RESTORING CO-ADAPTATION: AN ENDPOINT AUDIT OF ASYMMETRIC TRAINING IN MULTI-AGENT MARKET MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

When an endpoint produced by asymmetric multi-agent training is evaluated with all policies frozen, the evaluation disables the learning responses whose suppression helped shape that endpoint. We distinguish frozen persistence from algorithm- and horizon-relative co-adaptive response stability and introduce a restoration audit: resume all suppressed update channels from the learned endpoint and compare the resulting system-statistic change with a same-budget continuation from a co-adaptive reference. In a controlled three-agent market-making testbed, staged train-and-freeze training increases mean minimum quoted spread (MQS) relative to simultaneous co-adaptation by 0.02854 (95% CI $[0.02658, 0.02974]$; 20/20 seeds positive). After 600 restored episodes, the mean recovery fraction is 0.9953 (95% CI $[0.9922, 0.9981]$), whereas the matched simultaneous continuation drifts by only 6.67×10^{-5} (95% CI $[-9.88 \times 10^{-6}, 1.92 \times 10^{-4}]$). A 2×3 factorial audit shows that the tested spread shift is associated with the interaction of long active-agent blocks and elevated placeholders (primary interaction 0.0273, 95% CI $[0.0244, 0.0296]$; 20/20 seeds positive). The disadvantage follows the first-trained role across identities, and Independent DQN learners reproduce the prespecified block, placeholder, and restoration contrasts in the same testbed. Restoration is not an equilibrium or exploitability test; within this controlled setting, it shows that frozen persistence can overstate stability under the specified resumed-learning dynamics.

1 INTRODUCTION

Training schedules that freeze or stagger agents are useful ways to manage non-stationarity in multi-agent learning (Shoham et al., 2007; He et al., 2016; Foerster et al., 2017; 2018; Lanctot et al., 2017; Nekoei et al., 2023). They also create an evaluation problem. If an endpoint produced under asymmetric adaptation is later evaluated with every policy frozen, the evaluation suppresses the same response channels that were withheld during training. A frozen outcome can therefore be reproducible without being stable to renewed co-adaptation.

We separate these two questions explicitly. *Frozen persistence* means that an outcome is reproducible when learned policies no longer update. *Co-adaptive response stability* asks whether it remains after strategically relevant update channels are re-enabled under a specified learning procedure and continuation horizon. The distinction changes the evidentiary standard: a frozen endpoint is an observation, whereas co-adaptive response stability requires an intervention that restores the suppressed responses.

Relation to strategic robustness. Game-theoretic evaluation usually asks a different question: whether a player can improve by deviating unilaterally while opponents remain fixed. Best-response and exploitability methods operationalize this idea through deviation incentives, learned response oracles, or NashConv-style quantities (Lanctot et al., 2017; Timbers et al., 2022; Bighashdel et al., 2024). Our notion is deliberately algorithm- and horizon-relative. Restoration keeps the learning rule and objectives fixed, re-opens the update channels suppressed by the training curriculum for all learners, and asks whether a prespecified system-level outcome persists. Exploitability and restoration

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

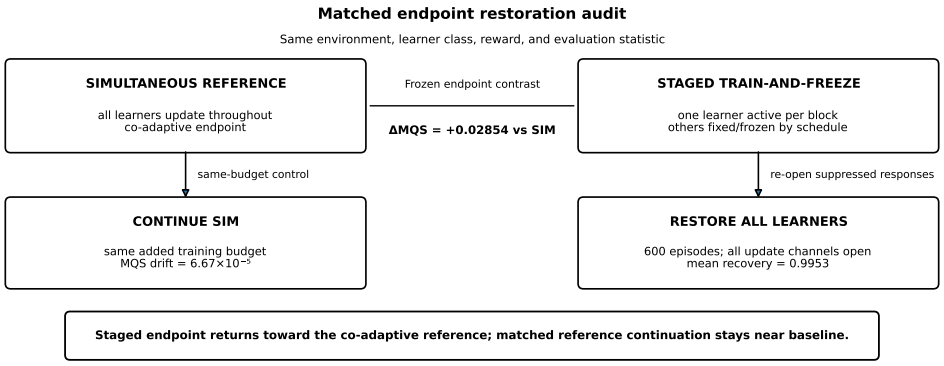

Figure 1: **Matched restoration audit.** The staged curriculum produces a frozen endpoint separated from the simultaneous co-adaptive reference. From these endpoints, the audit compares two equal-budget continuations: restore all learners from the staged endpoint versus continue simultaneous learning from the reference. The staged MQS gap nearly closes after restoration, while the matched reference continuation stays near baseline. The schematic reports confirmatory estimates but does not depict an empirical trajectory.

are therefore complementary: the former probes unilateral deviation against fixed opponents; the latter probes sensitivity of a learned endpoint to renewed joint learning.

We study this evaluation question in a deliberately small three-agent market-making environment, where spread competition makes adaptation asymmetries visible at both agent and system levels while preserving exact control over learner identity, update schedule, and opponent behavior. Three matched learners choose from the same eight quoted spreads; the market-level outcome is mean minimum quoted spread (MQS), the minimum contemporaneous quote across agents averaged over all greedy evaluation timesteps. The simulator is a strategic testbed, not a calibrated limit-order-book model.

The evidence is organized around the endpoint intervention. A 20-seed confirmatory comparison first establishes a staged-vs-simultaneous MQS shift. Restoring all update channels then reverses the staged endpoint, whereas an equal-budget continuation of the simultaneous endpoint stays near baseline. A factorial audit identifies the tested block$\times$placeholder interaction; identity permutations separate the first-trained role from fixed identity; and an Independent-DQN replication on a disjoint seed set reproduces the prespecified block, placeholder, and restoration contrasts within the same testbed. A reward-specification check identifies the regime in which active tight-spread participation is non-dominated; all confirmatory analyses use $\lambda = 10^{-4}$ within that regime.

Contributions. (i) We distinguish frozen persistence, unilateral strategic robustness, and algorithm- and horizon-relative co-adaptive response stability, and formalize restoration as an endpoint audit paired with a same-budget continuation control. (ii) In a controlled market-making testbed, the staged MQS shift reverses under restored co-adaptation but not matched reference continuation; a factorial audit shows a positive block$\times$placeholder interaction for elevated placeholders while the tight-placeholder block contrast is near zero, and role permutations separate training role from fixed identity. (iii) Independent DQN learners reproduce the prespecified block, placeholder, and restoration contrasts under the same environment and intervention, supporting learner-class transfer within the testbed without claiming equilibrium, cross-domain generality, or universal failure of sequential learning.

2 Related Work and Novelty Boundary

Asymmetric and sequential multi-agent learning. Independent learners face non-stationarity because other policies change the transition and reward process they experience (Tan, 1993; Hernandez-Leal et al., 2017; Lowe et al., 2017). Prior work mitigates this coupling through opponent modeling

(He et al., 2016), stabilized replay (Foerster et al., 2017), and structured update schedules. In cooperative MARL, HAPPO/HATRPO derive monotonic-improvement results from sequential policy updates (Kuba et al., 2022); A2PO further studies agent-by-agent optimization and update order (Wang et al., 2023a); and multi-timescale learning uses sequential best-response-style updates to reduce non-stationarity (Nekoei et al., 2023). These works show that asymmetric or sequential updating can be a useful training mechanism. They optimize how policies are learned; they do not ask whether a system-level outcome observed at a resulting frozen endpoint survives when suppressed co-adaptation is jointly resumed. Our contribution is an endpoint evaluation protocol, not a sequential optimizer or a claim that sequential updates are intrinsically harmful.

Opponent-aware learning and learning dynamics. A second line of work treats other agents' learning dynamics as part of the strategic problem. LOLA and Stable Opponent Shaping account for how an agent's update affects subsequent opponent learning (Foerster et al., 2018; Letcher et al., 2019); M-FOS learns long-horizon opponent-shaping meta-policies (Lu et al., 2022); COLA studies consistency under opponent-learning awareness (Willi et al., 2022); LOQA incorporates an opponent's Q-learning response into decentralized optimization (Aghajohari et al., 2024); and Shaper scales opponent shaping to higher-dimensional games (Khan et al., 2024). Recent theory also studies how an optimizer can anticipate a learner's future behavior in repeated games (Assos et al., 2024), while game-dynamics analyses characterize how interacting updates shape multi-player learning (Balduzzi et al., 2018). These methods establish that co-learner dynamics are strategically consequential. Restoration differs in objective and timing: it does not train an agent to shape or anticipate another learner; it is a post-hoc matched intervention that resumes the original suppressed learning channels from an already learned endpoint and measures persistence of a prespecified system statistic.

Strategic robustness and evaluation diagnostics. Response-based evaluation asks whether a learned strategy remains effective when counterpart behavior changes. PSRO iteratively trains approximate best responses to policy mixtures (Lanctot et al., 2017), approximate-exploitability methods learn unilateral best responses when exact computation is intractable (Timbers et al., 2022), and recent surveys organize response-oracle methods around strategy exploration and game-theoretic solution quality (Bighashdel et al., 2024). Complementary diagnostic work stresses that evaluation itself can hide strategically relevant failure modes: MADRID searches for diverse adversarial scenarios exposing weaknesses of pretrained multi-agent policies (Samvelyan et al., 2024), while ZSC-Eval constructs partner sets and metrics to reduce mismatch between evaluation and deployment partners in zero-shot coordination (Wang et al., 2024). Our target is narrower. We neither estimate NashConv nor search for an adversarial or unseen partner distribution. Starting from an endpoint created under suppressed adaptation, we resume all suppressed learners under the specified original learning rule and compare the system-statistic response with a same-budget continuation from a co-adaptive reference. Thus restoration complements unilateral, adversarial, and partner-distribution diagnostics with a finite-horizon test of endpoint dependence on the training path.

Learning market makers and closest financial precedents. Classical market making studies inventory-sensitive stochastic control (Avellaneda & Stoikov, 2008; Cartea et al., 2015); reinforcement-learning formulations introduced data-driven policies and reward design (Spooner et al., 2018). Recent work broadens strategic structure: Wang et al. (2023b) allow non-quoting/single-side actions, while Vadori et al. (2024) study interacting dealers in a multi-agent simulation. Wang et al. (2025) train heterogeneous later market makers against a pre-trained fixed market maker, and Cheridito et al. (2025) include single-agent retraining against fixed learned counterparties in a richer ABIDES environment. These works show that adaptation scope and fixed counterparts matter in financial multi-agent systems, but they do not perform the matched restore-all endpoint test studied here. Separately, algorithmic-pricing and dealer-market studies examine supra-competitive or collusive outcomes (Calvano et al., 2020; Cont & Xiong, 2024); our simulator is not designed to identify collusion or equilibrium selection. We therefore claim novelty only for the matched endpoint-audit chain, not for sequential learning, fixed-opponent training, later-agent advantage, opponent awareness, or the generic fact that agents respond strategically. Appendix C gives a compact claim-boundary table.

3 Endpoint Evaluation Problem and Controlled Setting

3.1 Frozen persistence versus co-adaptive response stability

Let a training protocol P map initialization and randomness ω to a joint policy $\boldsymbol{\pi}_{P,\omega}$, and let $M(\boldsymbol{\pi})$ be a system statistic measured with policies frozen. The endpoint contrast

$$\Delta_{\text{path}}(\omega) = M(\boldsymbol{\pi}_{\text{SEQ},\omega}) - M(\boldsymbol{\pi}_{\text{SIM},\omega}) \tag{1}$$

shows that two training paths end differently, but not whether the difference survives renewed learning. Let $\mathcal{R}_h^{\mathcal{A}}(\boldsymbol{\pi})$ denote joint policy after h episodes of resumed mutual adaptation under learning procedure $\mathcal{A}$. Define

$$D_h^{\mathcal{A}}(\boldsymbol{\pi}; M) = M\big(\mathcal{R}_h^{\mathcal{A}}(\boldsymbol{\pi})\big) - M(\boldsymbol{\pi}), \tag{2}$$

and, for the staged endpoint,

$$\Delta_{\text{restore}}(\omega) = M(\boldsymbol{\pi}_{\text{SEQ},\omega}) - M\big(\mathcal{R}_h^{\mathcal{A}}(\boldsymbol{\pi}_{\text{SEQ},\omega})\big)\,. \tag{3}$$

The negative control applies the same added simultaneous-training budget to the reference endpoint. With continuation operator $\mathcal{C}_h$,

$$\Delta_{\text{ctrl}}(\omega) = M(\mathcal{C}_h(\boldsymbol{\pi}_{\text{SIM},\omega})) - M(\boldsymbol{\pi}_{\text{SIM},\omega}). \tag{4}$$

For shifted conditions, we additionally report the descriptive recovery fraction $\Delta_{\text{restore}}/\Delta_{\text{path}}$. This normalization is omitted when the initial path gap is near zero because the denominator is then ill-conditioned.

We use *co-adaptive response stability* only relative to $\mathcal{A}$, horizon h, and statistic M. It is not a synonym for Nash equilibrium, low exploitability, or robustness to arbitrary opponent responses. Frozen evaluation estimates $M(\boldsymbol{\pi})$ but, because policies do not update, cannot estimate $D_h^{\mathcal{A}}$. Conversely, best-response/exploitability analysis fixes opponents and optimizes a unilateral deviation; restoration resumes all suppressed learning channels. Neither subsumes the other.

The staged protocol contains multiple ingredients, so we separately audit the block×placeholder interaction. Let $M(B, P)$ be MQS after a staged curriculum with active-agent block length B and fixed not-yet-trained placeholder P. For placeholder $P4$, the primary interaction is

$$I_{P4} = \{M(3000, P4) - M(1000, P4)\} - \{M(3000, P0) - M(1000, P0)\}. \tag{5}$$

A positive I_{P4} tests whether long blocks have a larger effect under an elevated placeholder than under the tight placeholder; it does not decompose every primitive cause of the curriculum.

3.2 Market-making testbed and outcome

Three market makers interact with one customer per step. Each agent selects a full quoted spread

$$\mathcal{D} = \text{linspace}(0.02, 0.20, 8). \tag{6}$$

The smallest spread wins, with random tie-breaking. Conditional on being best, spread d executes only if $d \le r$ for $r \sim \text{Uniform}(0, 0.26)$; trade direction is an independent fair coin flip. Learners observe inventory clipped to $[-15, 15]$ (the underlying inventory and its penalty are not clipped), and inventory resets to zero at the start of each 200-step episode. Reward includes post-trade inventory penalty λq^2; the primary confirmatory setting is $\lambda = 10^{-4}$, selected after the reward-validity audit in Section 4.

At greedy evaluation timestep t, let $s_t^{(a)}$ be agent a's quote and $m_t = \min_a s_t^{(a)}$. The market statistic is

$$\text{MQS} = \frac{1}{T}\sum_{t=1}^{T} m_t, \qquad T = 60{,}000. \tag{7}$$

MQS includes no-trade timesteps; it is a quote-side minimum, not transaction-based effective spread. Agent win rate and zero-inventory spread are secondary diagnostics.

Table 1: **Intervention logic.** Restoration and continued-SIM use the same added simultaneous-update budget but start from different endpoints.

Condition	Start	Adaptation enabled	Scientific role
SIM	fresh learners	all agents throughout	co-adaptive reference
Staged	fresh learners	one active learner per block	schedule treatment; frozen endpoint
RESTORE	staged endpoint	all agents	tests co-adaptive response stability
Continued SIM	SIM endpoint	all agents	controls for extra training budget
Factorial staged	fresh learners	one active; vary B, P	tests block$\times$placeholder interaction

3.3 TRAINING CURRICULA

The tabular study uses matched independent Q-learners (Watkins & Dayan, 1992; Sutton & Barto, 2018) with $\alpha = 0.15$, $\gamma = 0.9$, initial $\epsilon = 0.3$, and multiplicative decay 0.99999 after each active action. Each learner receives 3,000 episodes (600,000 updates). Under SIM, all three update concurrently for 3,000 episodes. Under the staged curriculum, exactly one learner updates per stage: not-yet-trained roles use a fixed placeholder; already-trained roles act greedily from persistent Q-tables and are frozen. Per-agent and total Q-update counts therefore match SIM, while temporal counterpart exposure and environment interactions differ by design.

The factorial audit crosses $B \in \{1000, 3000\}$ with $P \in \{P0, P4, P7\}$, corresponding to spreads $0.02, 0.122857, 0.20$. Restoration resumes simultaneous co-adaptation from the long-block staged endpoint, with exploration reset to $\epsilon = 0.05$ for all learners; reference continuation applies the same operator to the SIM endpoint for the same added budget. Identity permutations change which agent occupies the first-trained role while preserving schedule semantics.

What is matched and intentionally changed. Across the primary SIM and staged comparison we hold fixed learner class, state/action spaces, reward, hyperparameters, initialization distribution, and per-agent/total Q-update counts. The staged protocol intentionally changes the temporal distribution of counterpart policies and therefore the trajectories on which those updates are collected; it also requires more serial environment interaction. These exposure differences are the training-path treatment. Accordingly, Δ_{path} is an endpoint difference between curricula, not an isolated "update-order-only" causal effect.

For learner-class transfer, Independent DQN (Mnih et al., 2015) replaces the Q-table while holding environment, state, action grid, reward, MQS, and intervention structure fixed. Each DQN maps normalized inventory through two 64-unit ReLU layers to eight action values and uses Adam (10^{-3}), replay buffer 10^5, batch size 64, Huber loss, target update every 1,000 gradient steps, gradient clipping at 10, and the same $\gamma = 0.9$ and multiplicative ϵ decay, with an ϵ floor of 0.01; DQN restoration and continuation keep the decayed ϵ rather than resetting it. This is a function-approximation transfer test inside the same testbed, not arbitrary neural-MARL generalization.

4 INTERVENTION DESIGN AND INFERENCE

We organize experiments by scientific questions. **RQ1: Does staged training create a reversible endpoint?** The tabular confirmation uses 20 confirmatory seeds (101–120), disjoint from pilots, and pairs SEQ, SIM, restoration, reference continuation, and role permutations by seed. **RQ2: Which staged ingredients are necessary for the spread shift?** A disjoint 20-seed factorial audit (301–320) crosses the two block lengths and three placeholders and evaluates Equation 5 plus prespecified simple contrasts. **RQ3: Do the prespecified contrasts reproduce beyond a Q-table?** A DQN audit uses a third disjoint seed set (401–420) and tests block, placeholder, and restoration contrasts together with continued-SIM drift.

Interpretation criteria. The interpretation requires a conjunction: the staged endpoint must differ from the co-adaptive reference; restoration must move it toward the reference; and the same added simultaneous-training budget must not induce comparable movement from the reference endpoint. Role permutation asks whether disadvantage follows schedule position rather than fixed identity, and the factorial audit asks which tested ingredients are necessary. If the initial path gap is absent, the recovery ratio is not informative; if reference continuation drifts comparably, reversal cannot be attributed specifically to re-opened responses.

Reward-specification check. The inventory-penalty coefficient changes the strategic incentives of the simulator. At $\lambda = 10^{-3}$, abstention is preferable to active tight-spread participation; recent market-making work likewise treats non-quoting as a meaningful policy action (Wang et al., 2023b). We therefore evaluate $\lambda \in \{0, 10^{-5}, 3\times10^{-5}, 5\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}\}$ on five pilot seeds disjoint from all confirmatory seed sets. The tightest action is empirically non-dominated with positive reward for $\lambda \leq 10^{-4}$, whereas abstention dominates for $\lambda \geq 3 \times 10^{-4}$. All confirmatory analyses use $\lambda = 10^{-4}$. This check validates the intended active-competition regime of the simulator; it does not calibrate preferences to a real market.

Inference. The independent unit is the training seed, never an evaluation timestep. Continuous paired effects use 10,000 seed-level bootstrap resamples for 95% percentile intervals (Efron & Tibshirani, 1994). Two-sided Monte Carlo sign-flip tests with 10,000 random assignments are secondary; if no random assignment is at least as extreme as observed, we report the finite-resolution plus-one estimate $1/10001$. Binary role-support rates use exact Clopper–Pearson intervals (Clopper & Pearson, 1934); because the same 20 seeds are reused across role orderings, the descriptive 60/60 aggregate is not treated as 60 independent Bernoulli trials. No hyperparameter tuning follows inspection of confirmatory outcomes.

5 RESULTS

5.1 RQ1: THE STAGED ENDPOINT IS SYSTEM-LEVEL AND REVERSIBLE

The 20-seed confirmation establishes a system-level path effect: SEQ increases MQS relative to SIM on all 20 confirmatory seeds, with mean paired difference 0.02854 and 95% bootstrap CI $[0.02658, 0.02974]$ (median 0.02928; range $[0.01194, 0.03101]$). The first-trained agent also quotes more widely at zero inventory and loses order flow, but market-level MQS is primary.

Restoration changes the interpretation. After 600 additional episodes of simultaneous co-adaptation from the staged endpoint, the mean recovery fraction is 0.9953 (95% CI $[0.9922, 0.9981]$; range $[0.9762, 1.0067]$). Small values above one are overshoots, not evidence of a unique equilibrium. Applying the same added simultaneous-training budget to the SIM endpoint yields mean MQS drift 6.67×10^{-5} with CI $[-9.88 \times 10^{-6}, 1.92 \times 10^{-4}]$. **Takeaway:** the staged spread gap persists when policies are frozen but fails the specified co-adaptive restoration test over 600 episodes; matched continued-SIM rules out generic movement from merely adding the same simultaneous-training budget.

Role permutation argues against a fixed-identity explanation at the outcome level. When agent 0, 1, or 2 is assigned the first-trained role, that role has the lowest evaluation win rate in 20/20 seeds for each ordering (exact 95% CI $[0.832, 1]$ within each ordering). Widest zero-inventory spread follows the role in 19/20, 19/20, and 17/20 seeds, respectively.

5.2 RQ2: BLOCK LENGTH INTERACTS WITH PLACEHOLDER LEVEL IN THE TESTED GRID

At $B = 1000$, mean MQS is essentially at the low-spread reference for all three placeholders: 0.02006 (P0), 0.02004 (P4), and 0.02006 (P7). At $B = 3000$, P0 remains 0.02006, whereas P4 rises to 0.04733 and P7 to 0.04886. The primary interaction I_{P4} is 0.027288 with 95% paired-seed CI $[0.024392, 0.029555]$ and is positive on all 20 seeds; the sign-flip estimate is $1/10001$. The analogous P7 interaction is 0.028795 with CI $[0.026207, 0.031061]$, again 20/20 positive. Figure 2A reports these inferential contrasts directly rather than connecting the two discrete block-length levels.

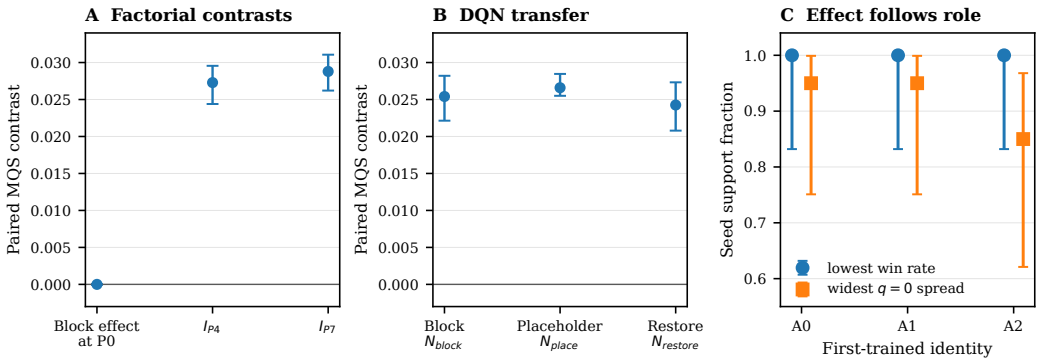

Figure 2: **Where the staged effect appears and whether it reproduces. A:** In the tabular 2×3 factorial audit, the block effect is near zero under P0, while the prespecified P4 and P7 block×placeholder interactions are positive; points show estimates with 95% paired-seed bootstrap CIs. **B:** Independent DQN learners reproduce the prespecified block, placeholder, and restoration contrasts (20/20 positive for each); bars show 95% paired-seed bootstrap CIs. **C:** Across tabular identity permutations, the first-trained role has the lowest win rate in 20/20 seeds for every identity; bars show exact 95% binomial CIs.

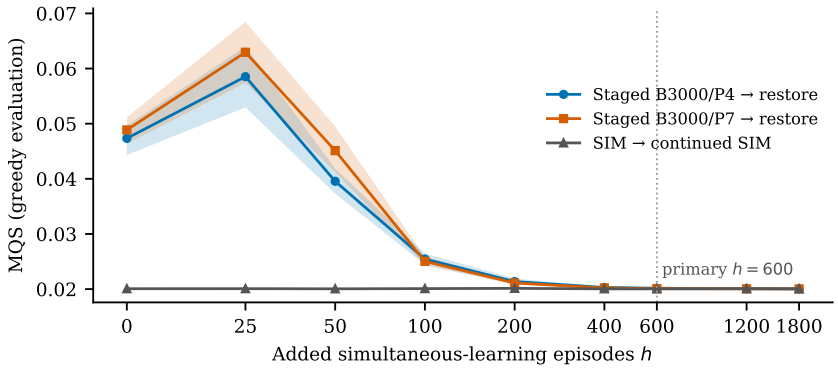

Figure 3: **Restoration remains closed beyond the primary horizon.** Mean MQS over 20 factorial seeds with 95% seed-level bootstrap bands. The staged gap widens transiently at very short horizons, closes by about $h = 200$, and stays closed through $h = 1{,}800$; the continued-SIM reference remains near baseline. This post-hoc descriptive trajectory does not create a new confirmatory claim.

The simple contrasts sharpen this interaction. The long-block effect under P0 is centered at zero (mean 0.00000, CI $[-0.00007, 0.00006]$), whereas under P4 it is 0.02729 (CI $[0.02441, 0.02954]$) and under P7 0.02880 (CI $[0.02621, 0.03107]$). At $B = 3000$, replacing P0 by P4 increases MQS by 0.02727 (CI $[0.02437, 0.02953]$). Restoration is nearly complete for shifted P4/P7 cells; the ratio is poorly conditioned for P0 because its staged gap is already near zero. A post-hoc trajectory check from the same saved P4/P7 staged endpoints (Figure 3; Appendix D) indicates that $h = 600$ does not sit on a still-moving part of the curve: mean recovery exceeds 0.95 by $h = 200$ and remains at 1.000 (P4) and 1.001 (P7) through $h = 1{,}800$, three times the primary horizon, while mean continued-SIM drift stays below 7×10^{-5} in absolute value at every checkpoint. The path is non-monotone: after 25 restored episodes, MQS is temporarily wider than at the staged endpoint. **Takeaway:** the evidence supports an interaction between sustained one-at-a-time exposure and elevated placeholder behavior; it does not support a generic claim that longer blocks alone widen spreads.

5.3 RQ3: THE PRESPECIFIED CONTRASTS REPRODUCE WITH DQN IN THE SAME TESTBED

The DQN audit reproduces all three prespecified contrasts on 20 new seeds. The long-block contrast $N_{\text{block}} = M(B3000, P4) - M(B1000, P4)$ has mean 0.02541 and CI $[0.02214, 0.02821]$; all 20 seeds are positive. The placeholder contrast $N_{\text{place}} = M(B3000, P4) - M(B3000, P0)$ has mean 0.02659 and CI $[0.02550, 0.02846]$, again 20/20 positive. Finally, $N_{\text{restore}} = M(B3000, P4) - M(\text{RESTORE}(B3000, P4))$ has mean 0.02425 and CI $[0.02079, 0.02732]$, also 20/20 positive (Figure 2B). Continued-SIM drift is zero at reported precision for all 20 seeds; no numerical failures are recorded and a deterministic rerun check passes. **Takeaway:** the three prespecified contrasts reproduce when the Q-table is replaced by nonlinear value-function approximation under the same environment and restoration operator; this supports learner-class transfer within the testbed, not cross-domain or arbitrary-MARL generalization.

Across the three questions, the evidence supports a bounded ladder: the specified staged curriculum changes the market endpoint; the resulting gap reverses under the specified restored co-adaptation but not matched reference continuation; the factorial audit localizes the shift to the tested block×placeholder interaction; role permutations separate schedule role from fixed identity; and the DQN study reproduces the prespecified contrasts across value-function representations within the same testbed. None of these results establishes equilibrium, arbitrary-response robustness, universal sequential-training harm, or real-market calibration.

6 DISCUSSION

What is identified. The matched continuation control shows that restoration is not merely "more training" at the market-level outcome, role permutations separate schedule role from agent identity, and the factorial audit shows that the staged effect is conditional on long blocks together with elevated placeholders. Together these interventions support a training-path-dependent interpretation of the tested endpoint. They do not identify a universal primitive such as "freezing causes wide spreads": the staged treatment also changes counterpart exposure through time, and the factorial study identifies a block×placeholder interaction within the tested grid rather than a complete causal decomposition.

Restoration and exploitability answer different questions. Best-response/exploitability analysis fixes an opponent profile and asks whether a unilateral responder can improve its utility, connecting directly to equilibrium deviation incentives (Lanctot et al., 2017; Timbers et al., 2022). Our restoration audit fixes the learning rule and environment, re-opens all response channels suppressed by the curriculum, and measures the change in a pre-specified system statistic. The present experiments therefore do not estimate NashConv and should not be read as an equilibrium test. Conversely, restoration adds information frozen evaluation cannot provide: the wide-spread endpoint is highly sensitive to renewed mutual learning even though it is reproducible when every policy is held fixed.

When the audit is applicable. The audit is conceptually applicable when three conditions hold: (i) training intentionally suppresses one or more strategically relevant update channels, (ii) the scientific claim is based on an endpoint evaluated with those policies frozen, and (iii) the suppressed learning dynamics can be resumed without changing the environment, reward, objective, or evaluation statistic. Staged opponent curricula, train-and-freeze protocols, and some population-learning pipelines can satisfy these conditions. This statement concerns applicability of the diagnostic, not empirical generalization: the present evidence validates it only in the market-making testbed studied here.

Why a small testbed is useful. A richer simulator can improve external realism while making intervention attribution harder because schedule changes interact with function approximation, role heterogeneity, replay, order-book state, and population composition. The small environment is used for the opposite purpose: matched learners and explicit schedule manipulation make the response channels auditable. The DQN replication reduces one obvious dependence on tabular representation while remaining intentionally inside the same environment. Richer financial MARL systems already study strategic adaptation and endogenous market outcomes (Vadori et al., 2024; Cheridito et al., 2025; Wang et al., 2025); the contribution here is controlled attribution of a restore-all endpoint diagnostic.

Economic interpretation. Persistently wide learned quotes can invite stronger labels. In pricing games, Q-learning can sustain supra-competitive prices (Calvano et al., 2020); in dealer-market models, learning dynamics can be compared against explicit competitive and collusive benchmarks (Cont & Xiong, 2024). Our evidence supports neither inference. The wide-spread endpoint largely disappears when reciprocal updating resumes, so in this testbed the protocol-induced response constraint is a warning against reading the frozen outcome as a stable market property. This does not contradict collusion results from models designed to identify them.

7 LIMITATIONS

The environment is stylized: three market makers, one-dimensional bounded inventory, discrete spreads, synthetic customer reservation prices, and no queue, latency, adverse-selection, or endogenous-price dynamics. We make no claim that numerical MQS effects calibrate to a real exchange. Tabular and DQN learners share observation, action grid, reward, and testbed; the DQN replication therefore supports learner-class transfer within this testbed, not broad neural-MARL generalization. The factorial audit does not exhaust curricula, opponent policies, or update-order designs. Restoration is finite-horizon; extending it to 1,800 episodes shows no re-divergence in the tested P4/P7 cells (Appendix D), but this cannot establish asymptotic convergence or equilibrium uniqueness. Response stability here is algorithm-relative: another learning rule, a unilateral best response, or an adversarial response oracle could reveal different behavior, and we do not estimate those quantities. The reward-validity sweep rules out a specific abstention degeneracy but does not make the reward economically canonical. Role permutations address fixed identity but not every implementation confound.

8 CONCLUSION

A multi-agent endpoint can be reproducible while frozen yet sensitive to renewed learning responses. In our controlled market-making testbed, a staged train-and-freeze curriculum produces a system-level spread shift that reverses when all learners resume the specified co-adaptation dynamics, while a same-budget simultaneous continuation does not move comparably. A factorial audit identifies the tested block$\times$placeholder interaction, the disadvantage follows the first-trained role across identities, and the prespecified block, placeholder, and restoration contrasts reproduce with DQN learners on a disjoint seed set. The contribution is not a new equilibrium concept or the claim that sequential learning is generally harmful. It is a complementary endpoint-evaluation principle: unilateral best-response tests ask whether fixed policies are exploitable, whereas restoration asks whether a path-induced system outcome survives renewed joint learning under the specified algorithm and horizon. For endpoints created by asymmetric training, frozen persistence alone is therefore insufficient evidence of co-adaptive response stability.

AI USE STATEMENT

Generative AI tools were used solely to improve the clarity, grammar, and readability of author-written text and to assist with the execution and orchestration of experiments under protocols, configurations, and evaluation criteria specified by the human authors. All research questions, hypotheses, core ideas, conceptual contributions, methodological choices, experimental designs, interpretations, and scientific conclusions were conceived and determined by the human authors. All AI-assisted experimental executions, outputs, and manuscript edits were subsequently reviewed and verified by the human authors, who take full responsibility for the correctness, integrity, and final content of the submission.

ETHICS STATEMENT

This work uses only a synthetic market-making simulator; it involves no human subjects, personal data, or real trading data, and no trading system was deployed. The simulator is not calibrated to any real market, and its results should not be used to assess or design real trading strategies. Because learned pricing agents can raise concerns about supra-competitive or collusive outcomes, we state explicitly that our simulator is not designed to identify collusion and that our findings neither establish nor rule out such outcomes in real markets.

Reproducibility statement

The main text specifies the market statistic, action grid, training semantics, restoration operator, factorial contrasts, seed-level inference, and DQN replication configuration. Appendix A records compact tabular and neural settings; Appendix B gives principal confirmatory statistics and exact role-support counts; Appendix C records the expanded prior-art boundary; Appendix D reports a post-hoc restoration-horizon check. The independent resampling unit is the training seed. Figures 1–3 summarize the reported intervention logic and outcomes. The anonymized supplementary material contains the environment and training code, raw per-seed outputs for the confirmatory and horizon studies, the analysis scripts that produce all reported statistics, and a README that maps each reported number to its source file.

References

Milad Aghajohari, Juan Duque, Timotheus Cooijmans, and Aaron Courville. LOQA: Learning with opponent q-learning awareness. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/402542c2341e5d2eadc1dd0891275901-Abstract-Conference.html.

Angelos Assos, Yuval Dagan, and Constantinos Daskalakis. Maximizing utility in multi-agent environments by anticipating the behavior of other learners. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4476dd7320e0eba63961990d73525064-Abstract-Conference.html.

Marco Avellaneda and Sasha Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2008.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 354–363. PMLR, 2018. URL https://proceedings.mlr.press/v80/balduzzi18a.html.

Ariyan Bighashdel, Yongzhao Wang, Stephen McAleer, Rahul Savani, and Frans A. Oliehoek. Policy space response oracles: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7951–7961, 2024. doi: 10.24963/ijcai.2024/880.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020. doi: 10.1257/aer.20190623.

Álvaro Cartea, Sebastian Jaimungal, and José Penalva. *Algorithmic and High-Frequency Trading*. Cambridge University Press, 2015.

Patrick Cheridito, Jean-Loup Dupret, and Zhexin Wu. Abides-marl: A multi-agent reinforcement learning environment for endogenous price formation and execution in a limit order book. *arXiv preprint arXiv:2511.02016*, 2025. URL https://arxiv.org/abs/2511.02016. Version 2, August 2026.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Rama Cont and Wei Xiong. Dynamics of market making algorithms in dealer markets: Learning and tacit collusion. *Mathematical Finance*, 34(2):467–521, 2024. doi: 10.1111/mafi.12401.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Jakob N. Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip H. S. Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1146–1155, 2017.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130, 2018.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1804–1813. PMLR, 2016. URL https://proceedings.mlr.press/v48/he16.html.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

Akbir Khan, Timon Willi, Newton Kwan, Andrea Tacchetti, Chris Lu, Edward Grefenstette, Tim Rocktäschel, and Jakob Foerster. Scaling opponent shaping to high dimensional games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 2024. URL https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p1001.pdf.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EcGGFkNTxdJ.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pp. 4190–4203, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3323fe11e9595c09af38fe67567a9394-Abstract.html.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1811.08469.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems 30*, 2017.

Christopher Lu, Timon Willi, Christian A. Schroeder De Witt, and Jakob Foerster. Model-free opponent shaping. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14398–14411. PMLR, 2022. URL https://proceedings.mlr.press/v162/lu22d.html.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Hadi Nekoei, Akilesh Badrinaaraayanan, Amit Sinha, Mohammad Amini, Janarthanan Rajendran, Aditya Mahajan, and Sarath Chandar. Dealing with non-stationarity in decentralized cooperative multi-agent deep reinforcement learning via multi-timescale learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 376–398, 2023.

Mikayel Samvelyan, Davide Paglieri, Mingi Jiang, Jack Parker-Holder, and Tim Rocktäschel. Multi-agent diagnostics for robustness via illuminated diversity. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 2024. URL https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p1630.pdf.

Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

Thomas Spooner, John Fearnley, Rahul Savani, and Andreas Koukorinis. Market making via reinforcement learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 434–443. IFAAMAS, 2018. URL https://ifaamas.csc.liv.ac.uk/Proceedings/aamas2018/pdfs/p434.pdf.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, 1993.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3487–3493, 2022. doi: 10.24963/ijcai.2022/484.

Nelson Vadori, Leo Ardon, Sumitra Ganesh, Thomas Spooner, Selim Amrouni, Jared Vann, Mengda Xu, Zeyu Zheng, Tucker Balch, and Manuela Veloso. Towards multi-agent reinforcement learning-driven over-the-counter market simulations. *Mathematical Finance*, 34(2):262–347, 2024. doi: 10.1111/mafi.12416.

Xihuai Wang, Zheng Tian, Ziyu Wan, Ying Wen, Jun Wang, and Weinan Zhang. Order matters: Agent-by-agent policy optimization. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=Q-neeWNVv1.

Xihuai Wang, Shao Zhang, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. ZSC-Eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1501. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54a7139c548c88e288aa0fcd2bcbeceb-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Ziyi Wang, Carmine Ventre, and Maria Polukarov. Robust market making: To quote, or not to quote. In *Proceedings of the 4th ACM International Conference on AI in Finance*, pp. 664–672. ACM, 2023b. doi: 10.1145/3604237.3626858.

Ziyi Wang, Carmine Ventre, and Maria Polukarov. Multi-agent reinforcement learning for market making: Competition without collusion. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 814–822, 2025. doi: 10.1145/3768292.3770388.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

Timon Willi, Alistair H. P. Letcher, Johannes Treutlein, and Jakob Foerster. COLA: Consistent learning with opponent-learning awareness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23804–23831. PMLR, 2022. URL https://proceedings.mlr.press/v162/willi22a.html.

A Environment, Training, and Neural Details

Table 2: Core reproducibility settings.

Setting	Value
Agents/state	3 independent learners; observed inventory clipped to $[-15, 15]$; reset to 0
Actions	$\mathrm{linspace}(0.02, 0.20, 8)$; full quoted spread
Episode/evaluation	200 steps; greedy 300×200; no updates
Tabular learner	$\alpha = .15$, $\gamma = .9$, $\epsilon_0 = .3$, decay .99999 per active action
Primary reward	spread revenue minus post-trade inventory penalty $10^{-4}q^2$
Staged semantics	one active learner; future fixed at placeholder; past frozen greedy
Per-agent training	3,000 episodes / 600,000 active Q-updates
Restoration/control	600 additional simultaneous episodes; tabular ϵ reset to 0.05
Factorial audit	$B \in \{1000, 3000\}$; P0/P4/P7 = $0.02/0.122857/0.20$
DQN network	normalized inventory $\to 64 \to 64 \to 8$, ReLU
DQN optimization	Adam 10^{-3}; replay 10^5; batch 64; Huber; target every 1,000; ϵ floor 0.01
Confirmatory seed sets	tabular 101–120; factorial 301–320; DQN 401–420
Inference	10,000 paired bootstrap; 10,000 sign flips; exact binomial intervals

B Principal Confirmatory Statistics

Table 3: Headline paired effects. All uncertainty is at seed level.

Contrast / diagnostic	Mean	95% CI	Seed support
Tabular SEQ$-$SIM MQS	0.02854	[0.02658, 0.02974]	$20/20 > 0$
Tabular restoration recovery fraction	0.9953	[0.9922, 0.9981]	$20/20 > 0$
P4 block$\times$placeholder interaction	0.027288	[0.024392,0.029555]	$20/20 > 0$
P7 block$\times$placeholder interaction	0.028795	[0.026207,0.031061]	$20/20 > 0$
DQN N_{block}	0.02541	[0.02214,0.02821]	$20/20 > 0$
DQN N_{place}	0.02659	[0.02550,0.02846]	$20/20 > 0$
DQN N_{restore}	0.02425	[0.02079,0.02732]	$20/20 > 0$

Table 4: Role-permutation support with exact 95% Clopper–Pearson intervals. The same seed set is reused across orderings, so 60/60 is descriptive rather than 60 independent trials.

First-trained identity	Lowest win rate	Widest $q = 0$ spread
Agent 0	20/20 [0.832,1.000]	19/20 [0.751,0.999]
Agent 1	20/20 [0.832,1.000]	19/20 [0.751,0.999]
Agent 2	20/20 [0.832,1.000]	17/20 [0.621,0.968]

C Claim Boundaries and Prior-Art Scope

Table 5: Expanded novelty boundary. Restoration is not a replacement for best-response/exploitability evaluation; it tests a different response mode and target quantity.

Work	Primary object	Response mode	Difference from restoration
Lanctot et al. (2017)	PSRO / policy generalization	approximate BR to policy mixtures	response-oracle/meta-game objective; not joint continuation from a path-induced endpoint
Timbers et al. (2022)	approximate exploitability	learned unilateral BR vs fixed policy	worst-case utility/equilibrium robustness; not mutual resumed learning
Nekoei et al. (2023)	cooperative optimization	sequential/multi-timescale active learner	training method, not endpoint response audit
Vadori et al. (2024)	multi-agent dealer simulation	interacting/shared-policy learning	richer market simulation; no restore-all endpoint test
Wang et al. (2025)	heterogeneous market-maker competition	later agents vs pre-trained fixed A	role/objective heterogeneity changes with training stage
Cheridito et al. (2025)	LOB MARL and execution	one role retrained vs fixed learned counterparts	targeted retraining, not all-agent restoration
This work	training-path endpoint audit	all suppressed learners resume the specified learning dynamics	system-statistic persistence with matched continuation control

Evidence-bounded claim map. The manuscript supports five linked statements: (C1) the specified staged curriculum alters the market endpoint; (C2) the gap fails the specified 600-episode co-adaptive restoration test while matched reference continuation stays near baseline; (C3) the tested block×placeholder interaction is positive for elevated placeholders while the corresponding tight-placeholder block contrast is near zero; (C4) the outcome-level disadvantage follows training role rather than fixed identity; and (C5) the prespecified block, placeholder, and restoration contrasts reproduce with DQN in the same testbed. These claims do not imply equilibrium, arbitrary-response robustness, universal sequential-learning harm, or real-market calibration.

D Restoration Horizon Trajectory

This check was added after the confirmatory studies and is descriptive; it does not create a new confirmatory claim. Its design and predictions were written and hashed before the full run; a one-seed timing pilot (seed 301, $h \leq 25$, reduced evaluation) was inspected beforehand and is not part of the result. For factorial seeds 301–320, we load the saved B3000/P4 and B3000/P7 staged Q-tables, regenerate the SIM endpoint (its MQS reproduces the factorial study exactly in 20/20 seeds), and apply the restoration operator of Section 4 (all learners updating, ϵ reset to 0.05) for up to 1,800 episodes with one continuing random stream. Greedy MQS is evaluated on a frozen copy at $h \in \{0, 25, 50, 100, 200, 400, 600, 1200, 1800\}$ with $T = 60{,}000$ timesteps and an evaluation stream shared across checkpoints and endpoints. Recovery is $[M_{\text{staged}}(0) - M_{\text{staged}}(h)]/[M_{\text{staged}}(0) - M_{\text{SIM}}(0)]$; no seed had a staged gap below the prespecified exclusion threshold of 0.005. All 20 seeds completed; none was dropped.

Table 6: Recovery fraction by horizon (mean, 95% seed-level bootstrap CI, seeds with recovery ≥ 0.9) and continued-SIM drift $M_{\text{SIM}}(h) - M_{\text{SIM}}(0)$ in units of 10^{-5}. Twenty seeds per cell.

h	P4 recovery	P4 ≥ 0.9	P7 recovery	P7 ≥ 0.9	Continued-SIM drift (10^{-5})
25	−0.429 [−0.600, −0.267]	0/20	−0.507 [−0.679, −0.343]	0/20	1.1 [−5.1, 9.4]
50	0.227 [0.068, 0.354]	0/20	0.126 [−0.004, 0.241]	0/20	−1.8 [−6.2, 2.7]
100	0.799 [0.765, 0.833]	2/20	0.829 [0.802, 0.854]	3/20	2.0 [−4.5, 9.8]
200	0.955 [0.938, 0.968]	19/20	0.962 [0.946, 0.975]	19/20	6.6 [−4.2, 22.0]
400	0.990 [0.981, 0.996]	20/20	0.995 [0.992, 0.997]	20/20	−2.2 [−6.5, 2.0]
600	0.997 [0.994, 1.000]	20/20	1.000 [0.997, 1.002]	20/20	−2.8 [−6.4, 0.7]
1,200	1.000 [0.997, 1.003]	20/20	1.001 [1.000, 1.003]	20/20	0.2 [−4.2, 4.6]
1,800	1.000 [0.998, 1.003]	20/20	1.001 [1.000, 1.002]	20/20	−2.5 [−5.7, 0.4]