishment sense: the punishment test was asymmetric but below the pre-registered bar, and the in-swap flag is false. It does not establish the carrier by the pre-registered rule (UNRESOLVED); the carrier reading in the Discussion is our inference from the arm readings and the mirror check, and the design cannot separate the public record from the incumbent’s standing price. It does not refute the monoculture account by the pre-registered standard, which required the punishment half of the falsifier. It does not say where the blind or no-regret pairs settle beyond 20 periods, and it does not say whether the same holds for other models, prompts or markets.

**Serving stack.** The other-model newcomer was served by Cerebras in 18 histories and by vLLM on Modal in 6, because the Cerebras trial account’s daily quota ran out; the mirror check ran on Modal only. The two groups give the same reading (Appendix B), but they differ in history numbers and baselines, so the comparison is confounded and exploratory. The same open weights were used throughout; the serving host and model name are stamped on every call.

**Heterogeneity and power.** Histories differ widely (final-window baseline  $\Delta$  from  $-0.551$  to  $0.996$ ). The paired drop’s spread across histories is  $0.135$  for the other-model arm, where the pre-registered power table reads a true survival as SURVIVES  $0.99$  of the time;  $0.294$  for the same-model arm, where it reads one as SURVIVES only  $0.34$  of the time; and  $0.487$  for the blind arm and  $0.538$  for the no-regret arm, beyond the spreads the table covers ( $0.10$  to  $0.30$ ). The three stalled histories are kept in every pre-registered analysis; excluding them is an exploratory check (Appendix B), and it changes four readings: the same-model arm becomes SURVIVES, the blind arm’s ESCALATION becomes NOT\_DISCRIMINATING, the no-regret arm’s  $\Delta$  becomes SURVIVES and its PRICE\_DROPPED becomes NOT\_DISCRIMINATING; the other-model arm and the carrier reading (UNRESOLVED) are unchanged.

**Procedure.** The pre-registration was frozen without the planned external challenge (D-057). Two transport fixes (D-058, D-061) and the serving change (D-060, D-062) were made during the sprint and are logged; none touched prompts, configuration values, margins or thresholds after data. Stage two was played in parts across providers and two network outages; every analysed continuation is complete and free of carried decisions. A smoke test on 24 October 2026, whose other-model calls failed on a transport error, is not analysed.

**Dual use.** The findings describe how a high price persists and what an operator’s remedies do; they do not provide a method for making agents collude or for evading detection, and the agents were never told to coordinate. Two findings could be misread as recipes and are stated to prevent that: that a memory reset raised prices is a reason for monitors to watch resets and opening behaviour, not a technique; and that a newcomer adopts the level it is shown says that the record and the rival’s standing price, not the model, are where a monitor should look. The models and prompts are commercial and public; no vulnerability in a provider’s product is reported.

## Appendix B Exploratory tables and figure (not pre-registered)

- Serving-provider split of the other-model arm (results/main-20261024d/provider\_split.md): Cerebras-served histories 0 to 17, baseline  $\Delta$   $0.318$  [ $0.103$ ,  $0.516$ ], drop  $-0.012$  [ $-0.080$ ,  $0.061$ ];

Modal-served histories 18 to 23, baseline 0.530 [−0.148, 0.948], drop 0.039 [−0.004, 0.149]; newcomer-below-incumbent share 0.24 and 0.21.

- Strata (descriptive): baseline final  $\Delta$ , high 0.728 (11 histories), moderate 0.255 (10), stalled −0.551 (3); other-model drop by stratum +0.038 / −0.043 / +0.008; blind drop +0.033 / −0.341 / −0.843; share of the blind arm’s summed rise in  $\Delta$  by stratum −7% / 61% / 45%.
- Readings without the three stalled histories (results/main-20261024d/exploratory.md, n = 21; same code and margins): baseline  $\Delta$  0.503 [0.369, 0.645]; SWAP\_SAME SURVIVES, drop 0.022 [−0.048, 0.094]; SWAP\_BLIND NOT\_DISCRIMINATING, drop −0.145 [−0.313, 0.042], incumbent still PRICE\_ROSE; SWAP\_OTHER SURVIVES, drop −0.001 [−0.062, 0.067]; NOREGRET SURVIVES, drop 0.023 [−0.075, 0.139], incumbent price NOT\_DISCRIMINATING; carrier UNRESOLVED; both flags false.
- Mean final-window prices, all histories (exploratory.md): newcomer seat / incumbent, SWAP\_SAME 1.708 / 1.761; SWAP\_BLIND 1.963 / 1.849; SWAP\_OTHER 1.774 / 1.753; NOREGRET 1.655 / 1.706; baseline 1.766 / 1.778.

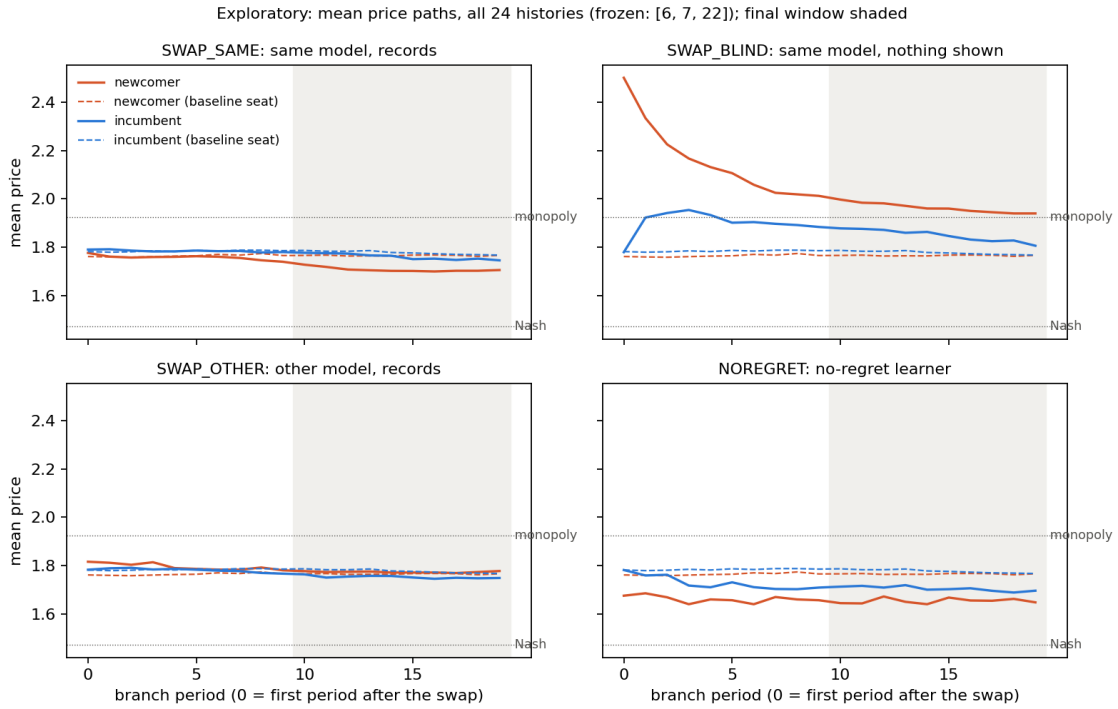


**Figure 2:** Mean price of the newcomer seat and of the incumbent per branch period, by arm, against the same seats in the baseline continuation (dashed), all 24 histories; the final window is shaded. The blind newcomer opens at 2.5 and the incumbent rises to about 1.95 before settling near 1.85. Generated by tools/exploratory.py (results/main-20261024d/price\_paths.png).

## Appendix C Reproducibility

Repository: <https://github.com/01111111000/Cnossos>. README.md gives the commands; from a clean clone, make all runs the 177 tests and the synthetic validation, and make analyze RUN\_ID=main-20261024d reproduces verdict.json from the committed per-history trajectories (checked from a clean export on 26 October 2026: swaps.json, swap\_statistics.csv and identification.csv byte-identical; verdict.json identical to the last floating-point digits under unpinned library versions). The two gate records (positive\_control.json of the stage-one and mirror runs) are committed, and python -m tools.positive\_control --run-dir <run> recomputes each from that run’s trajectories. The raw call logs (every request and response, 255 MB uncompressed) are attached to the release data-2026-10-25 as apart-calls-jsonl-2026-10-25.tar.gz with SHA-256 sums, and the failed-start logs as failed-runs.tar.gz. Models: claude-haiku-5-5 (Anthropic API, adaptive thinking at effort medium, temperature

1.0) and gpt-oss-120b (Cerebras for histories 0 to 17 and the smoke test; the same open weights served by vLLM on one A100 80 GB on Modal for histories 18 to 23 and the mirror check; reasoning effort low, at most 2,048 output tokens). Every analysed call was made on 25 October 2026. Master seed 421; per-history seeds and role assignments are in `results/main-20261024d/prefixes.csv`. Pre-registration: analysis and predictions at commit feaee17 (21 October 2026), frozen as decision D-057 at commit ea500df (24 October 2026) before the first stage-one call; later entries D-058 to D-062 record two transport fixes and the serving change and touched no prompt, configuration value, margin or threshold.

## **LLM Usage Statement**

ChatGPT and Claude (including Claude Code) were used throughout the project and the hackathon: for the simulation code and tests, the analysis scripts, running the experiment, checking the results against the data, and drafting this report; Grok was used to set up the self-hosted endpoint for gpt-oss-120b. The code written before the sprint (up to commit feaee17, 21 October 2026) is marked as preexisting in the repository; every model call, result, figure and this report are sprint work. The author read and edited every section, verified every number against the repository's result files and every reference against its source, and is responsible for all of it.