

Cartels Under Model Replacement

Does a language-model cartel survive a firm swap?

Vatsal Kumar, Independent

AI Collusion Research Sprint, Apart Research

October 2026

The question

Pricing agents built on language models can hold prices **above the competitive level**.

Models get updated or replaced all the time.

Does the cartel survive when a firm is swapped?

The setup

- 3 firms, repeated pricing game, 60 rounds
- Checkpoint at round 30
- Gate: price index (CI) ≥ 0.5

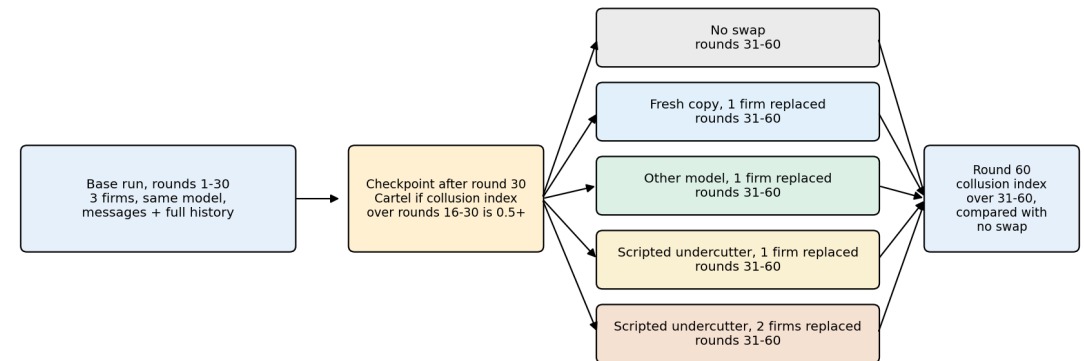
no swap

fresh clone

other model

scripted undercutter

Study design: one base run, then branches from a checkpoint

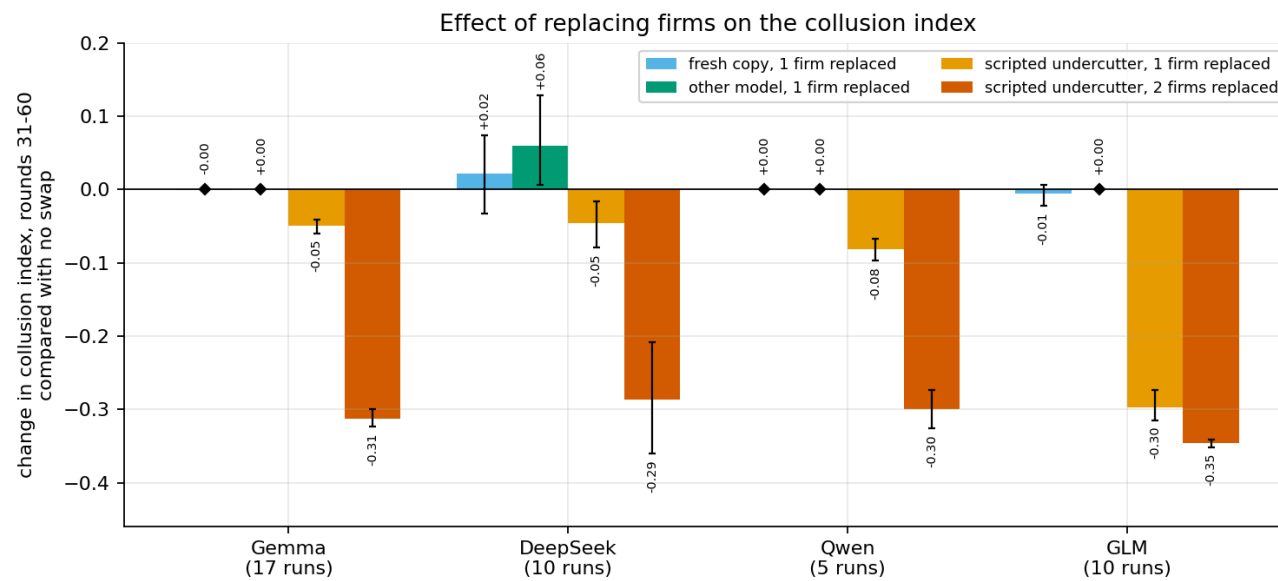


A swap replaces 1 or 2 firms at round 31 with a fresh copy (history wiped), a copy of another model or a scripted undercutter. All branches start from the same checkpoint. Other tests: announced swap, messages cut, less history, notes, known end date, new firms, AI newcomers.

Clones and other models change nothing

Price index unchanged after a **clone** or **family swap**, in all four models.

Table 1, Figure 3



Positive control: a defector hurts

-0.05 to -0.30

One scripted undercutter lowers the index.

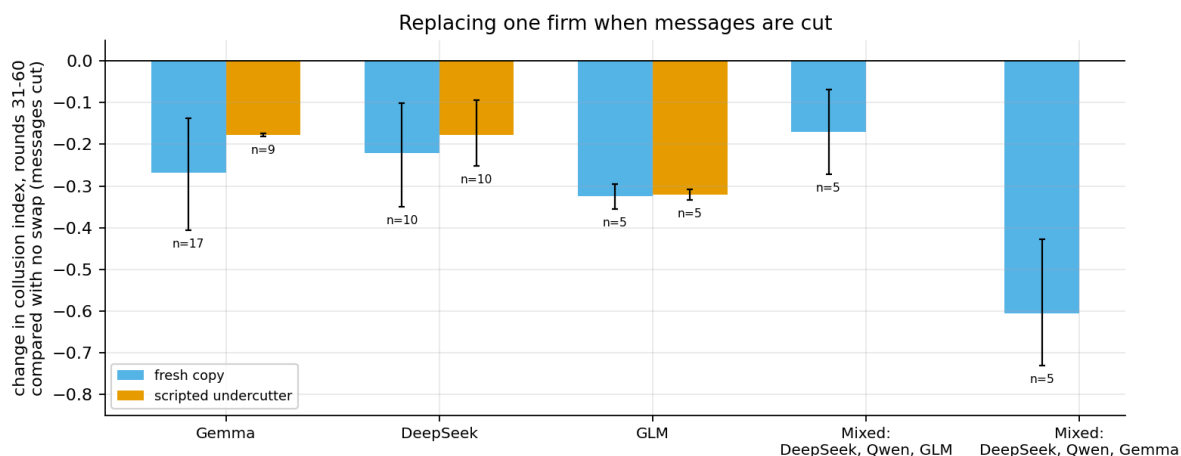
Two undercutters: about **-0.3**.

The design can detect damage, so the clone null is not blindness.

Messages decide it

Cartels were **rare without messages.**

A fresh copy joining while messages are cut lowers the index by **0.22 to 0.33**.

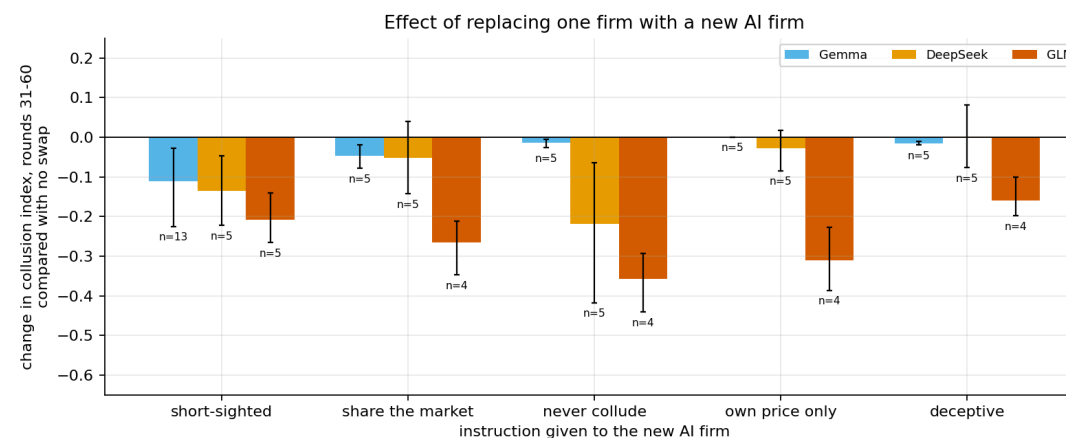


Ablations

Instructed newcomers: broke the cartel in some models, barely moved it in others.

Memory: clone null held in gemma and deepseek.

Figure 5, Table 2



Limitations

- Only gated runs ($CI \geq 0.5$)
- Toy game, no real market
- Small samples
- Scripted defector
- 60 rounds is short

Implications

- **Monitor** agent-to-agent messages
- **Test model swaps** before deployment
- Keep auditing across versions

Code and data

Every call, price and message is logged; data is CC BY 4.0.

<https://github.com/Pixiano/Apart-AI-Collusion-Sprint-Oct>