Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29489

On-Policy Vs. Off-Policy Reinforcement Learning in ConnectX: Seat-Stratified Performance and the Role of Action Masking

Jiayi Lai

Faculty of Engineering, McGill University, Montreal, Canada
jiayi.lai@mail.mcgill.ca

Abstract. This study evaluates on-policy and off-policy reinforcement learning for ConnectX under a unified, seat-stratified evaluation protocol, with particular attention to invalid-action masking. Policies are trained against a fixed opponent schedule and evaluated offline with frozen parameters. The test set distinguishes Seen opponents from Unseen, deeper search (noisy-minimax-3 as the primary adversary; clean minimax-3 as a stress test). Uncertainty is summarized with 95% Wilson intervals, and all outcomes are reported by seat to avoid pooling bias. Across Seen opponents, masking consistently accelerates policy-gradient learning and reduces variability: MaskablePPO reaches target win-rate levels with fewer samples and tighter intervals than PPO, while value-based baselines reach an early asymptote. Against the Unseen noisy-minimax-3 in the first seat, PPO exceeds MaskablePPO (71.9% vs. 37.5%; intervals non-overlapping), whereas in the second seat masking yields large margins where legal actions are scarce (e.g., 87.6% vs. 17.3% against noisy-minimax-2; 48.3% vs. 8.3% against blocker; intervals non-overlapping). Against clean minimax-3, win rates approach zero within CI in the second seat for all methods, indicating a persistent gap to deeper search. These results imply two practical guidelines for legality-constrained, turn-based games: (i) legality-aware sampling is a substantive stability lever for policy-gradient methods rather than a peripheral implementation detail; and (ii) seat-stratified reporting is required to disentangle structural initiative from algorithmic effects. Overall, masking improves in-distribution efficiency and second-seat robustness but does not by itself close the generalization gap to stronger, unseen search.

Keywords: Reinforcement learning, connectx, invalid-action masking, seat-stratified evaluation, on-policy vs. off-policy

1. Introduction

Discrete, turn-based games with strict legality constraints form a stringent testbed for reinforcement learning (RL) because naïve exploration assigns probability mass to illegal actions, wasting samples and injecting gradient noise; invalid-action masking addresses this by renormalizing the policy over legal moves and has both theoretical and empirical support [1]. ConnectX—a configurable Connect-Four variant hosted on Kaggle—provides a public, reproducible setting for controlled comparisons

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29489

under shared rules [2]. Because move order can bias outcomes in connect-style games, results are reported with first-/second-player stratification to ensure fair comparison [2, 3].

Recent studies further underscore the importance of legality handling. For instance, Hou et al. provide empirical evidence that invalid-action masking can improve sample efficiency and robustness in applied settings [4]. At the same time, production support in SB3-Contrib (MaskablePPO) has made masking readily accessible in modern RL pipelines [5]. On-policy algorithms such as PPO [4] remain widely used in discrete-action control, while off-policy value-based methods including DQN [6] and QR-DQN [7] have shown competitive results on Atari-style benchmarks. Yet these baselines were not originally designed for legality-constrained domains, and few evaluations consider masking under a standardized, seat-stratified ConnectX protocol with explicit Seen–Unseen splits [2-4].

This manuscript therefore focuses on On-Policy vs. Off-Policy performance and on the role of action masking in ConnectX, with all results reported by seat under the standardized Kaggle evaluation interface. Among distributional approaches, recent overviews and monographs [8, 9] emphasize both the promise and limitations of modelling return distributions, further motivating systematic evaluation against unseen search-based opponents. In settings where many actions are illegal at a given state, policy-gradient methods are especially sensitive to legality handling; production-grade support for masking is available in Stable-Baselines3 Contrib (SB3-Contrib) via MaskablePPO with a dedicated masking interface, which standardizes training and evaluation with legality masks [5]. Accordingly, PPO/MaskablePPO represent the on-policy family, whereas DQN/QR-DQN represent the off-policy family for the seat-stratified comparisons emphasized here [5-7, 10].

This study provides a seat-stratified comparison of on-policy (MaskablePPO, PPO) versus off-policy (DQN, QR-DQN) reinforcement learning under a single standardized protocol with an explicit Seen–Unseen split, evaluated through the official Kaggle ConnectX interface [2, 5-7, 10, 11]. It quantifies the stability and robustness effects of invalid-action masking using mask-aware evaluation consistent with official implementations for legality-constrained policies [1, 5, 10]. Finally, it establishes a reproducible and statistically grounded protocol with Kaggle's standardized environment, seat-stratified metrics, and Wilson 95% confidence intervals for both Seen and Unseen opponents [2, 3, 8, 12].

2. Related work

Invalid-action masking provides a principled fix for legality-constrained action spaces and yields empirical gains as illegal-move prevalence rises [1]; production support in SB3-Contrib (MaskablePPO) makes masking easy to integrate in practice, though the docs themselves are not a controlled evaluation and do not address fixed schedule or unseen-opponent testing [4, 5].

As on-policy baselines, Proximal Policy Optimization (PPO) is widely adopted; its clipped surrogate objective enables multiple minibatch epochs per batch with favorable stability and sample efficiency, making it an appropriate baseline for discrete-action games [10], while A2C—the synchronous variant of A3C—offers a lightweight actor–critic comparator; however, neither baseline by itself handles legality constraints or opponent-pool design [10, 13]. Within this study's framing, PPO and its mask-aware variant (MaskablePPO) instantiate the on-policy family for seat-stratified comparisons under the Kaggle protocol [2, 5, 10, 11]. On the value-based side, DQN established human-level control in high-dimensional discrete domains and distributional variants such as QR-DQN and IQN improve return modeling beyond mean critics [6, 7]; recent overviews consolidate these distributional advances for modern evaluation practice [8, 14, 15]. Beyond games,

DOI: 10.54254/2755-2721/2026.TJ29489

distributional/offline RL continues to be adopted; for example, BCD4Rec shows that batch-constrained distributional RL improves value estimation under large action spaces using offline logs, reinforcing the use of QR-DQN as a concise distributional comparator [16]. Yet these off-policy methods were developed primarily for single-agent Atari-style settings and, absent legality-specific mechanisms, do not natively address masked evaluation against stronger held-out opponents in ConnectX [2, 5-7, 11].

Prior work explores opponent diversity via fixed schedule and league/PSRO-style populations [17-19]; here, a single standardized schedule and held-out deeper minimax adversaries are used to isolate (i) on-policy vs. off-policy differences and (ii) the incremental effect of action masking under one public Kaggle protocol with seat-stratified reporting [2, 3, 5, 11, 17-19]. Taken together, these observations motivate a seat-stratified comparison of on- and off-policy methods with and without masking, conducted under a standardized ConnectX evaluation framework with Seen–Unseen separation.

3. Methods

3.1. Environment and observation/action spaces

The experimental domain is the standard ConnectX instance with six rows (R = 6), seven columns (C = 7) shown in the Figure 1, and a connect-four win conditions (see Section 3.5 for the evaluation platform and API details).To interface with learning algorithms, a Gymnasium-compatible wrapper (ConnectXGym) is used to expose explicit observation and action spaces [20]. The state is represented by a dense board vector $\mathrm{x} = \mathbb{R}^{RC}$, obtained by column-major flattening of a ternary encoding in which own stones are +1, opponent stones are −1, and empty cells are 0.

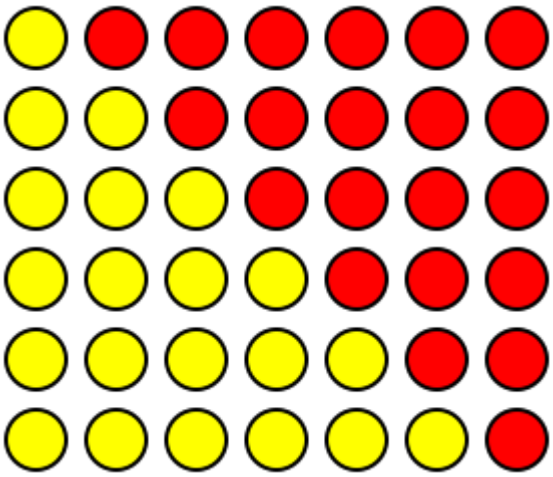


Figure 1. ConnectX board layout (R=6, C=7).

During training, horizontal mirroring is applied with probability $p = 0.5$. Let $\mathrm{M} \in \{0,1\}^{RC \times RC}$ denote the permutation matrix that reverses the column order; the augmented observation is $\tilde{x} = \mathrm{M}x$, and the legality mask is permuted identically (see Section 3.2). The action space is the finite set $A = \{0, \ldots, C-1\}$, where action aa corresponds to dropping a piece into column aa. Legality at state s depends solely on the occupancy of the top row. The legality mask as $m\left(s\right) \in \{0,1\}^{C}$ is defined componentwise

DOI: 10.54254/2755-2721/2026.TJ29489

$$m_a(s) = 1\{top\ cell\ of\ column\ a\ is\ empty\}, a = 0, \ldots, C-1 \quad (1)$$

To avoid confounding due to seat assignment, training may alternate first and second player deterministically across episodes, whereas evaluation fixes the seat and disables mirroring. Horizontal mirroring is applied with probability P_{mirror} = 0.5 during training and is disabled at evaluation.

3.2. Invalid-action masking and maskable policies

To enforce legality constraints and reduce variance, actions are sampled from a masked policy that renormalizes probabilities over legal moves. With m(s) from (Equation 1) and unconstrained logits $\tilde{\pi}_\theta(a \mid s)$, the executable policy is

$$\tilde{\pi}_\theta(a \mid s, m) = \frac{\tilde{\pi}_\theta(a|s) m_a}{\sum_{b \in A} \tilde{\pi}_\theta(b|s) m_b} \quad (2)$$

well-defined whenever $\sum_{b \in A} \tilde{\pi}_\theta(b \mid s)\, m_b$ >0. Equation 2 is implemented via a MaskablePPO wrapper that queries the environment's validity mask at each step, altering only sampling and log-prob computations while leaving the network unchanged [5].

Optimization follows PPO with clipped importance ratios computed under the masked policy. Let $\frac{\pi_\theta(a_t \mid s_t,\ m_t)}{\pi_{\theta^-}(a_t \mid s_t,\ m_t)}$. The update minimizes

$$L_{ppo}(\theta) = E[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t] - \beta E[KL(\pi_\theta \| \pi_{\theta^-})] \quad (3)$$

where $\epsilon > 0$ and $\hat{A}_t$ is the advantage estimate [10]. Because m_t gates both numerator and denominator in $r_t(\theta)$, gradients are taken only over legal actions—an approach with theoretical and empirical support in discrete domains with large invalid-action sets [1].

3.3. Opponent schedule and held-out adversary

Training uses a fixed stage-wise opponent schedule. Let $\{(O_k, n_k)\}_{k=1}^{K}$ denote stages where O_k is a distribution over opponents and n_k the number of environment steps in stage k. At episode start in stage k, $o \sim O_k$ is sampled and fixed; early stages use heuristic opponents (uniform-random, center, blocker), later stages introduce search-based adversaries (noisy and clean depth-2 minimax). The learning objective is unchanged—only the experience distribution shifts. To stress out-of-distribution generalization, depth-3 minimax is never used during training and is reserved as unseen at evaluation (see Section 3.5).

3.4. Learning algorithms and practical training

Benchmarking spans policy-gradient and value-based families implemented via Stable-Baselines3/Contrib interfaces [5]. On-policy methods include PPO and its maskable variant [5, 10]; off-policy baselines comprise DQN and the distributional QR-DQN [6, 7]. Policies are multilayer perceptrons over the 42-dimensional state; letting $_\theta(\mathrm{x})$ be action logits, the stochastic policy is

$$\pi_\theta\ (ax) = \ \mathrm{softmax}\left(\ _\theta(\mathrm{x})\ \right)_a, \quad (4)$$

DOI: 10.54254/2755-2721/2026.TJ29489

Maskable variants condition sampling and log-probabilities on the legality mask $m(s)$ defined in (Equation 1) (and renormalized as in Equation 2)). Value-based agents approximate $Q_\phi(s,a)$ with a target network; DQN minimizes squared TD error using

$$y(t) = r_t + \gamma \max_{a'} Q_{\phi^-}(s_{t+1}, a'), \tag{5}$$

whereas QR-DQN replaces the square loss with a quantile regression objective over N atoms using the pinball/Huber loss at quantile levels $\{\tau_i\}_{i=1}^{N}$[6, 7]. Hyperparameters follow established practice for PPO (e.g., clipped objective with Adam at a 3×10^{-4} learning rate and moderate rollout horizons) [10], while DQN/QR-DQN employ replay buffers and periodic target-network updates. Stage transitions checkpoint parameters (and, for value-based methods, optionally the replay buffer) to permit exact resumption; evaluation is strictly offline with frozen parameters and buffers.

3.5. Evaluation protocol

The trained policy is tested against Seen and Unseen opponents, using the official Kaggle connectx environment and the evaluate API under a fixed configuration, with parameters frozen and no learning during test episodes [2, 11]. The evaluation protocol isolates the two focal factors—policy family (on-policy PPO/MaskablePPO vs. off-policy DQN/QR-DQN) and the contribution of action masking—under one public protocol [5-7, 10].

The Seen opponent set mirrors the training schedule (uniform-random, center, minimax-2, noisy-minimax-2) as defined in Section 3.3. The unseen set is excluded from training, which defined as $U = \{noisy-minimax-3,\ minimax-3\}$; noisy-minimax-3 serves as the primary evaluation adversary, whereas minimax-3 functions as a stronger stress test. Here, noisy-minimax-3 denotes a depth-3 minimax adversary with controlled stochasticity at decision time (e.g., ε -greedy tie-breaking at the root and/or bounded noise within a top- k candidate set), which reduces brittleness to deterministic forcing lines and provides a more distributional probe of generalization; consequently, noisy-minimax-3 serves as the primary evaluation opponent. In contrast, minimax-3 remains a stress-test adversary, representing a stronger, clean depth-3 search used to assess worst-case robustness [2, 3].

For legality-aware baselines, evaluation is mask-aware, see Section 3.2 for the masking interface. For each agent–opponent–seat triple, N independent episodes are executed, and the empirical win rate $\hat{p}$ together with Wilson 95% confidence intervals to summarize binomial uncertainty at finite sample sizes [2, 8, 11].

3.6. Reproducibility and fairness

All training stages specify framework-level random seeds, and evaluation freezes parameters and buffers. Seat assignment is controlled explicitly, and first/second results are reported separately to avoid pooling asymmetries. The environment legality mask is applied consistently in both training and evaluation for maskable algorithms, while non-maskable baselines are safeguarded by a legality fallback that maps any illegal prediction to the nearest legal column without altering policy logits [5].

A Gymnasium-compatible interface exposes explicit action_space and observation_space declarations for compatibility and auditability [20]. The schedule is specified via a single --

DOI: 10.54254/2755-2721/2026.TJ29489

stages string with human-readable durations (random:40k,center:40k,blocker:80k,minimax2:120k,noisyminimax2:180k) and supports optional mixed pools by sampling opponents at reset() time. A front-end guard rejects stage names minimax3+ during training, ensuring the unseen set is never sampled; released configs/logs include the expanded opponent list for verification. The same mechanism that instantiates minimax2 enables drop-in evaluation against noisyminimax3 and minimax3 by name, rendering the seen/unseen split transparent and enforceable [2].

4. Experiments setups and results

4.1. Experimental setup

4.1.1. Environment and setup

Experiments follow the standardized platform in Section 3.5; episodes are instantiated via with kaggle_environments.make("connectx") and evaluated with evaluate() under frozen parameters. A Gymnasium-compatible wrapper (ConnectX Gym) exposes explicit observation_space/action_space for auditability [20].

Observations use the ternary board encoding from Section 3.1, and legality is enforced by the mask $m(s)$ in Equation 1; seat assignment alternates deterministically during training but is fixed during evaluation to avoid distribution shift [2, 20]. The opponent schedule is fixed across all agents (on-policy: PPO/MaskablePPO; off-policy: DQN/QR-DQN) so that any differences can be attributed to policy family and masking rather than exposure [5-7, 10]. Training opponents include random/center/blocker and depth-2 search (noisy-minimax-2, minimax-2), whereas depth-3 search (noisy/minimax-3) is held out for Unseen evaluation to probe out-of-distribution robustness under the known first-move advantage in Connect-Four [2, 3, 11].

4.1.2. Metrics

Primary metric. For each agent–opponent–seat triple report win-rate $\hat{p}$ and always separate first/second seats; overall is a weighted average as defined by the evaluation script schema [2, 11]. Uncertainty is summarized with Wilson 95% confidence intervals (per-opponent, per-seat, per-seed) to provide reliable coverage for binomial proportions [2, 8, 11].

Efficiency metrics. The AUC of win-rate versus steps and budget-matched deltas at fixed steps are computed to quantify online learning speed, matching the usage in the figures [2, 11]. All metrics are produced by the public pipeline that writes {win_rate, wilson_low, wilson_high} to CSV for auditability [2, 11].

4.1.3. Compared methods

On-policy (policy-gradient). PPO (on-policy) and MaskablePPO (legality-aware PPO) share architectures/hyper-parameters; the only difference is legality handling via mask-renormalization during sampling and log-probabilities [1, 5, 10].

Off-policy (value-based). DQN and QR-DQN represent distributional value learning comparators implemented with Stable-Baselines3/SB3-Contrib; QR-DQN improves return modeling beyond the mean critic [6-8]. These agents are evaluated under the same ConnectX protocol without legality-specific mechanisms at training time [2, 5, 11].

DOI: 10.54254/2755-2721/2026.TJ29489

4.1.4. Training and logging

Stage-wise training follows the fixed opponent schedule in Section 3.3; checkpoints are saved at stage boundaries (and, for off-policy methods, optional replay buffers) to permit exact resumption [2, 11]. Online curves are logged at each checkpoint and exported for AUC/Δ@budget computation; evaluation is performed offline with frozen parameters and buffers [2, 11]. Seat is alternated in training and fixed in evaluation; mirroring is disabled at test time [2]. For maskable methods, legality masks are consumed at inference per SB3-Contrib; non-maskable methods use a legality fallback that maps any illegal prediction to the nearest legal column without altering logits [2, 5].

4.2. Result and discussion

4.2.1. Offline evaluation seat-stratified outcomes

Table 1. Mean win rate with Wilson 95% CIs; 3000 games— first seat

	random	center	blocker	noisyminimax-2	minimax-2	noisyminimax-3	minimax-3
PPO	92.1% [91.0,93.0]	100.0% [99.9,100.0]	100.0% [99.9,100.0]	92.2% [91.2,93.1]	100.0% [99.9,100.0]	71.9% [70.2,73.4]	100.0% [99.9,100.0]
Maskable PPO	72.0% [70.3,73.5]	100.0% [99.9,100.0]	100.0% [99.9,100.0]	89.8% [88.6,90.8]	100.0% [99.9,100.0]	37.5% [35.8,39.3]	100.0% [99.9,100.0]
DQN	85.4% [84.1,86.6]	81.5% [80.0,82.8]	2.5% [2.0,3.2]	75.4% [73.8,76.9]	100.0% [99.9,100.0]	19.7% [18.3,21.1]	100.0% [99.9,100.0]
QR-DQN	78.4% [76.9,79.8]	89.6% [88.4,90.6]	8.1% [7.2,9.1]	84.5% [83.2,85.8]	100.0% [99.9,100.0]	27.3% [25.7,28.9]	100.0% [99.9,100.0]

Table 2. Mean win rate (%) by algorithm and opponent in 3000 games— second seat

	random	center	blocker	noisyminimax-2	minimax-2	noisyminimax-3	minimax-3
PPO	82.3% [80.9,83.6]	89.7% [88.6,90.7]	8.3% [7.4,9.4]	17.3% [16.0,18.7]	0.0% [0.0,0.1]	9.9% [8.9,11.0]	0.0% [0.0,0.1]
MaskablePPO	68.1% [66.4,69.8]	100.0% [99.9,100.0]	48.3% [46.5,50.1]	87.6% [86.3,88.7]	100.0% [99.9,100.0]	24.2% [22.7,25.8]	0.0% [0.0,0.1]
DQN	75.0% [73.5,76.5]	62.4% [60.7,64.1]	22.8% [21.4,24.4]	68.7% [67.1,70.4]	100.0% [99.9,100.0]	4.2% [3.5,5.0]	0.0% [0.0,0.1]
QR-DQN	63.0% [61.3,64.7]	69.6% [67.9,71.2]	3.1% [2.5,3.7]	68.1% [66.4,69.7]	100.0% [99.9,100.0]	5.4% [4.6,6.2]	0.0% [0.0,0.1]

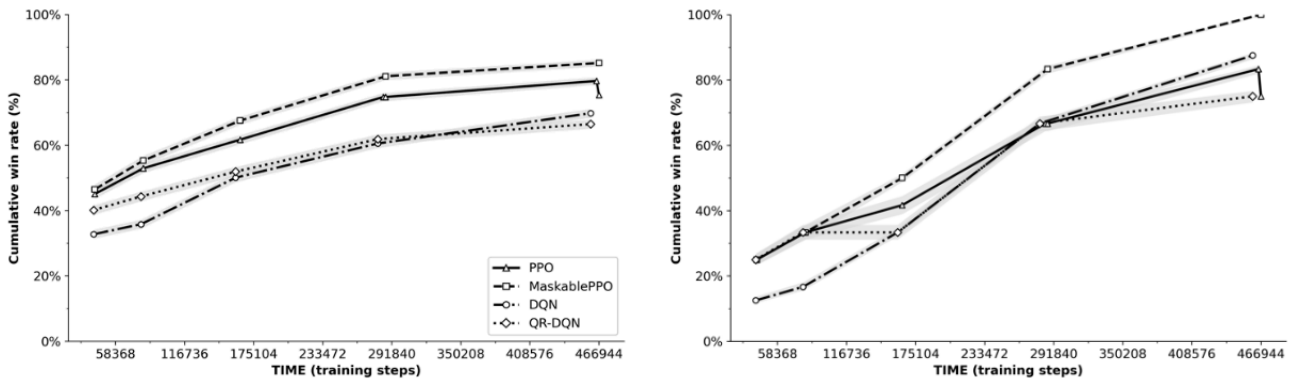
Entries report mean win-rate (%) with Wilson 95% Cis intervals; each cell aggregates 1000 games × 3 different seeds{123,456,789}. Across seen opponents (random, center, blocker, noisy-minimax-2, minimax-2), all algorithms reach a high plateau in the first seat. Against the unseen noisy-minimax-3, PPO is higher than MaskablePPO (71.9% [70.2, 73.4] vs. 37.5% [35.8, 39.3]), with DQN/QR-DQN in between (DQN 19.7% [18.3, 21.1], QR-DQN 27.3% [25.7, 28.9])—consistent with the well-known first-move advantage in Connect Four and motivating seat-stratified reporting rather than overall averages [3]. The Wilson intervals for PPO and MaskablePPO on noisy-minimax-3 are disjoint, indicating a statistically reliable separation.

DOI: 10.54254/2755-2721/2026.TJ29489

In the second seat, masking effects become salient: MaskablePPO attains 87.6% [86.3, 88.7] versus 17.3% [16.0, 18.7] for PPO against noisy-minimax-2, and 48.3% [46.5, 50.1] versus 8.3% [7.4, 9.4] against blocker. These pairs of intervals do not overlap, confirming large, statistically meaningful margins for MaskablePPO. By contrast, versus the unseen minimax-3, all algorithms drop to 0.0% with near-zero upper bounds (e.g., 0.0% [0.0, 0.1])in the second seat. Together, the seat-stratified numbers and their Wilson CIs confirm that a held-out stronger search opponent is necessary to probe out-of-distribution generalization without conflating structural asymmetry with algorithmic differences [2, 3]. When win rates sit at boundaries (e.g., 100.0% [99.9, 100.0] or 0.0% [0.0, 0.1]), the Wilson form appropriately expresses near-certainty while avoiding degenerate zero-width intervals.

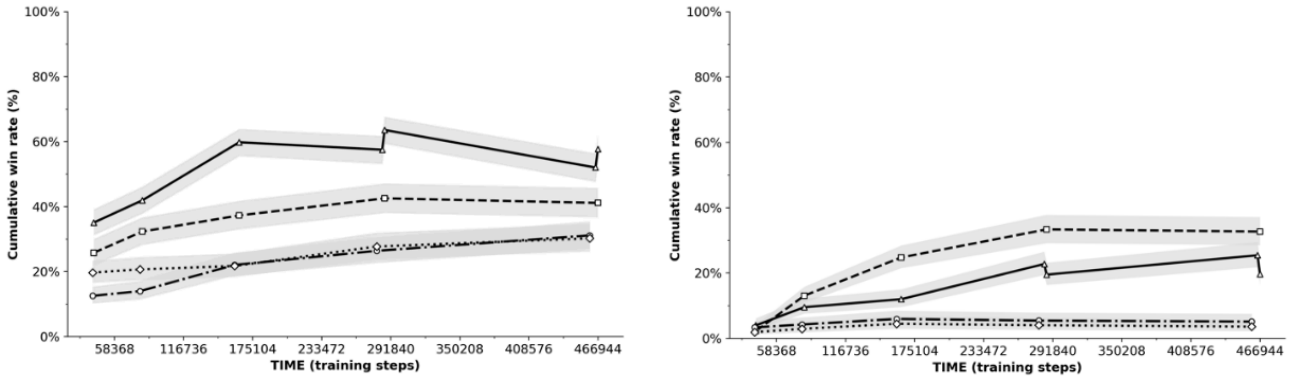
The first-seat dominance reflects the structural initiative inherent to Connect Four [3]; hence, first-seat gains should not be over-interpreted as universal robustness. The second-seat results isolate where invalid-action masking helps most: in high-pressure late-game states with few legal columns, the masked policy avoids allocating probability to illegal moves, improving effective search over admissible actions and yielding large margins over PPO against blocker and noisy-minimax-2. The non-overlapping Wilson intervals for these matchups (e.g., 48.3% [46.5, 50.1] vs. 8.3% [7.4, 9.4] on blocker; 87.6% [86.3, 88.7] vs. 17.3% [16.0, 18.7] on noisy-minimax-2) demonstrate that these gains are statistically reliable rather than artifacts of variance. The universal approach zero within CI with tight upper bounds against minimax-3 (second seat) indicates that masking alone does not overcome the depth advantage of a stronger, unseen search adversary; it improves efficiency and stability but does not by itself resolve distributional shift. Conversely, multiple 100.0% [99.9, 100.0] cells (e.g., first-seat vs. minimax-2) should be interpreted as ceiling effects consistent with strong first-move advantage rather than evidence of broad OOD robustness. These observations justify the emphasis on seat-stratified offline reporting and on using a held-out deeper opponent to assess generalization rigorously [2, 3].

4.2.2. Online evaluation learning efficiency



(a) (b)

DOI: 10.54254/2755-2721/2026.TJ29489



(c) (d)

Figure 2. Cumulative win rate versus training steps; shaded regions denote Wilson 95% confidence intervals; (a) seen-opponent average; (b) minimax-2 (seat-averaged); (c) noisy-minimax-3 (first seat); (d) noisy-minimax-3 (second seat); legend: PPO, MaskablePPO, DQN, QR-DQN

In the seen-opponent average (Figure 2a), all methods improve steadily with training, but MaskablePPO maintains the leading trajectory, reaching higher win-rate levels with fewer steps than PPO, while DQN and QR-DQN lag throughout. The Wilson 95% confidence bands narrow as training proceeds and become largely separated between MaskablePPO and PPO at later checkpoints, indicating a reliable efficiency gap (Section 3.6) [12]. Against minimax-2 (seat-average) (Figure 2b), MaskablePPO pulls away most decisively: its curve rises more steeply and ends near the top of the axis; PPO follows below, and value-based methods reach an early asymptote with wider bands, reflecting higher variability. For the unseen noisy-minimax-3, first seat (Figure 2c), the ranking reverses between the two policy-gradient baselines: PPO sits above MaskablePPO over most checkpoints (peaking around the mid-run, then slightly dipping at the final step), whereas MaskablePPO remains clearly ahead of DQN/QR-DQN. Finally, in the unseen noisy-minimax-3, second seat (Figure 2d), overall win rates are depressed, but MaskablePPO regains a noticeable edge over PPO; DQN/QR-DQN stay near the floor. Across (c)→(d), Wilson bands widen and means drop, signaling the combined difficulty of out-of-distribution evaluation and the disadvantageous second-seat condition.

Three consistent patterns emerge. (i) Efficiency benefit of masking in distribution. In Figures 2a–b, invalid-action masking accelerates learning and reduces variability—seen in faster climbs and tighter Wilson bands—matching the theoretical expectation that removing illegal-move probability mass lowers gradient noise and stabilizes policy-gradient updates [1]. The particularly strong separation on minimax-2 suggests masking is most beneficial once states frequently present few legal columns, so exploration is otherwise wasted on invalid actions. (ii) Seat asymmetry and OOD stress. Moving from the seen settings (a–b) to noisy-minimax-3 (c–d) systematically lowers ceilings and broadens bands; the first vs. second split is stark in (c) vs. (d), consistent with the structural first-move advantage of Connect Four and underscoring the need to report seat-stratified metrics rather than only overall averages [2, 3]. (iii) Where masking helps—and where it might not. The first-seat superiority of PPO over MaskablePPO on noisy-minimax-3 (Figure 2c) indicates that, against a deeper unseen adversary from a favorable seat, broader stochastic exploration by an unmasked policy can occasionally secure higher payoffs; by contrast, the second-seat curve (Figure 2d) shows masking regaining the advantage precisely in the harsher regime where legal actions are scarce and tactical errors are less forgivable. Together, Figures 2a–d

DOI: 10.54254/2755-2721/2026.TJ29489

show that masking reliably improves in-distribution efficiency and second-seat robustness, while generalization to deeper unseen search remains limited—an effect attributable to opponent strength and seat structure rather than training instability [1, 6, 7].

4.2.3. Online vs. offline policies

Across the online efficiency curves (Figures 2a–c), MaskablePPO consistently outperforms PPO by reaching equivalent win-rate thresholds with fewer training steps and maintaining narrower Wilson 95% confidence intervals [12]. This advantage is most visible against minimax-2, where MaskablePPO converges smoothly while PPO displays oscillations, reflecting the stabilizing effect of invalid-action masking on policy-gradient updates [1]. Against the unseen noisy-minimax-3, both policies achieve lower asymptotes with broader intervals, yet MaskablePPO retains an early-phase efficiency edge, showing particular strength in late-game states with few legal actions. In contrast, the offline value-based methods(Tables 1–2; Figure 2d) perform markedly worse: while QR-DQN occasionally surpasses DQN, both algorithms suffer from prolonged plateaus, wider confidence intervals, and weaker generalization to unseen opponents, especially in the second seat. Their collapse against minimax-3 highlights the instability of bootstrapped value learning in adversarial, legality-constrained environments [6, 7], and underscores the importance of seat-stratified reporting to distinguish structural asymmetry from algorithmic differences [2, 3].

4.2.4. Contribution of invalid-action masking

Invalid-action masking contributes along two axes. First, efficiency: by zeroing probability mass on illegal moves, masking reduces gradient noise and stabilizes policy-gradient updates, yielding faster, more reliable gains (tighter Wilson bands and earlier threshold-crossings) in Figures 2a–b [1]. Second, robustness under scarcity: when the legal action set is small—late game and especially the second seat—masking tightens exploration to admissible moves, producing the large margins observed for MaskablePPO over PPO against blocker and noisy-minimax-2 in Table 2, and mirroring the gap in Figure 2d. At the same time, the first-move advantage and opponent depth still dominate outcomes against noisy-minimax-3, particularly from the second seat (Figures 2c–d), showing that masking improves learning efficiency and seat-specific robustness but does not by itself solve out-of-distribution generalization to stronger search adversaries [2, 3, 17-19].

5. Conclusion

This study compared on-policy and off-policy reinforcement learning in ConnectX under a seat-stratified, reproducible protocol, with a focus on invalid-action masking. Under the fixed evaluation setup, masking consistently accelerated policy-gradient learning in-distribution and reduced uncertainty, while value-based baselines plateaued earlier. Mask-aware PPO reached target win-rate thresholds with fewer samples and exhibited tighter confidence bands than its unmasked counterpart.

Performance against the deeper, held-out noisy-minimax-3 was strongly seat-dependent: first seat favored PPO, whereas second seat highlighted the robustness benefits of masking when the legal action set is small and tactical errors are costly. Clean minimax-3 remained a stress test with near-zero win rates, indicating that masking improves efficiency and seat-specific robustness but does not close the generalization gap to stronger search. These results support mask-aware PPO and seat-aware reporting as default practice for legality-constrained, turn-based games.

DOI: 10.54254/2755-2721/2026.TJ29489

Limitations include a single standardized opponent schedule and an unseen set focused on depth-3 search. Future work should expand the unseen set (deeper or population-based adversaries), examine interactions between masking and lightweight test-time safeguards under frozen-parameter evaluation, and assess transfer across board sizes.

References

[1] Huang, S., & Ontañón, S. (2022). A Closer Look at Invalid Action Masking in Policy Gradient Algorithms. In FLAIRS-35 Proceedings. arXiv: 2006.14171.
[2] Kaggle. (n.d.). kaggle-environments: Agent and environment library for reinforcement learning competitions. GitHub repository. Retrieved from https: //github.com/Kaggle/kaggle-environments.
[3] Allis, L. V. (1988). A Knowledge-Based Approach of Connect-Four (Master's thesis). Vrije Universiteit Amsterdam. (Often cited with subtitle "The Game is Solved: White Wins".)
[4] Hou, Y., Liang, X., Zhang, J., Yang, Q., Yang, A., & Wang, N. (2023). Exploring the Use of Invalid Action Masking in Reinforcement Learning: A Comparative Study of On-Policy and Off-Policy Algorithms in Real-Time Strategy Games. Applied Sciences, 13(14), 8283. https: //doi.org/10.3390/app13148283.
[5] Stable-Baselines3 Contrib Team. (n.d.). Maskable PPO (v2.5.0) — Documentation. Retrieved from https: //sb3-contrib.readthedocs.io/en/v2.5.0/modules/ppo_mask.html.
[6] Dabney, W., Rowland, M., Bellemare, M. G., & Munos, R. (2017). Distributional Reinforcement Learning with Quantile Regression. arXiv: 1710.10044.
[7] Dabney, W., Ostrovski, G., Silver, D., & Munos, R. (2018). Implicit Quantile Networks for Distributional Reinforcement Learning. In ICML 2018. arXiv: 1806.06923.
[8] Bellemare, M. G., Dabney, W., & Munos, R. (2017). A Distributional Perspective on Reinforcement Learning. Website: distributional-rl.org.
[9] Cho, T., Han, S., Lee, H., Lee, K., & Lee, J. (2023). Pitfall of Optimism: Distributional Reinforcement Learning by Randomizing Risk Criterion. NeurIPS 2023. arXiv: 2310.16546.
[10] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv: 1707.06347.
[11] Kaggle. (n.d.). kaggle-api: Official Kaggle API. GitHub repository. Retrieved from https: //github.com/Kaggle/kaggle-api.
[12] McGrath, O., & Burke, K. (2021). Binomial Confidence Intervals for Rare Events. arXiv: 2109.02516.
[13] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous Methods for Deep Reinforcement Learning. ICML 2016. arXiv: 1602.01783.
[14] Nguyen, T. T., Gupta, S., & Venkatesh, S. (2020). Distributional Reinforcement Learning via Moment Matching. AAAI 2021. arXiv: 2007.12354.
[15] Cho, T., Han, S., Lee, H., Lee, K., & Lee, J. (2023). Pitfall of Optimism: Distributional Reinforcement Learning by Randomizing Risk Criterion. NeurIPS 2023. arXiv: 2310.16546.
[16] Garg, D., Gupta, P., Malhotra, P., Vig, L., and Shroff, G., "Batch-Constrained Distributional Reinforcement Learning for Session-based Recommendation, " arXiv: 2012.08984 [cs.LG], 2020. DOI: 10.48550/arXiv.2012.08984.
[17] Narvekar, S., Pereira, A., Leonetti, M., & Stone, P. (2020). Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. arXiv: 2003.04960.
[18] Bighashdel, A., Wang, Y., McAleer, S., Savani, R., & Oliehoek, F. A. (2024). Policy Space Response Oracles: A Survey. IJCAI-24. arXiv: 2403.02227.
[19] McAleer, S., Lanier, J. B., Fox, R., & Baldi, P. (2021). Pipeline PSRO: A Scalable Approach for Finding Approximate Nash Equilibria in Large Games. arXiv: 2006.08555.
[20] Farama Foundation. Gymnasium API — Env: attributes (action_space, observation_space). Official documentation. Available at: https: //gymnasium.farama.org/api/env/.