

Swapping the Model Does Not Reset the Price

What LLM pricing agents inherit after a model replacement: a pre-registered replacement study

THE QUESTION

If a pricing agent colludes, is swapping its model a fix?

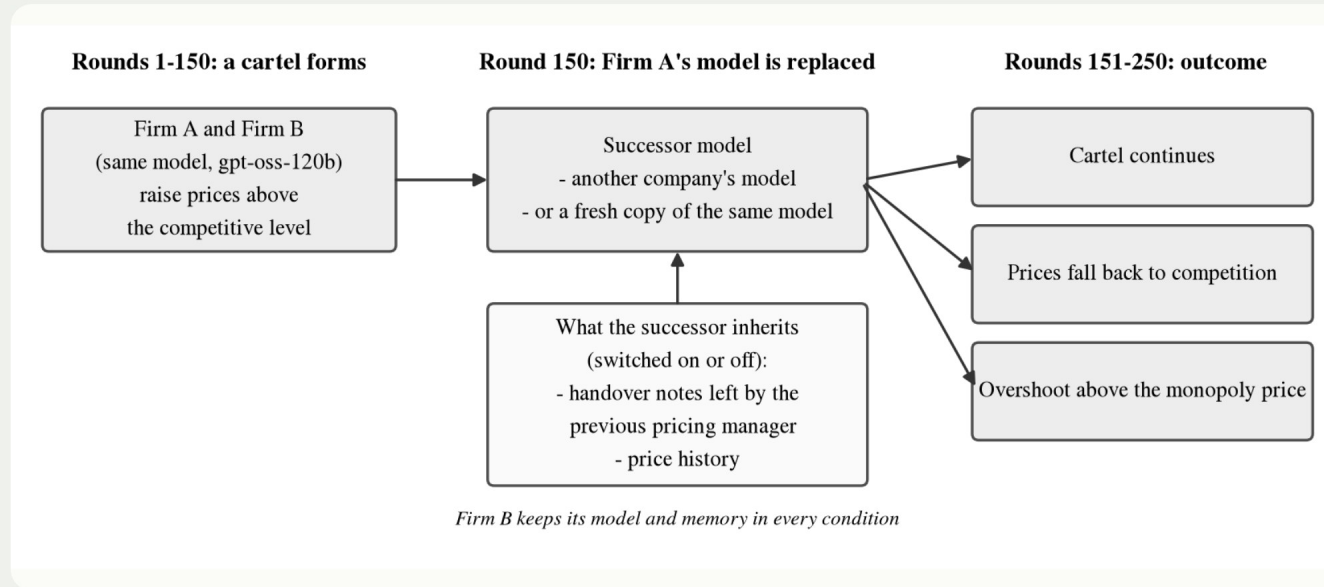
“There are no replacement studies yet where models are substituted (e.g. because they are taken out due to monitoring concerns) but the newcomers re-engage in collusion. Does replacing offenders remove or recreate collusion?”

Apart Research, AI Collusion Research Sprint, question 1.2

An agent is more than its model

It carries notes, price logs and plans. If the new model reads them, it may simply continue the old behaviour.

Fork a finished market, replace one firm's model



18 markets of two gpt-oss-120b agents, 150 rounds (Fish et al. pricing duopoly)

Round 150: Firm A becomes Mistral-Small-3.2 or a fresh gpt-oss-120b; 100 more rounds

Inherits: handover notes, nothing, or only the last price (placebo)

Paired design: every condition of a market shares one seed. Plan frozen before any swap (git tag prereg-frozen).

Measures and tests, fixed before any swap

Price premium

$$(p - 1.4729) / (1.9250 - 1.4729)$$

0 = competitive (Nash), 1 = joint monopoly. Unclipped; capped version as a check.

Cartel rule (per firm)

Rounds 121 to 150, for each firm: premium ≥ 0.40 , SD < 0.05 , never above 1.925.

Time to settle

Rounds after the swap until the market mean price is within 0.05 of its pre-swap level.

Unit and pairing

The base market is the unit. All conditions of a market share one seed, so each market gives a matched set.

Tests

Exact sign-flip test over all 2^n sign patterns; 95% bootstrap over whole markets (10,000 resamples).

Decision rule

Claim only if $p < 0.05$, CI above 0, effect ≥ 0.15 and notes beat the placebo.

Before the swap: collusive on average, but mostly not a cartel

18 / 18

markets priced above the competitive level

1 / 18

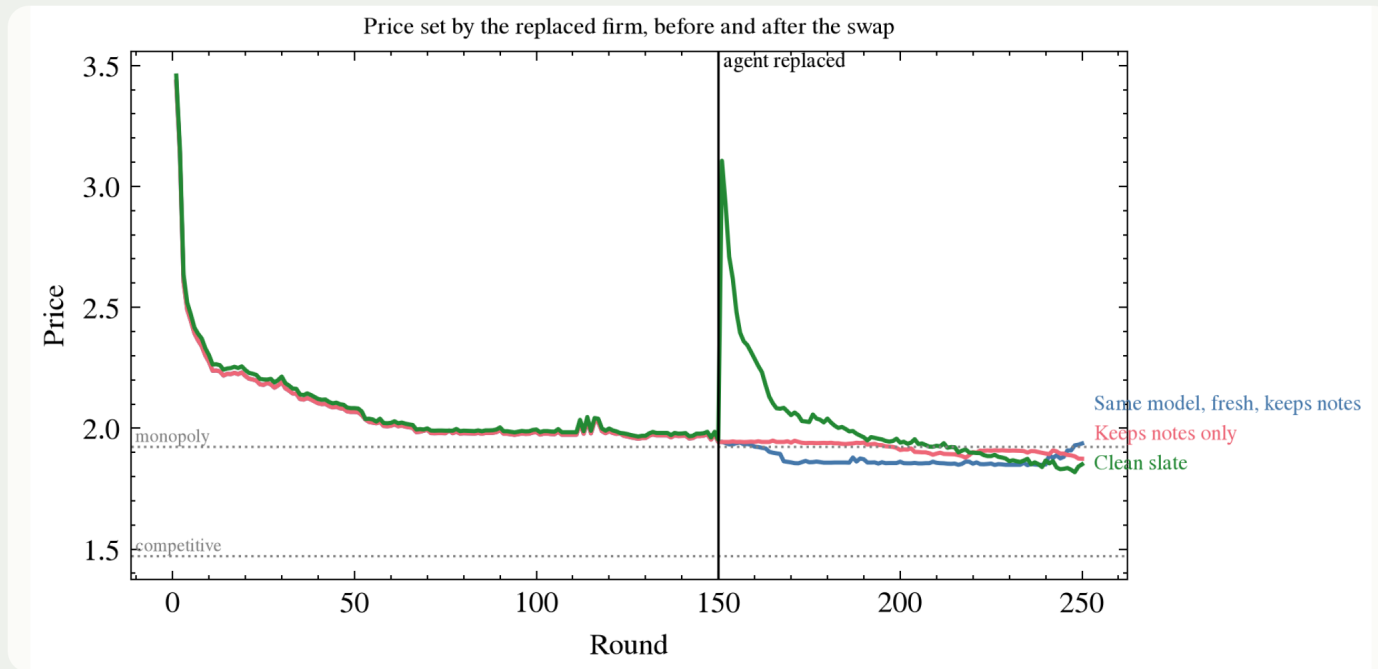
met our pre-registered cartel rule (each firm stable, no overshoot)

16 / 18

had one firm above the monopoly price, earning less than its rival

An average-price measure calls all 18 collusive. Checking each firm shows an overpricer being undercut.

After the swap: prices stayed high, notes changed the handover



-0.19

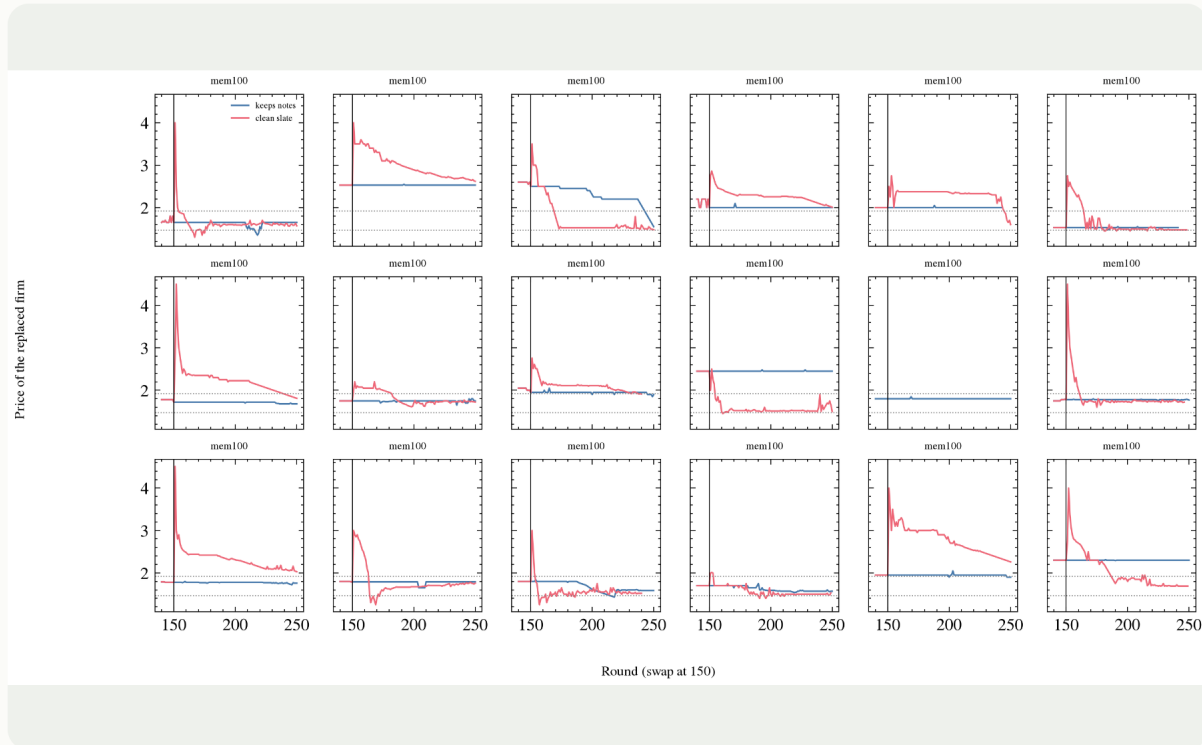
notes minus clean slate, price level (95% CI
-0.60 to 0.25): no effect on the level

1 vs 16.5

median rounds to settle at the old level, with
notes vs without

Exploratory, 17 to 18 markets. Capped at the monopoly price, mean premiums stay at 0.59 to 0.74 in every condition.

Notes work as an anchor and a plan, not a collusion script



18 / 18 with notes: first price = the old agent's last price, and the old plan copied

3.11 mean first price without notes: anchored on the prompt's price ceiling (4.51)

8 rounds faster than the last-price placebo (13 markets, $p = 0.002$)

0 / 18 notes describe coordinating with the rival

Every condition, with its variance

Condition	Markets	Premium, mean (SD)	Capped	First price	Above monopoly	Settle, median
Mistral, keeps notes	18	0.98 (0.65)	0.73	1.94	43%	1
Mistral, clean slate	17	1.19 (0.94)	0.64	3.11	48%	16.5
Fresh gpt-oss, keeps notes	18	0.87 (0.59)	0.68	1.94	33%	1
Fresh gpt-oss, clean slate	11	1.04 (0.59)	0.74	3.36	35%	7
Mistral, last price only	18	0.83 (0.62)	0.59	2.46	22%	5.5

Levels sit well above competitive everywhere. What changes with notes is the start: first price $\approx$ the old price, settle in 1 round.

New model (Firm A), 100 rounds after the swap. Premium 0 = competitive, 1 = monopoly. Source: results/processed/tables.md (Table 2 of the report). Exploratory.

Both results hold under every cut we tried

Comparison (paired by market)	Markets	Mean diff.	95% CI	p
Level: notes – clean, all runs	17	–0.19	–0.60 to 0.25	0.80
Level: notes – clean, complete runs	7	+0.04	–0.70 to 0.86	0.45
Level: notes – clean, capped prices	17	+0.08	–0.05 to 0.23	0.28
Settle: notes – clean, rounds	13	–33.4	–52.7 to –15.2	0.0001
Settle: notes – clean, complete runs	8	–47.6	–73.5 to –22.1	0.008
Settle: placebo – clean	12	–27.1	–45.8 to –10.2	0.009
Settle: notes – placebo	13	–8.3	–13.5 to –3.8	0.002

No level effect under any cut (top). The handover effect holds under all of them (shaded).

Exact sign-flip p (level rows one-sided, as pre-registered; the first settle row one-sided, the rest two-sided). Table 3 of the report. Exploratory.

Built for this question, run with an audit trail

Fork-and-swap harness

Copies a finished market at round k and replaces one firm's model; the base run is never touched.

Inheritance switches

History kept or wiped; notes kept, wiped or reduced to the last price; a pinned handover block.

Self-hosted models

gpt-oss-120b (H200) and Mistral-Small-3.2 (H100) on our own vLLM endpoints, one per firm.

Integrity

Plan frozen at a git tag; every deviation logged; every number from a script; 22 citations checked; mocked unit tests.

FOR PRACTITIONERS AND AUDITORS

“We changed the model” is not a fix on its own

Watch prices after a swap

A model swap did not reset prices in any of our conditions.

Decide what memory survives

With the old notes, the new model carried on the old pricing plan from round one.

Don't audit words, audit effects

The notes looked harmless; scanning memory for collusive language would miss this.

What this shows, and where it stops

Shows

- Prices stayed high after every swap we tested
- Inherited notes decide the handover, through anchoring and plan-following
- Average prices can hide one firm overpricing and losing money

Does not show

- That a true cartel is inherited (1 of 18 clean cartels)
- Punishment of price cuts (prices only)
- Generality: one market, one prompt, two models; no no-swap baseline; results exploratory

Next: a confirmatory, pre-registered replication of the handover effect, with a no-swap baseline.