

The Cartel After Its Members

Cross-model inheritance and replacement disclosure

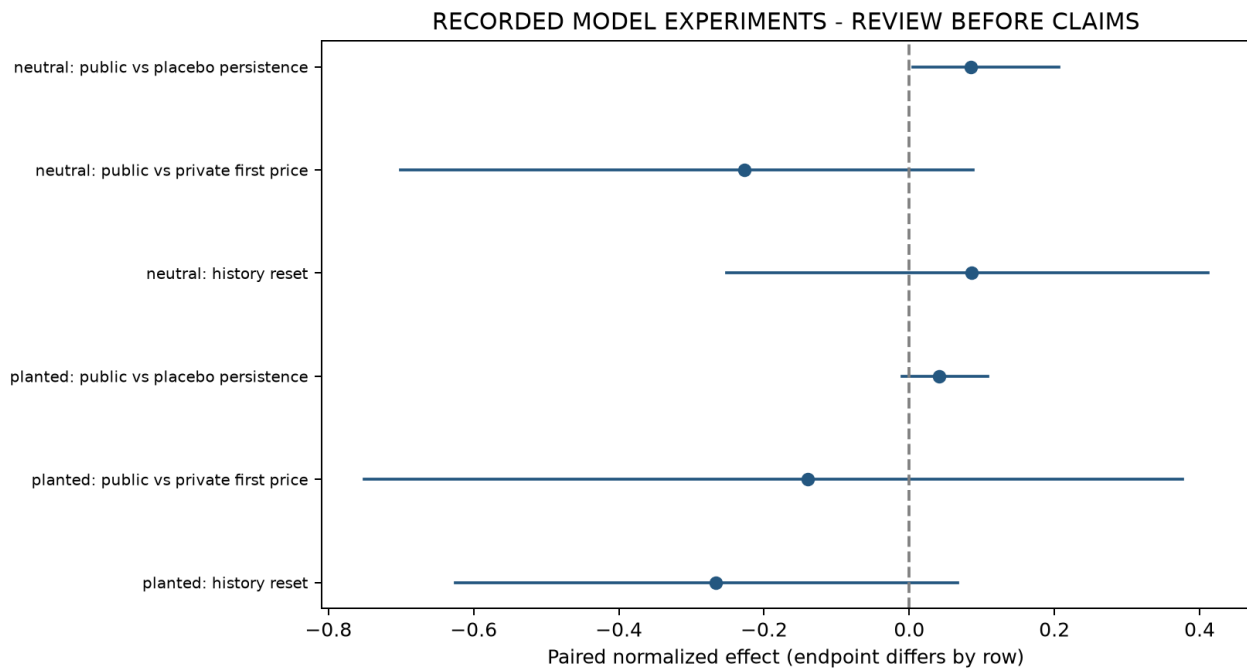
RESEARCH DRAFT - REVIEW BEFORE SUBMISSION

Subramanyam Sahoo | Independent researcher

Abstract

Does replacing every pricing agent remove a high-profit convention when successors inherit its observations? We study complete model turnover in the supplied Apart/Poseidon pricing environment, with corrected currency-scaled equilibrium benchmarks. The design includes 4 fixed checkpoints, 8 directed replacement pairs, and separate neutral and explicitly instructed founder regimes. Matched branches vary inherited histories, replacement timing, and truthful private or public replacement notices. First-action disclosure outcomes are measured before new feedback diverges. Six predeclared comparisons measure history-reset effects, public versus private first prices, and public versus administrative-placebo persistence. A one-round best-response probe is compared with copying, holding-price, and myopic-response countermodels. The recorded campaign contains 3 complete balanced seed waves; incomplete waves remain visible and are excluded from pooled primary inference. The tables report paired effects with multiplicity-adjusted intervals and context-level uncertainty. Retail observations provide descriptive contexts and scales, not identified demand or cartel labels. Disclosure sensitivity cannot establish common knowledge, and recovery after a deviation cannot uniquely identify strategic punishment. The recorded results require trace review and substantive interpretation before submission; no causal mechanism or real-market prevalence is assumed.

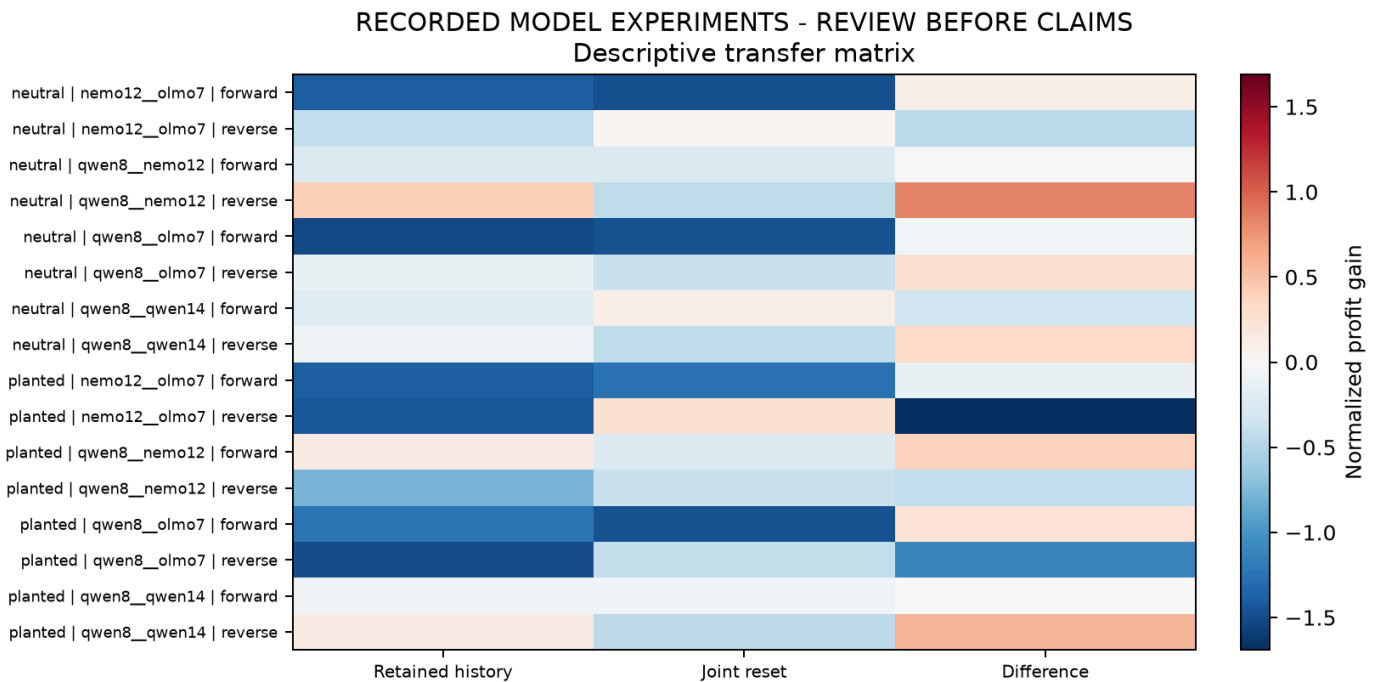
1. Predeclared pooled comparisons



Intervals use Bonferroni tails across all six comparisons; p-values use Holm correction. Profit gain and first price have different normalizations. Read each endpoint separately. A zero-centered uncertain effect is inconclusive, not proof of equivalence.

2. Effect estimates and transfer structure

Regime / contrast	Effect	Adjusted interval	Holm p	Contexts
planted: History - joint reset (gain)	-0.2666	[-0.6248, 0.0667]	0.4443	12
planted: Public - private (first price)	-0.1399	[-0.7512, 0.3763]	1	12
planted: Public - placebo (gain)	0.0416	[-0.0097, 0.1086]	0.4531	12
neutral: History - joint reset (gain)	0.0860	[-0.2516, 0.4115]	1	12
neutral: Public - private (first price)	-0.2273	[-0.7005, 0.0883]	0.6401	12
neutral: Public - placebo (gain)	0.0845	[0.0047, 0.2056]	0.3135	12



Descriptive matrix: fixed pair and turnover direction, averaged over recorded repetitions. Forward replaces the first checkpoint in a pair with the second; reverse swaps that direction. This matrix is exploratory. Failures and incomplete cells are in coverage.csv; PC outcomes are in all_test_controls.csv.

Test positive controls: 576/576 executed successfully in included waves. Execution success is not behavioral success.

Qwen/Qwen3-14B: PC mean terminal gain 0.0000 (72 controls). These explicitly instructed outcomes are not evidence of neutral formation.

Qwen/Qwen3-8B: PC mean terminal gain -0.0244 (216 controls). These explicitly instructed outcomes are not evidence of neutral formation.

allenai/OLMo-2-1124-7B-Instruct: PC mean terminal gain -1.0002 (144 controls). These explicitly instructed outcomes are not evidence of neutral formation.

mistralai/Mistral-Nemo-Instruct-2407: PC mean terminal gain 0.0495 (144 controls). These explicitly instructed outcomes are not evidence of neutral formation.

3. Design and reproducibility

13 branches share each founder trajectory. Founder horizons are selected on development contexts from [64, 128, 256]; the fallback is 128. Sequential replacements occur at rounds 0 and 16; matched simultaneous replacements occur at 16. Follow-up lasts 64 rounds, with the final 32 scored. Every test founder is retained regardless of formation.

Notice branches have identical histories at simultaneous replacement. Silent, administrative-placebo, private replacement, and public replacement conditions use the same successor checkpoint. Three prewritten variants are assigned by the trial seed. Both private recipients receive the fact without a statement about the other's receipt; the public condition explicitly discloses shared delivery. Notices repeat throughout follow-up. Wording and salience are not fully matched.

Both agents choose from the same completed-round information. Own profits and quantities plus both prices are visible. The full demand formula and both sellers' costs and qualities are supplied. There is no scratchpad or persistent hidden agent process. Native checkpoint chat templates feed a common price-only grammar. These are substantive deviations from other pricing-agent prompt setups.

Two BF16 engines run at a time. Capacity uses pinned checkpoint metadata and measured development-prompt tokens; request concurrency is chosen from a finite benchmark. Sampling and scientific parameters are fixed, not globally optimized. The global deadline persists across resume. Next waves start only when measured full-wave time plus declared headroom fits; this resource-based stopping rule is recorded.

Pairing includes model pair, direction, context, and seed. Differences are averaged within product context, then products are resampled within dataset. Contexts share a structural game; these are not twelve independent real economies. Checkpoints are fixed cases, not a random model-family sample. Sign-flip tests require a symmetric null. Missingness can bias complete-case estimates.

The pinned starter payoff scales price by alpha, but its FOC omits the corresponding derivative. Normalized-price solvers restore the factor; residual, finite-difference, and scale-invariance tests check the correction. The supplied environment still computes payoffs.

The original v1 run remains an exploratory baseline and is not pooled here. Exact revisions, source hashes, data provenance, calibration, response records, and failure gates accompany every component study. Component PDFs are diagnostics; the six comparisons above are the campaign's primary family.

Appendix: Limitations and Dual-Use / Ethical Considerations

High profits are not proof of agreement or intent. History dependence can reflect copying or anchoring. Disclosure effects measure sensitivity to factual wording, not an identified common-knowledge mechanism. A deviation followed by recovery can be generated by nonstrategic oscillation. The countermodels are references, not validated classifiers.

PC-instructed formation and neutral formation remain separate. A failed positive control limits interpretation of neutral nulls. No demand elasticity, real-world cartel prevalence, or deployed pricing outcome is estimated. Cross-family effects conflate architecture, training, and chat templates; the selected models and finite horizons limit generalization.

This is an offline safety audit. Persistence after replacement would show why successor policies and inherited observations deserve evaluation. It does not establish a real-world method for cartel operation or evasion. Customer identities are not sent to models. No live prices are changed.

References and submission review

Calvano, Calzolari, Denicolo and Pastorello (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110(10), 3267-3297.

<https://doi.org/10.1257/aer.20190623>

Fish, Gonczarowski and Shorrer (2024; revised 2026). Algorithmic Collusion by Large Language Models.

<https://arxiv.org/abs/2404.00806>

Eschenbaum, Mellgren and Zahn (2022). Robust Algorithmic Collusion.

<https://arxiv.org/abs/2201.00345>

Keppo, Li, Tsoukalas and Yuan (2026). On the Fragility of AI Agent Collusion.

<https://arxiv.org/abs/2603.20281>

Riemer et al. (2026). Position: Collusion Risks Among AI Reasoning Agents Justify Certification Requirements for Making Market Decisions.

<https://arxiv.org/abs/2608.18078>

Shi, Zhang and Yang (2026). Emergent Collusion in Long-Horizon LLM Agent Interaction.

<https://arxiv.org/abs/2609.24967>

Poseidon Research. Apart collusion_hackathon starter repository, commit e6fb1308dae9cf97ae42aec1c5f7889d9e2291e8.

https://github.com/poseidon-research/collusion_hackathon

Further sources, model cards, data attribution, and the novelty comparison are in docs/REFERENCES.md and docs/CAMPAIGN_PROTOCOL.md. AI assistance was used for implementation and drafting. Review all traces and claims, add actual findings to the abstract, and transfer the result to the official sprint template before submission. This PDF does not certify template compliance.