
Algorithmic Diversity Can Fail: A Tolerant Q–LLM Collusion Counterexample¹

Arad Maleki · Mohammadpayam Taebi
The Chinese University of Hong Kong

With
Apart Research

Abstract

Algorithmic diversity has been proposed as a safeguard against AI price collusion. We report a counterexample using tolerant Q-learning, which downweights negative value-prediction errors and makes disappointing outcomes less influential during learning. We compare it with ordinary Q-learning, which weights positive and negative errors equally. Both optimize only their own monetary profit. Each learner faces three LLM rivals in a four-seller market without direct communication. We observe that ordinary Q improves profit without sustained collusion. Tolerant Q sustains cooperative pricing and develops a small late-game profit advantage over its LLM partners. The result challenges treating architectural diversity as sufficient protection against collusion. A reinforcement learner can learn to benefit from a cooperative LLM market, making adaptation and the distribution of gains essential targets for safety audits. This shifts the focus from architecture labels to the behavior learned through repeated interaction.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

A reinforcement learner can learn to join an LLM pricing coalition. We compare two versions of Q-learning, which learns price values from profit feedback. Ordinary Q-learning weights positive and negative value-prediction errors equally. Tolerant Q-learning downweights negative errors, making disappointing outcomes less influential during learning.

In our four-seller market, tolerant Q sustained cooperative pricing with three LLM rivals and gained a small late-game profit advantage over them. Ordinary Q improved profit but failed to sustain collusion. This challenges the assumption that introducing a different learning architecture is enough to prevent collusion.

Both learners optimize their own monetary profit; tolerance changes how feedback shapes learning. The contrast focuses attention on learning dynamics within an architectural family, which a simple LLM-versus-Q classification misses. Diversity-based safeguards must therefore be tested after adaptation to the other market participants.

We make two contributions: a mixed-algorithm counterexample to treating diversity as a sufficient safeguard, and an analysis of asymmetric cooperation in which the Q seller gains more than its LLM partners near the end of a recorded game. Audits must therefore measure who benefits from cooperation as well as whether prices remain elevated.

2. Related Work

Keppo et al. [1] found that algorithmic heterogeneity disrupted sustained collusion when a tabular Q learner faced an LLM in duopoly, motivating diversity as a policy lever. Our result identifies a failure mode for that proposed safeguard: a recurrent tolerant Q learner can sustain cooperation with LLM rivals after longer training. Their transferred tabular learners and our four-seller recurrent setting differ; the contribution challenges the general sufficiency of diversity, while preserving the scope of their empirical finding.

Calvano et al. [2] established collusion among Q-learning agents. The new question here is cross-algorithm cooperation and the unequal gains it supports. Tolerant learning relates to hysteretic Q-learning [3], which attenuates negative updates in multi-agent learning. Here that treatment operates on own-profit learning signals and distinguishes a colluding learner from an ordinary-Q comparison that fails to sustain cooperation.

3. Methods

Four sellers simultaneously select grid prices, receive logit demand, and earn their own profit. Each observes completed market history; current choices and rivals’ private notes are unavailable. There is no shared reward or direct messaging. We compare ordinary Q plus three LLMs, tolerant Q plus three LLMs, and an all-LLM positive control. Profit comparisons use seller 4 in the reference market and mean contemporaneous LLM-rival profit within each mixed market.

Ordinary and tolerant Q-learning

The temporal-difference (TD) error is the target return minus the predicted Q-value. Positive errors indicate outcomes better than predicted; negative errors indicate outcomes worse than predicted. Ordinary Q-learning gives both signs equal loss weight (1). Tolerant Q-learning assigns weight 0.25 to negative errors and weight 1 to nonnegative errors. It is tolerant of disappointing feedback; its reward remains own monetary profit.

Both variants use a 128-unit GRU, a 128-unit ReLU head, a Double DQN target, and smooth-L1 TD loss. Adam uses learning rate 0.0003. From the 64th transition, each fresh transition triggers four replay minibatches of size 8. Full settings appear in Appendix D.

A game contains 3,000 rounds; its endpoint is unannounced. Under the fixed-suite runner’s convention, 120 games are run per learner per seed, giving 360,000 fresh transitions per seed and 1,080,000 per variant across seeds 0, 1, and 2. Learners retain weights, optimizer, and recent replay across games; market history and LLM notes reset. Training exploration declines from 0.20 to 0.05 over the first 1,000 transitions and then stays at 0.05.

Outcomes and evaluation

Total profit sums round profits; discounted profit uses $\gamma = 0.95$. Collusion round share counts rounds in qualifying sustained windows. Following the implementation, a 100-round window qualifies if price dispersion (maximum minus minimum, divided by minimum) is at most 5% each round, mean price exceeds the competitive benchmark by at least 5%, and market profit index is at least 0.10. That index scales average seller profit between competitive (0) and joint-profit-maximizing (1) benchmarks.

Final evaluation in the runner uses fresh games, greedy Q actions, and no learner updates; the default is one game per variant per seed, including the all-LLM control. Tail summaries use the final 100 rounds, following the supplied reporter. Independent training seeds are the replication unit; LLM API randomness is not seed-controlled.

4. Results

The central result is sustained collusion across learning architectures. Tolerant Q learned cooperative pricing with the LLM sellers and gained a small late-game profit advantage over its partners. Ordinary Q improved its profit but failed to sustain collusion. Figures 1–2 organize this contrast: learning to earn more and learning to maintain a collusive market are distinct outcomes.

Table 1. Sustained collusion and focal-seller profits after training.

Focal strategy	Sustained collusion	Total profit	Discounted profit
Ordinary Q	Not observed	180	0.90
Tolerant Q	Observed	430	1.90
LLM reference	Observed	450	2.30

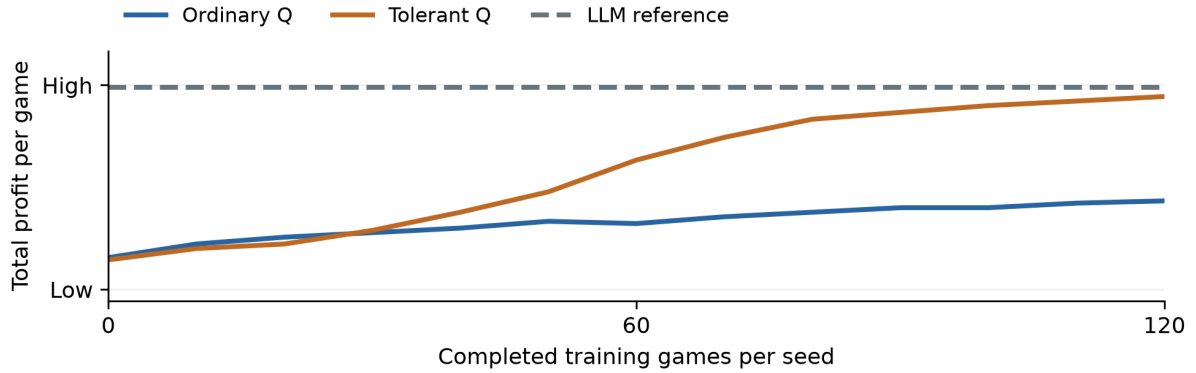


Figure 1. Total focal-seller profit over training. Schematic trajectories for 120 games of 3,000 rounds per seed, comparing ordinary Q, tolerant Q, and the LLM reference.

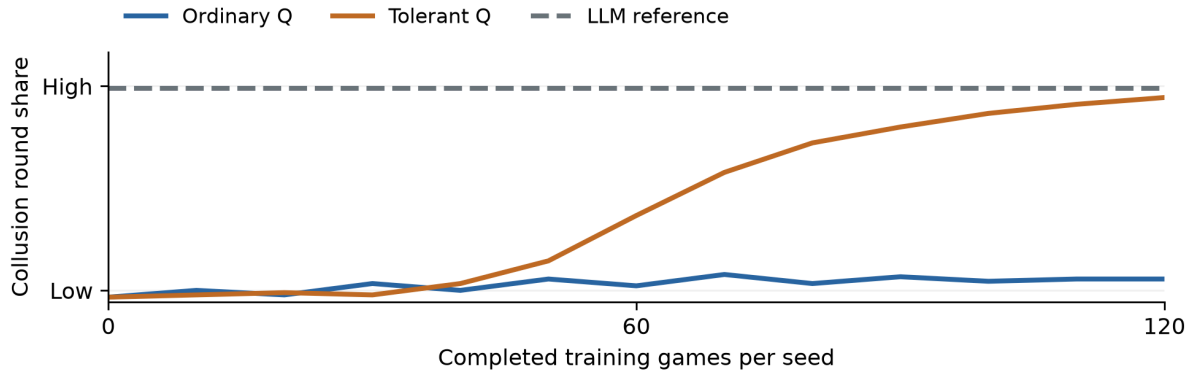


Figure 2. Sustained-collusion round share over training. Schematic trajectories illustrate the ordinary-versus-tolerant contrast over the same training exposure.

5. Discussion and Limitations

Algorithmic diversity can fail as a safeguard because agents adapt. The tolerant learner’s outcome shifts the audit question from which architectures are present to what those architectures learn through repeated interaction. The ordinary-Q comparison is decisive for that framing: the same architectural family contains both a learner that disrupts sustained cooperation and one that joins it.

Asymmetric cooperation adds a second safety concern. The Q seller can profit from a cooperative regime while taking a larger late-game share than its LLM partners. Monitoring equal prices or equal gains would miss that form of participation. Appendix B gives a concrete payoff mechanism through which small pricing differences can support a relative advantage within near-aligned prices.

Limitations

Differences in game length and training exposure prevent attributing the local-to-server change to one factor. This is somewhat troublesome in real-world markets where the length of the game cannot be too long before profit is made, otherwise the seller might go bankrupt. Also, late relative gains do not establish higher full-game profit than the all-LLM reference. broader generalization and infinite-training guarantees remain outside the claim.

Future Work

One idea is to try and come up with different methods of defining tolerance, such that maybe they can achieve profit in shorter game horizons. One could also extend the frozen evaluation across more fresh markets and model families. Then force a brief price deviation and measure punishment and recovery.

6. Conclusion

A tolerant Q learner learned to collude with LLM pricing agents and gained a small late-game advantage over its partners, while ordinary Q failed to sustain collusion. This is a counterexample to treating algorithmic diversity as sufficient protection. Diversity-based safeguards need to survive adaptation; architecture labels alone do not establish that protection.

Code and Data

Code: github.com/aradmaleki02/q-llm-collusion. All LLM responses were gathered through the official OpenAI API and the gpt-6-luna model.

References

- [1] Keppo, J., Li, Y., Tsoukalas, G., and Yuan, N. (2026). On the Fragility of AI Agent Collusion. arXiv:2603.20281, v1. <https://arxiv.org/abs/2603.20281>
- [2] Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
- [3] Matignon, L., Laurent, G. J., and Le Fort-Piat, N. (2007). Hysteretic Q-Learning: An Algorithm for Decentralized Reinforcement Learning in Cooperative Multi-Agent Teams. *IEEE/RSJ International Conference on Intelligent Robots and Systems*. <https://gjlautent.github.io/assets/papers/Matignon2007hysteretic.pdf>

Appendix

A. Why we need longer training horizons

In this setup we followed [1] and showcased the issue with a lower number of games. Each learner completed seven 1,000-round training games and one fresh, frozen 1,000-round development evaluation. Both retained 7,000 environment steps and 31,332 gradient updates; the freeze audit verified unchanged weights and counters. These runs are separate from the final study.

Table A1. 1000 round profits. The LLM comparator is seller 4 in a historical reference, not a matched-seed server control.

Focal strategy	Total profit	Discounted profit ($\gamma = .95$)
Ordinary Q	58.09125	0.66747
Tolerant Q	64.06638	0.87096
Historical LLM seller 4	149.29352	2.30673

Table A2. 1000 round full-market outcomes. The LLM reference’s 43.2% collusion round share uses the sustained-window criterion in Methods.

Market	Price lift	Profit index	Longest alignment
Ordinary Q + 3 LLM	11.12%	14.70%	3 rounds
Tolerant Q + 3 LLM	0.99%	−3.13%	1 round
Historical 4 LLM	28.46%	56.27%	324 rounds

Neither Q evaluation sustained collusion. Tolerant Q out-earned ordinary Q, but both earned less than the historical LLM seller in total and discounted profit.

B. A possible source of unequal gains

Assume identical logit demand parameters and marginal cost c . Three LLM sellers charge p , and Q charges $p - d$, where $0 < d < p - c$. Since demand shares have a common denominator, the Q-to-LLM profit ratio is:

$$\frac{\pi_Q}{\pi_L} = \frac{p - d - c}{p - c} e^{\frac{d}{\mu}}$$

Here $\mu > 0$ is the logit price-sensitivity scale. When $p - c > \mu$, a small undercut raises relative profit: the log-ratio derivative at zero is $1/\mu - 1/(p - c) > 0$. Alignment passes threshold η if $d \leq \eta p / (1 + \eta)$. A suitable grid point is required. This shows relative payoff feasibility, not the learned mechanism, absolute gains, or stable cooperation.

C. Limitations and Dual-Use / Ethical Considerations

Scope. The claims concern tacit pricing behavior in a simulated market. The ordinary-Q negative result, the shorter pilot, and the differences from Keppo et al. are retained to delimit the contribution. A high price or a single positive payoff gap does not establish a stable collusive policy. The late-game observation is descriptive; no known-horizon causal effect has been established.

Dual use. Adaptation mechanisms and pricing traces could help improve real-world collusive agents or conceal unequal participation in a cartel. The intended use is auditing and evaluating safeguards. Release should emphasize the controlled environment, detection metrics, and countermeasure tests, with API credentials and unrelated private notes removed. No agents were deployed to set prices in a real market.

Reproducibility. All results are verifiable by running our code following the instructions on our github project.

D. Implementation specification

Table D1. Effective implementation settings.

Component	Setting
Training exposure	3,000 rounds/game; 120 games/variant/seed; seeds 0, 1, 2
Market	Four sellers; quality 2; outside quality 0; marginal cost 1; logit scale 0.25; market mass 1; unlimited inventory
Actions and seat	51 equally spaced prices from 1 to 3 (step 0.04); focal Q seller 4
LLM policy	gpt-6-luna, low reasoning; full current-game history; 160-word private notes; explicit cache blocks
Observation / network	7 inputs: four normalized prices, own demand, own profit, remaining-time input (zero at $\gamma = 0.95$); GRU 128; concatenate latest input; ReLU layer 128; 51 action values
Objective / update	Own profit; $\gamma = 0.95$; Double DQN; smooth-L1 loss; $\kappa = 1$ ordinary / 0.25 tolerant; Adam lr 0.0003; gradient norm cap 5
Replay / target	Capacity 5,000 transitions; warmup 64; batch 8; four minibatches/fresh transition; target sync every 100 updates; original history prefix per replay item
Exploration	ϵ decreases linearly $0.20 \rightarrow 0.05$ over first 1,000 total training steps; stays 0.05 across later games
Evaluation	Fixed-suite default: one fresh 3,000-round game per variant/seed, including four-LLM reference; greedy Q; no updates; fresh notes/history
Seeds / game boundary	Seed controls Q initialization, exploration and replay, not LLM API sampling. Weights/optimizer/replay persist; $\gamma = 0.95$ bootstrap retained at unannounced cap.

LLM Usage Statement

OpenAI Codex assisted with implementation, literature checks, analysis, drafting, and figure layouts. Final results and text are verified by the authors.