The Expanded Othello AI Arena: Evaluating Intelligent Systems Through Constrained Adaptation to Unseen Conditions

Anonymous authors
Paper under double-blind review

Abstract

The ability to rapidly adapt to environmental changes is a core requirement for Artificial General Intelligence (AGI), yet most AI benchmarks evaluate performance in static environments. We present the Expanded Othello AI Arena, a benchmark designed to measure Skill-Acquisition Efficiency — the rate at which agents discover latent objectives and converge to effective strategies within a limited interaction budget. The Arena formalizes a spectrum of 56 environments using a parametric framework $\mathcal{E} = (\mathcal{L}, \mathcal{C})$, where $\mathcal{L}$ defines Othello board geometries and $\mathcal{C}$ represents latent winning conditions via a disc-ratio threshold K. This parameterization requires agents to decipher terminal rules through direct interaction while simultaneously interpreting the opponent's behavior — in narrow regimes, agents must strategically induce the opponent into violating the hidden threshold to secure victory. Unlike traditional evaluation, the Arena imposes a strict 2,000-game interaction budget to prioritize sample efficiency over asymptotic optimization. We establish the benchmark's utility through a neuroevolutionary adaptive-Minimax baseline that utilizes meta-learned spatial priors and adaptive weighting. Our empirical analysis reveals that while this baseline achieves competitive performance in standard and inverse regimes, it fails in narrow-interval regimes that demand adversarial inducement, exposing a substantial efficiency gap that gradient-based reinforcement learning cannot bridge even with five times the interaction budget. Released as an extensible Python-based research toolkit, the Arena provides a standardized platform for exploring research directions including test-time learning, in-context learning, and world models. The code is available at: `https://anonymous.4open.science/r/ExpandedOthello/`.

1 Introduction

The capacity for rapid strategic adaptation to novel environments has been one of the primary goals for Artificial General Intelligence (AGI) (Chollet, 2019). In recent years, specialized artificial agents have achieved remarkable milestones in various domains. For instance, reinforcement learning (RL) agents have reached superhuman proficiency in high-complexity board games such as Go, Chess, and Shogi (Silver et al., 2016; 2017; 2018), and large language models (LLMs) have demonstrated their reasoning abilities in text-based games (Yang et al., 2024; Zhuo & Murata, 2024; Hu et al., 2025; Phan et al., 2025) and complex, long-horizon video games such as Minecraft (Wang et al., 2023). However, these successes are predominantly anchored in knowledge-saturated environments, where the agents benefit from either an extensive amount of training interactions or decide actions zero-shot, purely based on prior knowledge.

To mitigate such properties of traditional environments, various studies have proposed environments that incorporate the variation concept, such as ProcGen (Cobbe et al., 2020) and MiniHack (Samvelyan et al., 2021). These benchmarks challenge agents to generalize across procedurally generated levels, requiring agents to learn to adapt to novel environments. However, the novelty or the variation introduced by these environments is limited. For instance, ProcGen focuses on the variation of visualizations while the set of games remains the same, and MiniHack offers varying level layouts and tasks within a fixed rule-set. Nonetheless, while these environments serve as a great testbed for traditional meta-learning, they typically do not deal with changes to core game dynamics, such as reversed winning conditions. Furthermore, most existing benchmarks evaluate adaptation through single-player interactions with the environment alone, thereby overlooking the

necessity for social or adversarial reasoning where an agent must interpret and influence the behavior of other intelligent actors. This oversight limits the evaluation of an agent's ability to identify hidden reward structures and objective functions from a few sparse terminal signals, which is a critical facet of general intelligence that involves not only understanding environmental structures but also identifying the fundamental objectives of the interaction itself.

Therefore, we introduce the Expanded Othello AI Arena, a benchmark designed to move beyond structural or visual perturbations by incorporating objective-level novelty within a competitive framework. Unlike existing platforms, the Arena compels agents to infer terminal game rules and adapt their strategic behavior within a strictly constrained interaction budget while contending with an opponent, thereby shifting the focus toward genuine skill-acquisition efficiency in an adversarial setup. Consider an agent that dominates a game with 44 discs against 20, yet unexpectedly loses because the hidden rule penalizes excessive occupation. Can an intelligent system, receiving only win/loss signals, identify such a non-obvious objective within a limited number of trials? The Expanded Othello AI Arena is designed to answer precisely this question.

To systematically evaluate these capabilities, we formalize the environment space of the Expanded Othello AI Arena as a parametric framework $\mathcal{E} = \mathcal{L} \times \mathcal{C}$, defined as the Cartesian product of the layout space $\mathcal{L}$ and the condition space $\mathcal{C}$. Each environment instance $E \in \mathcal{E}$ is represented as a tuple (L, C), where $L \in \mathcal{L}$ specifies the board geometry and obstacle placements fully observable to the agent. In contrast, $C \in \mathcal{C}$ denotes the hidden winning conditions, parameterized via a disc-ratio threshold K, which must be inferred solely through a limited number of interactions with the environment E. This K-parameterization generalizes standard, inverse, and narrow-interval victory states. Furthermore, since Othello works as a 1-vs-1 zero-sum setup, successful adaptation demands not only identifying K but also employing a primitive form of Theory of Mind (ToM) (Sap et al., 2022) to anticipate and induce opponent moves toward favorable occupancy ratios or strategically leading them to violate winning conditions. Consequently, the Arena serves as a sophisticated micro-world that compels agents to engage in strategic information-seeking — forming a strategy not only to optimize immediate performance but to actively uncover the latent rules governing the task while coping with the dynamic unpredictability of an adversary.

By imposing a strict 2,000-game interaction budget, the Arena shifts the evaluation focus toward the velocity of strategic convergence rather than asymptotic optimization. This constraint exposes a critical efficiency gap that current optimization-heavy agents often fail to bridge. Successfully navigating this budget necessitates meta-cognitive strategy synthesis, compelling agents to comprehend novel geometries (L) and latent winning conditions (C) from an extremely restricted stream of experience. This requirement mirrors the human ability to achieve functional play in unfamiliar game variants after only a handful of trials (Lake et al., 2017), posing a rigorous test for the development of truly adaptive agents (Chollet, 2019; Reed et al., 2022).

To demonstrate the practical utility and discriminative power of the Arena, we provide a foundational baseline utilizing a neuroevolutionary adaptive-Minimax architecture. This hybrid system leverages meta-learned spatial priors combined with a dynamic utility-weighting mechanism that adapts to latent winning conditions ($\mathcal{C}$). Our empirical analysis reveals that while this baseline achieves significantly higher sample efficiency than traditional reinforcement learning agents within the limited interaction budget, a substantial performance ceiling remains in the most complex environment variants. By releasing the Arena as a modular, local Python-based research toolkit, we provide a standardized platform for exploring emerging research frontiers, such as test-time training (TTT), in-context learning (ICL), and the development of World Models (Ha & Schmidhuber, 2018a;b).

In summary, the primary contributions of this work are:

- **Formalization of a Parametric Environment Space**: We introduce a rigorous framework for evaluating adaptive intelligence by formalizing the environment space as $\mathcal{E} = \mathcal{L} \times \mathcal{C}$, decoupling observable board geometries (L) from latent winning conditions (C).

- **Integration of Strategic ToM in Adaptation**: We propose a benchmark where adaptation requires the simultaneous deciphering of latent rules and the interpretation of opponent intentions in a zero-sum context, demanding Theory of Mind to induce favorable outcomes or rule violations.

- **The Expanded Othello AI Arena Toolkit**: We release a modular, Python-based research toolkit designed to quantify Skill-Acquisition Efficiency under strict interaction budgets and analyze the velocity of strategic convergence.

- **Empirical Baseline and Efficiency Analysis**: We establish the benchmark's utility through a neuroevolutionary baseline, identifying a substantial "Efficiency Gap" compared to human-level adaptation in low-data regimes.

2 The Expanded Othello AI Arena Environment

To evaluate an agent's capacity for rapid strategic adaptation, we formalize a parametric environment space $\mathcal{E} = \mathcal{L} \times \mathcal{C}$ built upon the core mechanics of Othello. The Arena is designed as a rigorous proving ground for Skill-Acquisition Efficiency, specifically evaluating an intelligent system I — a combination of a learning algorithm and a corresponding model — on its ability to adapt to novel rules and adversarial challenges.

2.1 Adaptation Formalization

The adaptation task in the Arena is defined as the process by which a system I interacts with a specific environment instance $E = (L, C)$ to adapt its policy π into an environment-specific strategy π_E. This transition is strictly limited by an interaction budget of $N_{games} = 2,000$ episodes. We formalize this transformation as:

$$I(E) : \pi \xrightarrow{N_{games}} \pi_E \tag{1}$$

In this framework, the primary subject of evaluation is the efficiency of system I. The goal is to measure how effectively the system's underlying algorithm can transition from its general baseline state π to an optimized state π_E by deciphering terminal signals and navigating the strategic uncertainty of the environment within the allowed scarce interaction opportunities.

This formulation shifts the evaluation from asymptotic performance to the velocity of strategic convergence, focusing on how an intelligent system I identifies environmental conditions with respect to the opponent's unpredictable actions to produce a functional, task-specific policy π_E.

2.2 Game Dynamics and Substrate Rationale

Othello (Reversi) is a classic zero-sum, perfect-information board game characterized by the strategic reversal of opponent discs through spatial sandwiching mechanics. We select Othello as the foundational substrate for our benchmark due to its unique combination of simple local rules and high global state-change dynamics. Unlike games with static piece values, Othello involves frequent state reversals (flipping), making it an ideal environment where the geometric layout of the board directly affects core strategies. Furthermore, the game's terminal state naturally maps to a continuous disc occupancy ratio $\rho \in [0, 1]$, providing a parameterizable winning condition variation.

The Expanded Othello framework maintains the two fundamental mechanics of traditional Othello as its invariant core:

Flipping Mechanism: A player places a disc on an empty cell to sandwich one or more opponent discs between the new disc and an existing disc of their own, along any of the eight possible directions.

Legal Moves and Termination: A move is only legal if it flips at least one opponent disc. The game ends when neither player can make a legal move.

While these core dynamics remain constant, the Expanded Othello introduces novelty and complexity by perturbing the spatial layout and the terminal victory conditions, forcing the agent to adapt and discover the dynamics of the novel environment it is confronted with.

2.3 Formalization of the Environment

We define the environment space $\mathcal{E}$ as the Cartesian product of a layout space $\mathcal{L}$ and a condition space $\mathcal{C}$:

$$\mathcal{E} = \mathcal{L} \times \mathcal{C} \tag{2}$$

Each environmental instance $E \in \mathcal{E}$ is represented as a tuple $E = (L, C)$. A critical feature of the Expanded Othello is the asymmetry of information, where the spatial layout L is fully observable to the agent, while the victory condition C remains latent.

2.3.1 The Observable State and the Layout

In the Expanded Othello, the observable state at turn t is defined by the spatial configuration of discs and obstacles on the board. This state is represented as a grid $G_t \in \{-1, 0, 1, 2\}^{W \times H}$, where the cell values are assigned as follows: **1** and **-1** denote the discs of the agent and the opponent, respectively; **0** represents an empty playable cell; and **2** signifies an immutable obstacle. Obstacles are permanently unplayable and obstruct the flipping logic; specifically, discs cannot be sandwiched across an obstacle.

The layout L is formally defined as the initial board configuration G_0, and thus remains static throughout the game. While the positions of discs evolve, the distribution of obstacles is fixed upon environment initialization, introducing significant spatial diversity. By incorporating irregular board boundaries and non-interactive obstacles, the Expanded Othello effectively disrupts the traditional heuristics of standard Othello, as such methods are topology-dependent on the classical Othello (Rose, 2005; Yoshioka et al., 1999; Lucas, 2008; Jaśkowski, 2014; Szubert et al., 2013; Buro, 1995). This forces agents to generalize spatial value-mapping across diverse and novel topologies.

2.3.2 The Interval-Based Victory Condition

While the layout L provides the spatial context, the condition space $\mathcal{C}$ defines the strategic constraints of the environment. Unlike traditional games where dominance strictly correlates with victory, Expanded Othello parameterizes success through a latent threshold $K \in \mathbb{R} \setminus \{0.5\}$, introducing a relative victory logic based on the occupancy ratio $\rho \in [0, 1]$, which is defined as follows for each player's number of discs on the board n:

$$\rho = \frac{n_{agent}}{n_{agent} + n_{opponent}} \tag{3}$$

Specifically, K defines an upper bound on the majority player's occupancy (if $K > 0.5$) or a lower bound on the minority player's occupancy (if $K < 0.5$). Victory is determined by whether the constrained player remains within the permissible range. If the constrained player violates this boundary, victory is automatically forfeited to the opponent, regardless of the opponent's occupancy ratio. Let Player i and j be the two participants. The victory of player i is determined by the following regime-based logic:

1. Majority-Constraint Regime ($K > 0.5$): In this regime, the majority player ($\rho > 0.5$) must not exceed the threshold K. Player i wins if either of the conditions is met:

- **Successful Restraint**: $0.5 < \rho_i \leq K$
- **Opponent Forfeit**: $\rho_j > K$ (The opponent occupied the majority but violated the constraint).

2. Minority-Constraint Regime ($K < 0.5$): In this regime, the minority player ($\rho < 0.5$) must not fall below the threshold K. Player i wins if either of the conditions is met:

- **Successful Restraint**: $K \leq \rho_i < 0.5$
- **Opponent Forfeit**: $\rho_j < K$ (The opponent occupied the minority but violated the constraint).

The final reward $R_i \in \{+1, 0, -1\}$ is assigned based on these conditions, where $R_i = +1$ and -1 is assigned for the victory and defeat of the player i, respectively. $R_i = 0$ is reserved for a perfect draw ($\rho_i = \rho_j = 0.5$).

By varying K, this formulation recovers several distinct Othello variants within a single parameter:

Standard (Positive) Othello ($K > 1.0$): By setting $K > 1.0$, the opponent forfeit condition becomes inaccessible, and the effective successful restraint becomes $0.5 < \rho_i$, as $\rho \in [0, 1]$. This recovers the traditional majority-win rule, where the objective is to simply maximize disc occupancy.

Narrow-Majority Othello ($0.5 < K \leq 1.0$): This regime requires the agent to seek a majority ($\rho > 0.5$) while avoiding over-domination leading to an automatic forfeit ($\rho_j > K$).

Inverse (Negative) Othello ($K < 0$): By setting $K < 0.0$, the successful restraint condition effectively becomes $\rho_i < 0.5$, and the opponent forfeit option is prohibited. This recovers the simple minority-win rule, where the objective is to minimize the disc occupancy without further constraints.

Narrow-Minority Othello ($0 \leq K < 0.5$): This regime requires the agent to maintain a specific minority ratio ($K \leq \rho < 0.5$) while avoiding the forfeit condition.

The inclusion of the opponent forfeit logic introduces a sophisticated tactical layer of strategic forfeit inducement. In narrow regimes, a player who anticipates an inability to remain within the permissible bounds of K may pivot to a strategy of intentional disc concession. By exploiting the mandatory flipping mechanics of Othello, an agent can deliberately surrender board space to force the opponent into a position that violates the threshold K. This allows the agent to secure a victory ($R = +1$) from what would traditionally be considered a losing board state. Furthermore, in extreme cases such as $K = 0.52$, the Successful Restraint interval becomes nearly impossible to maintain — requiring an occupancy of exactly 33 discs on a standard 8×8 board ($\rho \in (0.5, 0.52] \Rightarrow \#\text{discs} \in (32, 33.28]$). In such scenarios, the strategic priority shifts entirely from optimizing one's own discs to actively engineering an opponent's forfeit via adversarial reasoning.

Consequently, the Expanded Othello environment is transformed from a stationary optimization task into a rigorous test of online system identification and adversarial reasoning. This transition necessitates the functional integration of Theory of Mind—the capacity to model and anticipate an opponent's strategic intentions—into the process of environmental adaptation. Because the winning condition C remains latent, the agent must not only decipher the hidden success criteria from sparse terminal feedback but also reason about the adversary's perspective to secure the victory.

For instance, consider a majority-constraint regime with $K = 0.8$. In the mid-game phase, consider two players i and j, having disc occupancy ratios as $\rho_i = 0.3$ and $\rho_j = 0.7$. Low occupancy player i may conclude that reaching the target winning range ($0.5 < \rho_i < 0.8$) is mathematically improbable through standard play. Instead, they might pivot to a strategy of strategic concession, attempting to induce the opponent into a forfeit by forcing them to exceed the $K = 0.8$ threshold. Conversely, the opponent j, anticipating this attempt to engineer an over-domination, should adapt by safely reducing their own occupancy to remain within the permissible bounds. This recursive strategic interplay demonstrates that successful adaptation in the Arena requires deciphering not just the environment's rules, but the opponent's reactive behavior.

2.4 Interaction Protocol

The interaction between an agent and the environment is designed to isolate the challenge of strategic adaptation. While the mathematical properties of the state G_t and the reward R_i are established in Section 2.3, this section defines the functional interface and the specific flow of information governing the interaction loop.

Action Space ($\mathcal{A}$): At any given turn t, the set of legal actions $\mathcal{A}_t$ is determined by the current board state G_t. An action $a_t \in \mathcal{A}_t$ is a coordinate (x, y) within the grid boundaries $W \times H$ where a disc can be placed. A move is legal if both conditions are met: First, the target cell is empty ($G_t(x, y) = 0$). Second, the placement results in the flipping of at least one opponent disc, while not being obstructed by immutable obstacles. If $\mathcal{A}_t$ is empty, the agent is forced to pass its turn. The interaction continues until the termination condition, where neither player has a legal move. Note that the full-disc board also results in termination, as neither player has space to play their disc.

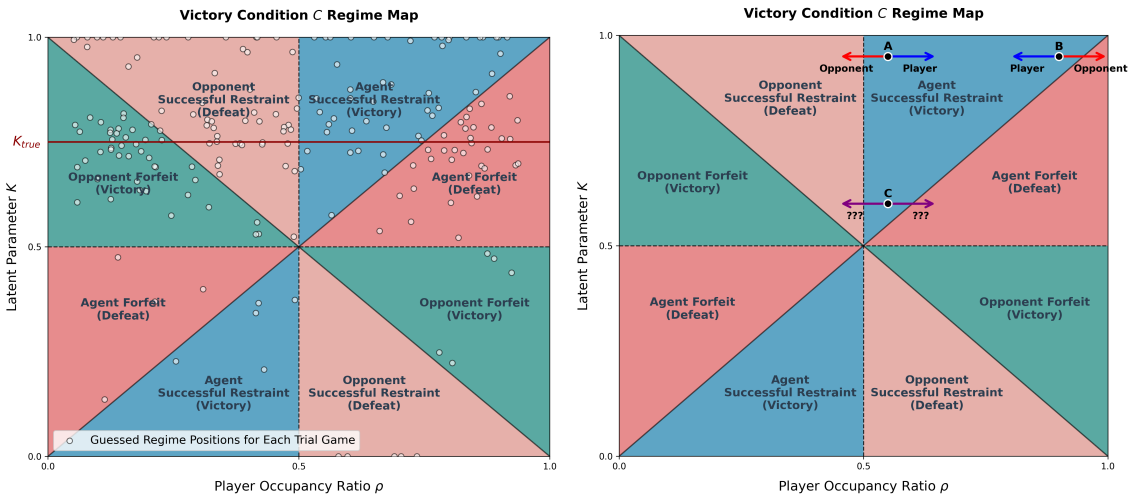


Figure 1: The regime map of the victory condition space $\mathcal{C}$ with respect to the latent parameter K and the player disc occupancy ratio ρ, and the two major goals of the Expanded Othello AI Arena. (Left) The intelligent system I should inductively infer the latent winning condition via trials, with the interactions bounded to N_{games} games. (Right) ToM should be considered to gain victory. For instance, in A and B, the strategies for the player and the opponent remain logically clear. However, for C, there is a dilemma of whether remaining in narrow successful restraint intervals with smaller ρ modifications or trying to gain safer forfeit victory conditions with significant ρ modifications; such actions are dependent on the opponent's strategy, as they share the same dilemma.

Asymmetric Observability: The protocol enforces a strict boundary between observable spatial data and latent objective data to evaluate the agents' online system identification and ToM capabilities to infer opponent behaviors:

- **Observable Input**: The agent receives the full grid G_t as input, providing complete information regarding the layout L with obstacle distributions and the current positions of all discs. Alongside the spatial observations, the set of legal action positions $\mathcal{A}_t$ is also given to the agent.
- **Latent Constraint**: The winning threshold K and the resulting victory regime are never explicitly provided to the agent. The agent should operate under the assumption that the objective function is unknown and may vary between environment instances.

Feedback Mechanism: To prevent the agent from relying on dense rewards that may convey cues about the latent condition of the environment, the Expanded Othello provides a sparse terminal reward system. Throughout the episode, all intermediate rewards are given as zero. Only at the terminal state T, the agent receives the trinary reward $R_i \in \{+1, 0, -1\}$, calculated based on the final occupancy ratio ρ and the latent parameter K as defined in Section 2.3.

This protocol ensures that any performance gain is a direct result of the agent's ability to correlate its actions and the resulting spatial configurations with the sparse terminal signals, effectively measuring its efficiency in deciphering latent strategic constraints.

3 Standardized Benchmarking Setup

To rigorously evaluate the adaptation capabilities of AI agents, we establish a standardized benchmarking framework. This setup is designed to move beyond static optimization, focusing instead on the efficiency gap — the disparity between rapid system identification and asymptotic convergence under severe interaction budget constraints.

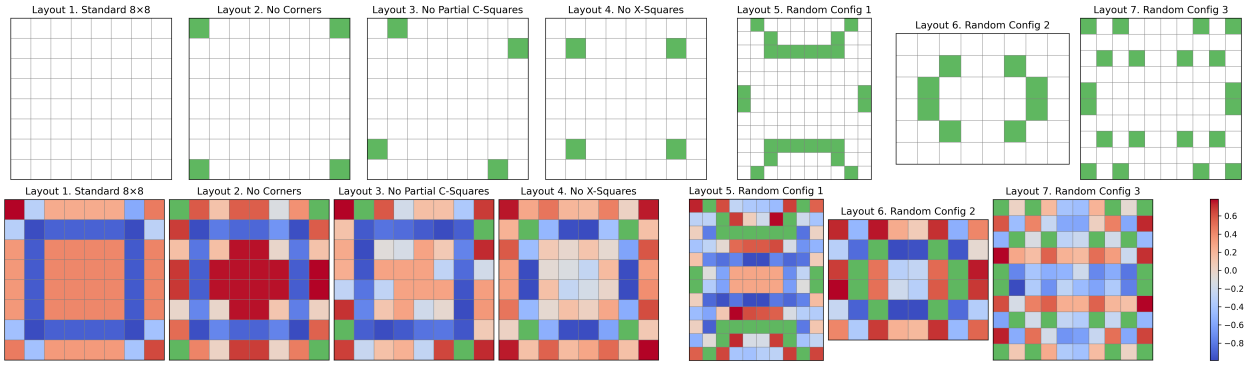


Figure 2: (Top) Layouts of the proposed standard benchmarking environments. (Bottom) The positional value tables generated via neuroevolutionary positional heuristics (PosNet) in standard condition ($K > 1.0$), where higher values are marked as red and lower values are marked as blue. Obstacles are marked as green.

3.1 Benchmarking Instances

The evaluation suite consists of 56 independent environmental instances, derived from a systematic 7×8 cross-product of spatial and objective variables. This ensemble ensures a comprehensive assessment across diverse topological and teleological challenges.

Layout Diversity ($\mathcal{L}$): We employ 7 distinct board layouts (L) as illustrated in Figure 2 (top). These include: (1) a standard 8×8 grid without obstacles; (2) obstacles on corners; (3) obstacles on partial C-squares (horizontal and diagonal neighbors of corners); (4) obstacles on X-squares (diagonal neighbors of corners); and (5-7) randomly generated obstacle distributions with varying board sizes. These irregular boundaries disrupt traditional topology-dependent heuristics, forcing agents to generalize spatial value-mapping.

Latent Strategic Objectives ($\mathcal{C}$): For each layout, we apply 8 distinct victory conditions (C). These include the standard Othello ($K > 1.0$), the inverse Othello ($K < 0$), and various narrow-interval regimes ($K \in \{0.8, 0.6, 0.4, 0.2\}$). Additionally, we include time-constrained variants of 10-turn limits with $K > 1.0$ and $K < 0$ conditions to evaluate adaptation under shorter match configurations.

3.2 Evaluation Constraints

To focus on sample efficiency rather than asymptotic convergence, we enforce strict operational limits:

Interaction Budget: Each agent is allocated a total of 2,000 episodes per environment ($N_{games} = 2000$). Within this limit, the agent must perform both latent rule discovery and policy optimization.

Knowledge Reset: Each of the 56 sessions is independent. All learned parameters or inferred knowledge on the evaluation environment are reset upon transitioning to a new environment, ensuring the measurement of immediate online adaptation.

Prior Knowledge: As the benchmark evaluates online adaptation to latent objectives, the winning condition C of each evaluation instance must be discovered solely through the 2,000-game interaction budget. However, spatial prior knowledge about board topologies (L) may be acquired in advance, as obstacle placement are a standard variation one could assume—for example, by meta-learning positional heuristics across a distribution of layouts, as demonstrated by our PosNet baseline (Section 4.2). Similarly, the use of general-purpose foundation models (e.g., LLMs) that encode broad reasoning capabilities is permitted, as such knowledge is not specific to the evaluation instances.

4 Baselines

To evaluate skill-acquisition efficiency in the Expanded Othello AI Arena, we establish a set of baselines. The defining challenge of this benchmark is the latency of the victory condition C. Because the objective

is unknown, defining a computable fitness function or a dense reward signal is fundamentally impossible. Consequently, we focus on survival-only evolution as the primary mechanism for both prior knowledge acquisition and online adaptation.

4.1 Survival-Only Evolution Framework

The core of our approach is the evolutionary framework driven solely by terminal match outcomes (Victory/Defeat) following the conditions of Section 2.3.2. In this pure-survival scheme, agents do not receive intermediate rewards or gradient information. Instead, they undergo a comparative-selection process: individuals within a population compete in matches, where the parameters of winners are duplicated and mutated while those of losers are discarded. This ensures that adaptation is driven strictly by survival in the face of latent environmental rules, providing a robust path to strategic discovery when explicit fitness functions are unavailable. The detailed procedure for survival-only evolution is in Appendix B.

4.2 Meta-Learning of Spatial Priors

The historical strategies for the classical Othello have been heavily dependent on the topology of the board (Rose, 2005). For instance, corners have been considered as the most important cells, as the discs on the corners cannot be flipped. It is widely accepted that the cells in the grid of the board do not share equal values. Consequently, the early artificial agents for Othello adopted the idea of a positional value table (Yoshioka et al., 1999), a static table that scores each cell's value.

Expanding upon this concept, we seek a universal value table generator conditioned by the layout L. We propose the Positional Network (PosNet), a fully-convolutional network designed to generate these positional value tables as a foundation for spatial reasoning across diverse geometries. Crucially, as the winning condition C is latent and an explicit fitness function cannot be defined, PosNet is trained via survival-only neuroevolution. It meta-learns generalized spatial expertise across a wide distribution of 1,000 randomly generated layouts with C conditioned as $K > 1.0$. In our benchmarking framework, PosNet serves as a non-adaptive baseline (Positional Agent) by remaining static during the 2,000-episode interaction phase, providing a rigorous measure of the performance limits of pure spatial expertise in the absence of objective-level adaptation. The detailed algorithm for PosNet is detailed in Appendix C.

4.3 Adaptive Minimax Agent

Building upon the architectural principles of high-performance Othello agents like Logistello (Buro, 1995), we propose an adaptive Minimax agent capable of real-time objective identification under latent winning conditions. The agent performs a depth-limited tree search where Minimax state nodes s are evaluated by a dynamic utility function $U(s)$, defined as:

$$U(s) = w_{pos}(s) \cdot \text{PosScore}(\text{PosNet}(L), s) + \sum_{i=1}^{3} w_i(s) \cdot f_i(s) \tag{4}$$

where $f_i(s)$ represents fundamental sub-utilities including mobility, cornerstone occupancy, and player-opponent disc difference. The positional score (PosScore) is calculated through a weighted piece counter (WPC), which utilizes the positional value table generated by PosNet and the placements of the discs to score the current board state.

Traditional minimax agents rely on static, hand-crafted utility weights; however, such weights cannot be utilized when the environmental condition C is latent, forcing a dynamically trainable utility function. To address such an aspect, our agent employs WeightNet, a neural network designed to output the optimal weight vector $\mathbf{w}(s) = [w_{pos}(s), w_1(s), w_2(s), w_3(s)]$ conditioned on the current state. WeightNet takes the current game progress and disc occupancy as inputs, allowing the agent to adjust its strategic priorities dynamically as the game evolves. Similar to the PosNet, as the victory condition C is latent and an explicit fitness function cannot be defined, WeightNet is trained via survival-only neuroevolution during the 2,000-episode interaction

phase. Technical details regarding the WeightNet architecture and the evolutionary operators are provided in Appendix D.

4.4 Comparative Baselines

To provide a comprehensive evaluation of our adaptive approach, we include several comparative baselines that represent different levels of environmental knowledge and learning efficiency. These baselines serve as reference points to measure the relative success of survival-based adaptation against traditional search and learning methods.

Monte Carlo Tree Search (MCTS) (Coulom, 2006) functions as a computation-intensive, simulation-based baseline in our Arena. Unlike other agents that must infer the latent winning condition C from sparse terminal signals, MCTS implicitly samples C by performing extended simulated playouts to anticipate potential outcomes. Because these simulations are governed by the true environment rules, MCTS gains statistical access to the reward structure through rollout-based look-ahead during its search process. This makes MCTS a strong empirical reference point, illustrating the level of strategic performance achievable when the objective can be sampled via extensive simulation. We use MCTS with versions of 30, 50, and 100 simulations per move, and detailed specifications of the MCTS implementation are provided in Appendix E.

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is evaluated to quantify the efficiency gap as a representative of modern gradient-based RL. In our experiments, each PPO agent is trained independently for each specific environment for a total of 10,000 games, as PPO agents fail to achieve meaningful performance under 2,000-game budget constraints. PPO often requires dense feedback to optimize its policy network; however, in our Arena, the latent nature of the objectives and the sparsity of rewards pose significant challenges for convergence. By providing PPO with five times the interaction budget of our adaptive minimax agent, we illustrate that RL struggles to perform rapid system identification and strategic convergence under latent rules and sparse rewards. Technical specifications regarding the PPO architecture and reward shaping are detailed in Appendix F.

Finally, we utilize the Random agents, which select actions uniformly at random, given valid action sets. This baseline works as the performance lower limit.

5 Results and Analysis

5.1 Evaluation of Meta-Learned Spatial Priors

The core intelligence of our framework is rooted in generalized prior expertise regarding board topology, as represented by the meta-learned PosNet. Setting aside the complexities of adapting to the latent winning condition $\mathcal{C}$, we first demonstrate that a generalized spatial mapping of strategic importance across diverse board topologies $\mathcal{L}$ is effectively achievable.

Figure 2 illustrates that PosNet is capable of gaining a generalizable expertise in spatial prior using only terminal rewards, showing logical value tables of prioritizing stable positions and penalizing risky positions. In Layout 1 (classical Othello board), PosNet prioritizes the corners most, which are the non-flippable positions. Furthermore, PosNet penalizes X-squares and the C-Squares, which are the neighbors of corners that risk the opponent getting corner cells when the player's discs are placed on them. Such results meet the general strategic consensus of Othello (Rose, 2005).

Furthermore, PosNet demonstrates remarkable generalization to the other layouts. A representative example of this topological reasoning is observed in Layout 2 (No Corners). Unlike the classical board, where both C-squares and X-Squares are penalized, in Layout 2, the C-squares become non-flippable positions, while X-squares still remain risky, as placing discs on X-squares still allows opponents to take C-squares. The PosNet generates the value table that logically follows such changes, as C-squares are prioritized and the X-squares remain penalized.

Board Layout	**Random**	**PosNet**	**MCTS-30**	**MCTS-50**	**MCTS-100**
1 (Standard 8×8)	72.50 ± 36.06	62.50 ± 36.35	65.00 ± 39.19	61.25 ± 37.20	64.38 ± 40.57
2 (No Corners)	68.13 ± 37.88	58.75 ± 37.39	53.75 ± 31.93	53.13 ± 31.27	60.63 ± 27.83
3 (No Partial C-squares)	67.50 ± 34.43	61.88 ± 42.08	48.75 ± 30.33	48.13 ± 27.89	42.50 ± 24.78
4 (No X-squares)	65.63 ± 34.27	62.50 ± 44.32	52.50 ± 31.96	56.25 ± 30.68	46.25 ± 28.13
5 (Random Configuration 1)	71.25 ± 31.14	60.63 ± 42.38	43.75 ± 29.12	31.25 ± 30.21	33.75 ± 22.48
6 (Random Configuration 2)	58.13 ± 32.40	43.75 ± 41.73	38.75 ± 25.04	35.63 ± 23.21	25.00 ± 21.71
7 (Random Configuration 3)	64.38 ± 29.09	62.50 ± 35.36	53.13 ± 24.78	45.63 ± 22.11	40.63 ± 21.29
Winning Condition	**Random**	**PosNet**	**MCTS-30**	**MCTS-50**	**MCTS-100**
1 ($K > 1.0$, Standard Othello)	95.71 ± 7.32	78.57 ± 29.82	85.00 ± 10.80	84.29 ± 17.18	78.57 ± 17.96
2 ($K = 0.8$, Narrow-Majority)	54.29 ± 17.66	47.86 ± 24.98	47.14 ± 15.24	44.29 ± 18.13	44.29 ± 25.24
3 ($K = 0.6$, Narrow-Majority)	19.29 ± 10.97	19.29 ± 22.44	14.29 ± 10.58	9.29 ± 7.32	13.57 ± 14.35
4 ($K = 0.4$, Narrow-Minority)	18.57 ± 12.49	3.57 ± 6.27	7.14 ± 8.09	10.71 ± 7.87	8.57 ± 12.49
5 ($K = 0.2$, Narrow-Minority)	77.86 ± 13.18	56.43 ± 31.85	60.71 ± 22.63	48.57 ± 20.96	45.00 ± 19.79
6 ($K < 0.0$, Inverse Othello)	94.29 ± 5.35	100.00 ± 0.00	65.71 ± 26.84	59.29 ± 29.22	55.71 ± 24.40
7 ($K > 1.0$, 10 turns limit)	90.00 ± 5.77	83.57 ± 23.22	65.00 ± 15.81	64.29 ± 15.12	56.43 ± 25.45
8 ($K < 0.0$, 10 turns limit)	84.29 ± 8.86	82.14 ± 23.78	61.43 ± 18.19	57.86 ± 14.68	55.71 ± 17.42

Table 1: Winning rate of the adaptive Minimax agents against various baselines after the training is finished for each environment E. The Minimax agent is trained for a 2,000-game session in each evaluation environment. The results show the mean winning rate averaged over each L (top) and C (bottom), tested on the 56 benchmarking instances. For each environment E, 20 games are played, where the side is swapped after 10 games.

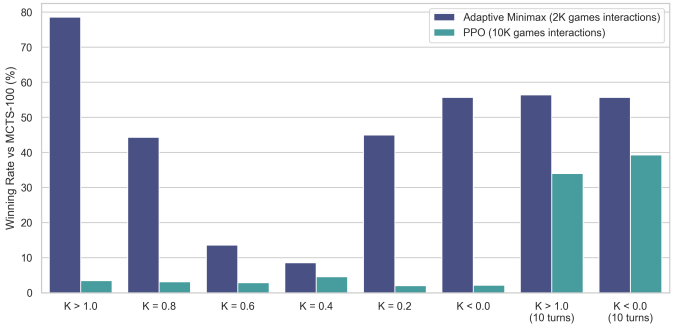


Figure 3: Winning rates of the adaptive minimax agent and the PPO agent against the MCTS-100 agent, the strongest baseline. The results show the mean winning rate averaged over each C, tested on the 56 benchmarking instances.

5.2 Adaptation Performance

The primary objective of the Expanded Othello AI Arena is to evaluate the skill-acquisition efficiency of agents, i.e., the ability of strategic convergence under a strictly constrained interaction budget. As conceptualized in the interaction framework of Figure 1, the defining challenge is deciphering the latent winning condition C from sparse terminal signals in an adversarial setting. Our empirical results reveal a substantial efficiency gap between adaptive Minimax and traditional gradient-based reinforcement learning, as illustrated in Figure 3. Despite receiving an interaction budget of 10,000 games — five times that of the adaptive Minimax agent — the PPO agent consistently fails to achieve meaningful convergence in environments with latent rules. This failure underscores the limitations of gradient-based optimization in zero-information settings where sparse rewards provide an insufficient landscape for policy updates in a restricted budget. While our adaptive minimax agent effectively performs system identification within the 2,000-game limit for standard and inverse regimes, the MCTS-100 baseline serves as a compute-intensive search reference, using repeated rollouts to approximate the latent win condition through terminal outcome statistics.

However, the adaptive Minimax agent encounters significant challenges in narrow-interval regimes, where its winning rate declines as the victory condition C becomes increasingly constrained ($K = 0.4$ and $K = 0.6$). This performance degradation primarily stems from the lack of a functional theory of mind, which is essential for adversarial inducement — the strategic concession to force an opponent into a forfeit. In these regimes, the agent must model the opponent's perspective to engineer forfeit traps; however, the adaptive Minimax fails to achieve this against the established baselines for different reasons. Against the MCTS-100 baseline, the agent struggles because MCTS is able to statistically look ahead to the results to some degree through its extended rollout-based simulation approach. Conversely, against Random or PosNet agents, insufficient strategic rationality in the opponents' moves prevents the adaptive agent from establishing a reliable model of their intentions. Since these baselines do not behave with sufficient rationality, the adaptive agent cannot effectively induce specific occupancy ratios, highlighting that true strategic mastery in latent environments requires the recursive modeling of both rational and stochastic adversaries.

6 Future Directions for Research with the Expanded Othello AI Arena

The empirical analysis in Section 5 demonstrated that while the adaptive Minimax agent achieves competitive performance in several environments, it fails in narrow victory regimes due to the absence of a functional ToM. Building upon these findings, this section outlines future research directions that can be facilitated by the Expanded Othello AI Arena.

LLM-Based Rapid Strategy Synthesis via TTT and ICL. Recent successes on AGI benchmarks such as ARC-AGI (Chollet et al., 2024; 2025) highlight LLMs' ability to rapidly adapt to novel abstract rules, leveraging their vast prior knowledge and symbolic reasoning capabilities (Wei et al., 2022; Achiam et al., 2023; Chollet et al., 2025). This motivates investigating LLM-based methods in the Expanded Othello AI Arena. Identifying unseen layouts (L) and latent conditions (C) through minimal interaction aligns naturally with paradigms such as test-time training (TTT) and in-context learning (ICL). TTT, which optimizes parameters during inference, has proven effective for handling distribution shifts and long-horizon sequence modeling through specialized TTT-layers (Sun et al., 2020; 2024). Meanwhile, ICL enables few-shot adaptation without explicit gradient updates, facilitating LLM-based decision-making process (Raparthy et al., 2023; Feng et al., 2025). By treating each game trial trajectory and terminal reward as a demonstration sample, TTT and ICL can be utilized to achieve the rapid adaptation scheme of the Expanded Othello AI Arena.

Universal Strategic Representations through World Models. The world model is a predictive representation of environment dynamics that internalizes state transitions and rewards within a latent space to facilitate long-term planning (Ha & Schmidhuber, 2018a; Hafner et al., 2019; 2023). These architectures have demonstrated high sample efficiency and superior performance in complex domains by allowing agents to optimize strategies through latent imagination rather than continuous environmental interaction (Schrittwieser et al., 2020; Kaiser et al., 2019). The Expanded Othello AI Arena offers a specialized testbed for world models thanks to the explicit separation of board topologies L and latent victory conditions C, where such a structure necessitates world models that can distinguish between invariant environmental dynamics and varying objectives.

7 Related Works

7.1 Benchmarks for Artificial General Intelligence

A prominent recent direction in AGI-oriented evaluation is to emphasize few-shot skill acquisition and novel problem solving rather than broad memorized knowledge. The ARC benchmark was introduced to measure a human-like form of fluid intelligence via abstract grid-based tasks solvable from a small number of examples under explicit priors (Chollet, 2019). Building on this philosophy, ARC-AGI (Chollet et al., 2024) became a widely studied benchmark for measuring progress on rapid generalization, and the ARC Prize technical reports document the evolving landscape of approaches—often combining program induction, search, and test-time adaptation—along with limitations of the original task set. ARC-AGI-2 (Chollet et al., 2025) was then proposed as a next-generation iteration intended to increase difficulty and provide a finer-grained stress test for frontier reasoning systems while remaining solvable for humans. Most recently, ARC-AGI-3 (ARC

Prize Foundation, 2026) shifts the evaluation from static input–output puzzles toward interactive, game-like environments that require exploration and adaptation, explicitly targeting agentic generalization beyond one-shot prediction.

In parallel, many widely used general capability benchmarks target complementary axes such as breadth of academic knowledge (Hendrycks et al., 2020), broad task coverage across diverse skills (Srivastava et al., 2023), holistic multi-metric evaluation across scenarios and risk (Liang et al., 2022), or expert-written, high-difficulty science QA designed to resist superficial retrieval (Rein et al., 2024). Relative to these primarily static question-answering or puzzle settings, the ARC-AGI series underscores generalization under novelty, and ARC-AGI-3 in particular highlights the importance of interactive adaptation. Motivated by the same generalization goal but in a competitive zero-sum game decision-making setting with the concept of ToM, our Expanded Othello Framework introduces latent objectives under adversarial play to directly measure how efficiently an agent can identify the objective and acquire effective strategies within a strict interaction budget.

7.2 Adaptive Game Playing under Latent Rules and Outcome-Only Feedback

A substantial body of work studies agents that can operate across many games rather than being tailored to a single fixed specification. In General Game Playing (GGP), agents receive an explicit formal game description and must quickly synthesize effective play, motivating standardized languages, infrastructures, and competitions (Genesereth et al., 2005). Similarly, the General Video Game AI (GVGAI) framework evaluates agents across unseen games and emphasizes general-purpose decision-making and planning under limited game-specific knowledge (Perez-Liebana et al., 2019). General game systems such as Ludii further broaden the scope of cross-game evaluation by providing efficient, human-readable game descriptions and a large library of games (Piette et al., 2019).

Complementary to general playing given rules, another line of research considers the rule discovery: inferring governing mechanics or objectives from observations of gameplays (Björnsson, 2012). On the algorithmic side, classical game AI has long combined search with learned or tuned evaluation components (Samuel, 1959), and even in board games such as Othello, prior work explored evolving neural networks to focus minimax search (Moriarty & Miikkulainen, 1994); more broadly, competitive coevolution provides a natural framing for adversarial settings where improvement is measured relative to opponents rather than dense intermediate supervision (Klijn & Eiben, 2021). Building on these perspectives, the Expanded Othello Framework instantiates a controlled family of adversarial environments where the board layout is observed, but the win condition is latent, enabling direct measurement of how efficiently an agent can identify the objective and acquire effective strategies under a strict interaction budget.

8 Conclusions

The Expanded Othello AI Arena focuses on the rapid strategic convergence concept by imposing a strict 2,000-game interaction budget. By decoupling the observable layout (L) from the latent winning conditions (C), the framework formalizes and parameterizes various expansions of Othello within a unified structure. Furthermore, it compels agents to discern valid victory paths in real-time based solely on sparse terminal feedback. Particularly in critical regimes, the presence of the forfeit mechanic necessitates a strategic pivot, requiring agents to move beyond simple occupancy maximization and instead employ complex tactics that induce opponent violations of the hidden threshold.

To further enhance the benchmark's discriminative power, the environment can be expanded to perturb the fundamental mechanisms currently held as constants. While the Arena successfully varies layouts and winning thresholds, the core flipping logic and alternating turn structure remain fixed parameters. Future iterations can introduce consecutive turns or modified obstacle rules, such as allowing discs to be flipped through immutable barriers, to test an agent's ability to adapt immediately to shifting environmental physics. Such expansions can be utilized for verifying whether a system can maintain logical decision-making under arbitrary transition dynamics without relying on a fixed game substrate.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

ARC Prize Foundation. Arc-agi-3. `https://arcprize.org/arc-agi/3/`, 2026. Accessed 2026-02-11.

Yngvi Björnsson. Learning rules of simplified boardgames by observing. In *ECAI 2012*, pp. 175–180. IOS Press, 2012.

Michael Buro. Logistello: A strong learning othello program. In *19th annual conference gesellschaft für Klassifikation eV*, volume 2. Citeseer, 1995.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pp. 72–83. Springer, 2006.

Yuanjun Feng, Vivek Choudhary, and Yash Raj Shrestha. Noise, adaptation, and strategy: Assessing llm fidelity in decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7704–7717, 2025.

Michael Genesereth, Nathaniel Love, and Barney Pell. General game playing: Overview of the aaai competition. *AI magazine*, 26(2):62–72, 2005.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. 2018b. doi: 10.5281/ZENODO.1207631. URL `https://zenodo.org/record/1207631`.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Yao Mu, Hongyuan Zhang, Wenqi Shao, and Ping Luo. Text2world: Benchmarking large language models for symbolic world model generation. *arXiv preprint arXiv:2502.13092*, 2025.

Wojciech Jaśkowski. Systematic n-tuple networks for othello position evaluation. *ICGA Journal*, 37(2):85–96, 2014.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

Daan Klijn and AE Eiben. A coevolutionary approach to deep multi-agent reinforcement learning. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 283–284, 2021.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Simon M Lucas. Learning to play othello with n-tuple systems. *Australian Journal of Intelligent Information Processing*, 4:1–20, 2008.

David E Moriarty and Risto Miikkulainen. Evolving neural networks to focus minimax search. In *AAAI*, pp. 1371–1377, 1994.

Diego Perez-Liebana, Jialin Liu, Ahmed Khalifa, Raluca D Gaina, Julian Togelius, and Simon M Lucas. General video game ai: A multitrack framework for evaluating agents, games, and content generation algorithms. *IEEE Transactions on Games*, 11(3):195–214, 2019.

Long Phan, Mantas Mazeika, Andy Zou, and Dan Hendrycks. Textquests: How good are llms at text-based video games? *arXiv preprint arXiv:2507.23701*, 2025.

Eric Piette, Dennis JNJ Soemers, Matthew Stephenson, Chiara F Sironi, Mark HM Winands, and Cameron Browne. Ludii–the ludemic general game system. *arXiv preprint arXiv:1905.05013*, 2019.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

Sharath Chandra Raparthy, Eric Hambro, Robert Kirk, Mikael Henaff, and Roberta Raileanu. Generalization to new sequential decision making tasks with in-context learning. *arXiv preprint arXiv:2312.03801*, 2023.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Brian Rose. *Othello: A Minute to Learn... A Lifetime to Master*. Anjar Co., 2005.

Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Kuttler, Edward Grefenstette, and Tim Rocktäschel. Minihack the planet: A sandbox for open-ended reinforcement learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=skFwlyefkWJ`.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 3762–3780, 2022.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Marcin Szubert, Wojciech Jaśkowski, and Krzysztof Krawiec. On scalability, generalization, and hybridization of coevolutionary learning: a case study for othello. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(3):214–226, 2013.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26275–26285, 2024.

Taku Yoshioka, Shin Ishii, and Minoru Ito. Strategy acquisition for the game othello based on reinforcement learning. *IEICE TRANSACTIONS on Information and Systems*, 82(12):1618–1626, 1999.

Binggang Zhuo and Masaki Murata. Utilizing gpt-4 to solve textworld commonsense games efficiently. In *Proceedings of the 10th Workshop on Games and Natural Language Processing@ LREC-COLING 2024*, pp. 76–84, 2024.

	C					C	
C	X					X	C
			W	B			
			B	W			
C	X					X	C
	C					C	

1.00	-0.25	0.10	0.05	0.05	0.10	-0.25	1.00
-0.25	-0.25	0.01	0.01	0.01	0.01	-0.25	-0.25
0.10	0.01	0.05	0.02	0.02	0.05	0.01	0.10
0.05	0.01	0.02	0.01	0.01	0.02	0.01	0.05
0.05	0.01	0.02	0.01	0.01	0.02	0.01	0.05
0.10	0.01	0.05	0.02	0.02	0.05	0.01	0.10
-0.25	-0.25	0.01	0.01	0.01	0.01	-0.25	-0.25
1.00	-0.25	0.10	0.05	0.05	0.10	-0.25	1.00

Figure 4: (Left) Locations of the C-squares (C), X-squares (X), and the initial black (B) and white (W) disc positions. (Right) The positional value table of the standard weighted piece counter heuristic agent created by Yoshioka et al. (1999).

A Strategies and Heuristics for Classical Othello

A.1 Human Strategies for Othello

There exists a set of generally accepted consensus strategies among the Othello experts, which is used as the fundamental guides for the classical Othello game (Rose, 2005). A central concept is the *stability*, finding the non-flippable cells. Most notably, discs placed on corners are not flippable for the rest of the game, and corner control often anchors stable edges that are difficult to challenge. Accordingly, strong play prioritizes securing corners when possible, while carefully avoiding moves that allow opponents to gain the corners.

Two well-known risky regions are the *C-squares* and *X-squares*, which illustrated in Figure 4 (Left). The C-square is the orthogonally adjacent cell to a corner, and the X-square is the diagonally adjacent cell. Placing a disc on these cells is often dangerous because it can enable the opponent to immediately capture the corner. As a result, C- and X-squares are typically penalized unless the corresponding corner is already secured or a concrete tactical justification exists.

Another major principle is *mobility* — the number of legal moves available. Maintaining high mobility while restricting the opponent's mobility improves positional control and reduces the risk of being forced into unfavorable moves, especially those that open access to corners and stable edges. On the other hand, early-game disc count is not a reliable objective as Othello shows high board state changes, due to its signature flipping mechanism. Hence, strong players often accept temporary material loss to preserve mobility and long-term stability, while disc advantage becomes decisive mainly in the late game.

A.2 Heuristic Agents

Before modern learning-based agents, strong Othello play was commonly built on hand-designed (or statistically fitted) evaluation functions coupled with adversarial search. A canonical family is the *positional heuristic* that assigns a fixed weight to each board square and evaluates a position by a weighted sum over occupied squares, often called a *weighted piece counter* (WPC), which calculates a value of the state s given the positional value table V, following Equation 5. Note that $s_{i,j} = 1$ if player disc is on the cell (i, j), -1 if opponent's disc is on the cell, and 0 if empty.

$$\mathrm{WPC}(V, s) = \sum_{i=1}^{W} \sum_{j=1}^{H} V_{i,j} s_{i,j} \tag{5}$$

Classic positional value tables typically emphasize stable regions such as corners and penalize risky near-corner squares (e.g., C- and X-squares), encoding well-known human principles in a simple linear form (Yoshioka et al., 1999). Despite its simplicity, WPC-style evaluation has served as a long-standing baseline and a convenient expert proxy in Othello research, especially when paired with shallow minimax search.

A major milestone in heuristic Othello was *Logistello*, which demonstrated that high-level performance can be achieved by combining deep search with evaluation functions learned from data. Logistello popularized constructing evaluation functions from large collections of game positions and outcomes, fitting parameters statistically rather than solely relying on hand-tuned tables (Buro, 1995). In practice, Logistello-style approaches can be viewed as a progression from fixed positional heuristics (static WPC tables) to richer feature sets and statistically optimized weights, with search (e.g., minimax/alpha–beta) acting as the decision-time optimizer. This line of work remains directly relevant to our setting because it clarifies what is achievable with strong *positional* priors and search-based planning, even before introducing more adaptive mechanisms.

B Detailed Procedures of Survival-Only Evolutionary Learning

This appendix details the survival-only evolutionary procedure used throughout our framework. The key constraint is that the winning condition C is latent, so we cannot define a reliable scalar fitness or shaped reward. Hence, the only trustworthy signal is the terminal match outcome. Accordingly, we adopt the *survival-only* selection mechanism in which individuals compete against each other and only the winners are duplicated, followed by mutation and reshuffling. This procedure is used (i) to meta-learn PosNet parameters over layout distributions, and (ii) to train environment-specific WeightNet parameters of the adaptive Minimax within the 2,000-game interaction budget. Algorithm 1 summarizes the survival-only evolution, which is explained in detail from Section B.1 to Section B.4.

Algorithm 1 Survival-only evolution

Require: Population size $2N$, generations G, mutation scale σ, environment E
Initialize population $P = \{\theta_i\}_{i=1}^{2N}$
for $g = 1$ to G **do**
  Instantiate agents $\{A_i\}_{i=1}^{2N}$ from $\{\theta_i\}$ in environment E
  Pair individuals $(i, i+N)$ for $i = 1, \ldots, N$
  For each pair, play two games with swapped colors and compute $\{\mathrm{match}(A_i, A_{i+N})\}_{i=1}^{N}$
  Construct selected multiset S by Eq. (6)
  Mutate: $P_{\mathrm{mut}} = \{\theta'_i + \epsilon_i\}$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$
  Shuffle: $P \leftarrow \mathrm{Shuffle}(P_{\mathrm{mut}})$
end for
return final population P (to be ensembled/normalized as in Appendix B.3)
=0

B.1 Survival-Only Selection via Paired Matches

Let $P = \{\theta_i\}_{i=1}^{2N}$ denote a population of size $2N$ (even-sized for pairing). Each θ_i parameterizes an instantiated agent (e.g., neural network parameter weights of PosNet and WeightNet). We evaluate individuals via paired matches. Let A_i denote an agent with θ_i. Then, for each pair $(i, i+N)$, we play two games between A_i and A_{i+N}, with swapped colors to reduce first-move bias. Let $\mathrm{match}(A_i, A_j) \in \{1, 0, -1\}$ be the outcome from A_i's perspective (victory/draw/defeat). We then construct a *selected* multiset S of size $2N$ by duplicating

the winner:

$$S = \bigcup_{i=1}^{N} \begin{cases} \{\theta_i, \theta_i\} & \text{if match}(A_i, A_{i+N}) = 1, \\ \{\theta_{i+N}, \theta_{i+N}\} & \text{if match}(A_i, A_{i+N}) = -1, \\ \{\theta_i, \theta_{i+N}\} & \text{if match}(A_i, A_{i+N}) = 0. \end{cases} \tag{6}$$

Intuitively, this is a pure-survival scheme; selection pressure is induced only by head-to-head outcomes, without assuming any explicit fitness function.

B.2 Mutation, Shuffle, and Generational Update

After selection, each retained parameter vector is mutated by additive Gaussian noise:

$$P_{\text{mut}} = \{\theta'_i + \epsilon_i\}_{i=1}^{2N}, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2 I). \tag{7}$$

We then randomly shuffle the mutated population to form the next generation

$$P \leftarrow \text{Shuffle}(P_{\text{mut}}), \tag{8}$$

and repeat the cycle (agent instantiation → paired matches → comparison-only selection → mutation → shuffle) for G generations.

B.3 Ensembling and Normalization at Inference

We use a population ensemble at the end of evolution to reduce variance within the agents in the population P. The concrete aggregation depends on which object is being evolved:

PosNet / positional value tables. Each individual θ_i induces a value table V_i over the board. We normalize each table by affine rescaling so that its minimum becomes -1 and maximum becomes 1, then average the normalized tables to obtain the final table $\bar{V}$ used by the positional baseline (PosNet).

WeightNet / utility re-weighting. Each individual θ_i induces a WeightNet that outputs a weight vector $w_i(s)$ for a given state s. For stability, we normalize each vector by an L_1-style normalization, $\text{Normalize}(w) = w/\sum_k |w_k|$, and use the mean normalized weights $\bar{w}(s) = \frac{1}{2N}\sum_{i=1}^{2N} \text{Normalize}(w_i(s))$ over the final population.

B.4 Budget accounting and default hyperparameters

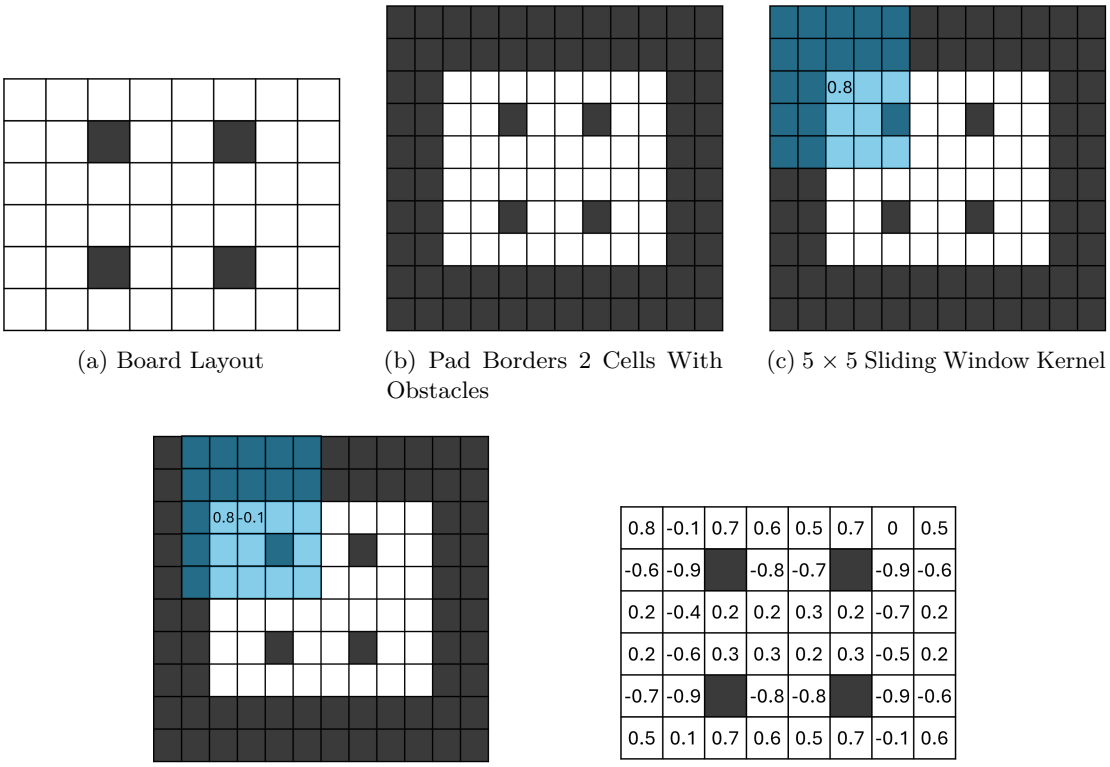
All training in a single environment E is constrained by a fixed interaction budget ($N_{games} = 2000$). For the adaptive Minimax, this corresponds to $G = 20$ generations with population size $N = 50$ (thus $2N = 100$ individuals), where each selection pairing uses two games (side-swapped) per match, resulting in 100 games per generation, yielding 2,000 total games across 20 generations.

C Detailed Procedures of PosNet

This section explains the detailed procedure of crafting the positional network (PosNet) to meta-learn spatial prior knowledge from diverse layouts (L). Figure 5 shows an example procedure of the value table (V) generation using PosNet.

C.1 Network Architecture and Input Representation

The positional network (PosNet) is implemented as a lightweight fully-convolutional network (FCN), specifically engineered to generate positional value tables V for any board layout $L \in \mathcal{L}$. Unlike traditional Othello agents that rely on fixed 8×8 tables, the FCN structure allows PosNet to process boards of arbitrary geometries — such as those found in our random configuration layouts (Layouts 5–7) — by applying the same learned spatial filters across the entire grid.



(a) Board Layout

(b) Pad Borders 2 Cells With Obstacles

(c) 5×5 Sliding Window Kernel

(d) Repeat for Every Blank Position

0.8	-0.1	0.7	0.6	0.5	0.7	0	0.5
-0.6	-0.9		-0.8	-0.7		-0.9	-0.6
0.2	-0.4	0.2	0.2	0.3	0.2	-0.7	0.2
0.2	-0.6	0.3	0.3	0.2	0.3	-0.5	0.2
-0.7	-0.9		-0.8	-0.8		-0.9	-0.6
0.5	0.1	0.7	0.6	0.5	0.7	-0.1	0.6

(e) Finished Positional Value Table

Figure 5: An example procedure of the positional value table generation.

The architecture comprises two convolutional layers with 5×5 kernels, followed by a final 1×1 convolutional layer. We purposefully selected a 5×5 kernel size based on the following strategic and technical rationales:

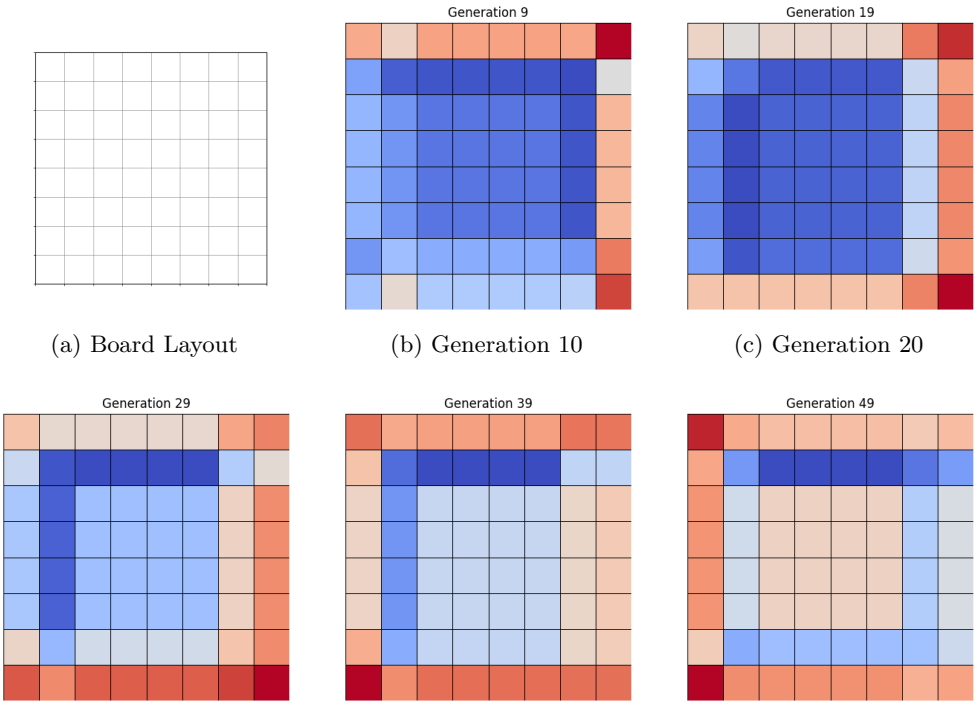
- **Local Context for Heuristic Identification**: In classical Othello, the strategic value of a square is primarily defined by its proximity to corners and edges. The most critical squares (C-squares and X-squares) are located exactly 1 cell (orthogonally or diagonally) away from a corner. Furthermore, as the expert-generated value table in Figure 4 (Right) shows, the major cell value differences occur within the 2 cell distances from the corners or the edges. The 5×5 window provides a 2-cell radius, which is the minimum sufficient area to perceive a corner or edge from any of these critical positions.

- **Inductive Bias of Spatial Stability**: Human expertise (Rose, 2005) suggests that stability — the inability of a disc to be flipped — is the most important factor in positional valuing. Since stability is a property determined by the local topology (e.g., being in a corner), the 5×5 kernel acts as an inductive bias that forces the network to learn these local patterns rather than memorizing global board states.

- **Architectural Compatibility for Variable Geometries**: By utilizing a fully-convolutional architecture devoid of fixed-size components such as fully-connected layers, PosNet remains inherently invariant to the input dimensions. This design ensures that the meta-learned spatial filters can be applied zero-shot to any board size ($W \times H$) or the placement of the obstacles.

To ensure consistent behavior at the board boundaries, we apply a 2-cell padding of obstacles around the input. This ensures that the 5×5 kernel always has full context, even when evaluating squares on the very edge of the board, effectively treating out-of-bound cells as impassable obstacles.

C.2 Stabilization via Orthogonal Transformation



(a) Board Layout (b) Generation 10 (c) Generation 20

(d) Generation 30 (e) Generation 40 (f) Generation 50

Figure 6: Positional Value Table from PosNet trained separately on the standard Othello (8×8) environment ($K > 1.0$), without meta-learning or stabilization process. A population size of 50 is used, and the agents are trained for 50 generations. Red means higher positional values, while blue means lower values. The drifting effect is evidently shown; directional imbalance can be observed.

To counteract the *drifting* effect (e.g., Figure 6) — where comparison-only selection leads to unstable directional biases — we employ a rigorous stabilization protocol.

- **Symmetric Nature of Othello**: Conceptually, an Othello board has no inherent orientation. Hence, PosNet's output should ideally be equivariant to orthogonal transformations. Consider an orthogonal transformation F from the eight-fold symmetry group (D4 dihedral group), which consists of rotations and reflections. Due to orthogonal transformation equivalence, Equation 9 should hold.

$$F(\text{PosNet}(L)) = \text{PosNet}(F(L)) \tag{9}$$

  Then, for some orthogonal transformation F and its inverse function F^{-1}, Equation 10 should hold.

$$\text{PosNet}(L) = F^{-1}(\text{PosNet}(F(L))) \tag{10}$$

- **Training and Inference Protocol**: During meta-learning, each agent in the population is evaluated under a randomly chosen transformation F, forcing the weights to generalize across all orientations. Hence, the positional value table V is generated as Equation 11.

$$V_{\text{train}} = F^{-1}(\text{PosNet}(F(L))) \tag{11}$$

At inference time, we apply all eight transformations to the layout L, pass them through PosNet, and average the inverse-transformed outputs to produce a symmetrical value table, following Equation 12.

$$V_{\text{infer}} = \frac{1}{8}\sum_{i=1}^{8} F_i^{-1}(\text{PosNet}(F_i(L))) \tag{12}$$

C.3 Meta-Learning and Spatial Generalization

PosNet evolves via a meta-learning scheme, using the survival-only evolution over a distribution of 1,000 randomly generated board layouts. By keeping the winning condition fixed ($K > 1.0$) during this phase to isolate the spatial reasoning from the influence of variant winning conditions (C), the network learns a universal spatial prior — interpretation of stable cells and risky cells on an arbitrary layout L — which then serves as both a robust foundation for the adaptive Minimax agent during environment-specific adaptation and the non-adapting positional baseline in the experiments. Note that the PosNet is trained for 1,000 generations with a population size of 200.

D Detailed Procedures of Adaptive Minimax

D.1 Adaptive Utility Function and Sub-Utilities

Motivated by Logistello (Buro, 1995), the proposed adaptive agent baseline utilizes a Minimax search algorithm with alpha-beta pruning, governed by a dynamic utility function $U(s)$. To enable adaptation to unknown winning conditions C, the utility is defined as a learnable linear combination of four fundamental sub-utility functions following Equation 13, which is a formal definition of Equation 4.

$$U(s) = w_{\text{pos}} \cdot \text{PosScore}(\text{PosNet}(L), s) + w_{\text{mobility}} \cdot \text{Mobility}(s) + w_{\text{corner}} \cdot \text{Corner}(s) + w_{\text{disc}} \cdot \text{discDiff}(s) \tag{13}$$

The weights $\mathbf{w} = [w_{\text{pos}}, w_{\text{mobility}}, w_{\text{corner}}, w_{\text{disc}}]$ are dynamically generated by WeightNet based on the game state. The sub-utilities u are defined as follows:

- **PosScore**: The positional value score of state S, calculated through WPC, using the value table generated by the meta-learned PosNet, following Equation 14. Note that $s_{i,j} = 0$ for obstacles.

$$\text{PosScore}(\text{PosNet}(L), s) = \frac{1}{n_{agent} + n_{opponent}} \text{WPC}(\text{PosNet}(L), s) \tag{14}$$

- **Mobility**: The difference between the number of legal moves available to the agent and the opponent. Let $\phi_i(s)$ denote the number of valid moves for player i in state s. Then the mobility is defined as Equation 15.

$$\text{Mobility}(s) = \frac{\phi_{agent}(s) - \phi_{opponent}(s)}{\phi_{agent}(s) + \phi_{opponent}(s)} \tag{15}$$

- **Corner**: The difference between the number of corners occupied by the agent and the opponent, following Equation 16.

$$\text{Corner}(s) = \frac{\#\text{Corners Occupied by Agent} - \#\text{Corners Occupied by Opponent}}{4} \tag{16}$$

- **discDiff**: The difference between the number of the agent's discs and the opponent's discs, following Equation 17.

$$\text{discDiff}(s) = \frac{n_{agent} - n_{opponent}}{n_{agent} + n_{opponent}} \tag{17}$$

To maintain numerical stability across diverse board configurations, all sub-utilities are normalized to a range of $[-1, 1]$.

D.2 WeightNet and State Representation for Strategy Adaptation

WeightNet is a shallow MLP designed for rapid strategy synthesis within the restricted 2,000-game interaction budget. Its objective is to map the temporal and adversarial context to gain an optimal weight vector $\mathbf{w}$, following Equation 18.

$$\mathbf{w} = \text{WeightNet}(\text{progress}, \rho, 1 - \rho) \tag{18}$$

The input features provide a functional basis for environmental adaptation:

- **Progress**: The ratio of total discs to the board area, representing the progress of the game, defined as Equation 19.

$$\text{progress} = \frac{n_{agent} + n_{opponent}}{WH} \tag{19}$$

- **Occupancy ratios (ρ, $1 - \rho$)**: Following Equation 3, ρ represents the agent's disc occupancy and $1 - \rho$ represents the opponent's disc occupancy.

Such state representation works as a simplified, primitive foundational basis for adversarial reasoning.

D.3 Online Adaptation via Neuroevolution

WeightNet is trained independently for each environment E using the survival-only neuroevolutionary framework. For each environment E, the WeightNet is trained with a population size of 100, for 20 generations, i.e., $2000 = \#\text{generations} \times \#\text{games per generation} = 20 \times (100 \times 0.5 \times 2)$. During the training phase, the agent instantiation is done via generating an adaptive Minimax agent for each WeightNet in the population.

The adaptive Minimax uses a depth-3 Minimax search during training for computational efficiency. For the evaluation, the depth is increased to depth-5 to rigorously assess the converged policy π_E with the ensemble formed by averaging the normalized weights of the population from the final generation.

E Specifications of MCTS agents

E.1 Simulation-Based Look-Ahead Baseline

In the Expanded Othello AI Arena, Monte Carlo Tree Search (MCTS) serves as a computation-intensive, rollout-based reference baseline. While other adaptive agents in this benchmark must inductively infer the latent winning condition C from sparse terminal signals, the MCTS agent is provided with a simulation environment that explicitly utilizes the true parameters of the environment, including the latent threshold K. By performing extended rollouts governed by these actual rules, MCTS gains statistical look-ahead into the terminal rewards during its search process. This allows it to approximate effective strategies through repeated sampling of the true outcome space, rather than through implicit rule identification from interaction history.

E.2 Tree Search Mechanism

The MCTS agents follow the standard four-stage cycle to navigate the game tree:

1. **Selection**: Starting from the root, the agent traverses the tree by selecting child nodes that maximize the upper confidence bound for trees (UCT) (Kocsis & Szepesvári, 2006), following Equation 20.

$$\text{UCT}(s, a) = \bar{Q}(s, a) + c\sqrt{\frac{\ln N(s)}{n(s, a)}} \tag{20}$$

where $\bar{Q}(s, a)$ is the average reward (i.e., mean win rate in case of the Expanded Othello) from simulations, $N(s)$ is the total visit count of the parent, $n(s, a)$ is the visit count of the child, and c is the exploration constant.

2. **Expansion**: If a selected node is not terminal and has unvisited legal moves, a new child node is added to the tree.
3. **Rollout**: From the expanded node, the agent performs a random playout to evaluate potential trajectories. These simulations are executed using the environment's ground-truth rules and K, enabling the agent to sample potential rewards from the resulting state transitions during the search process.
4. **Backpropagation**: The final outcome is propagated back up the tree to update the mean win rates ($\bar{Q}(s, a)$) and node visit counts of all ancestral nodes.

E.3 Implementation Details

To provide diverse reference points for computational budgets, we evaluate three variants: MCTS-30, MCTS-50, and MCTS-100, where the number denotes the total simulations performed per move. For the exploration constant, we set $c = 1.4 \approx \sqrt{2}$. For the rollout depth limit, each simulation is limited to a maximum depth of 30 steps to ensure computational efficiency. If the game does not terminate within this limit, the state is evaluated based on the current occupancy ratio ρ relative to K. At the conclusion of the search, the agent selects the action a that maximizes the visit count $N(s, a)$ rather than the average win rate, ensuring robustness against stochastic simulation noise and outliers.

F Specifications of PPO agents

The Proximal Policy Optimization (PPO) agents are implemented using the `stable-baselines3` (Raffin et al., 2021) library with the default `MlpPolicy` configuration. A unique PPO agent is initialized and trained independently for each environment E in the Arena to ensure that no cross-environmental transfer or prior knowledge influences the learning process. Each agent is trained against a *RandomAgent* for a total of 10,000 games. The reward structure is strictly terminal, with a reward of 1.0 for a victory, -1.0 for a defeat, and 0.0 for a draw, provided only at the conclusion of a game. This configuration serves as a baseline to measure the performance of standard model-free reinforcement learning without simulation capabilities or prior knowledge.

G Further Analysis of the Results

In this section, we will further discuss whether the adaptive Minimax tries to adapt to narrow-Regime cases and find the reason behind its failure, via analyzing the change of the output $\mathbf{w}$ with respect to the player disc occupancy ρ in various environments. Note that the results are gained using the trained WeightNet by interpolating ρ.

Figure 7 and 8 show the output $\mathbf{w}$ of WeightNet in layout 1 (Standard 8×8 without obstacles) and layout 2 (Obstacles on corners). It is shown that the WeightNet is trained differently for different K. For instance, in layout 1, $K > 1.0$ showed high positive importance on corners and mobility, while low importance on PosScore. On the other hand, for $K < 0$, PosScore has high positive importance, and the discDiff and Corner have high negative importance (having low values is critical), while mobility is shown to have less importance than in the $K > 1.0$ case.

For narrow regimes, if the adaptive Minimax is able to adapt well to the forfeit conditions, the results should show steep slopes as the weights should be sensitive to the occupancy ρ. It is shown that for some cases, such as mobility increase on $K = 0.2$ in Figure 7 and PosScore decrease in $K = 0.6$ of Figure 8. We can expect that the Minimax agent is able to get some gist of adaptation to narrow regimes. However, for many cases like $K = 0.4$ of Figure 7 and $K = 0.2$ of Figure 8, there is no evident weight changes with respect to ρ, meaning that the WeightNet was not able to learn any useful relationship between ρ and victory condition C. Such non-adaptive results are likely to support the low performances in Table 1.

One plausible speculation on why the WeightNet is able to learn something in only some of the narrow-regime environments is the nature of evolutionary methods, especially due to the survival-only selection. Since the survival-only selection happens between Minimax agents, there is no guarantee that an agent with a

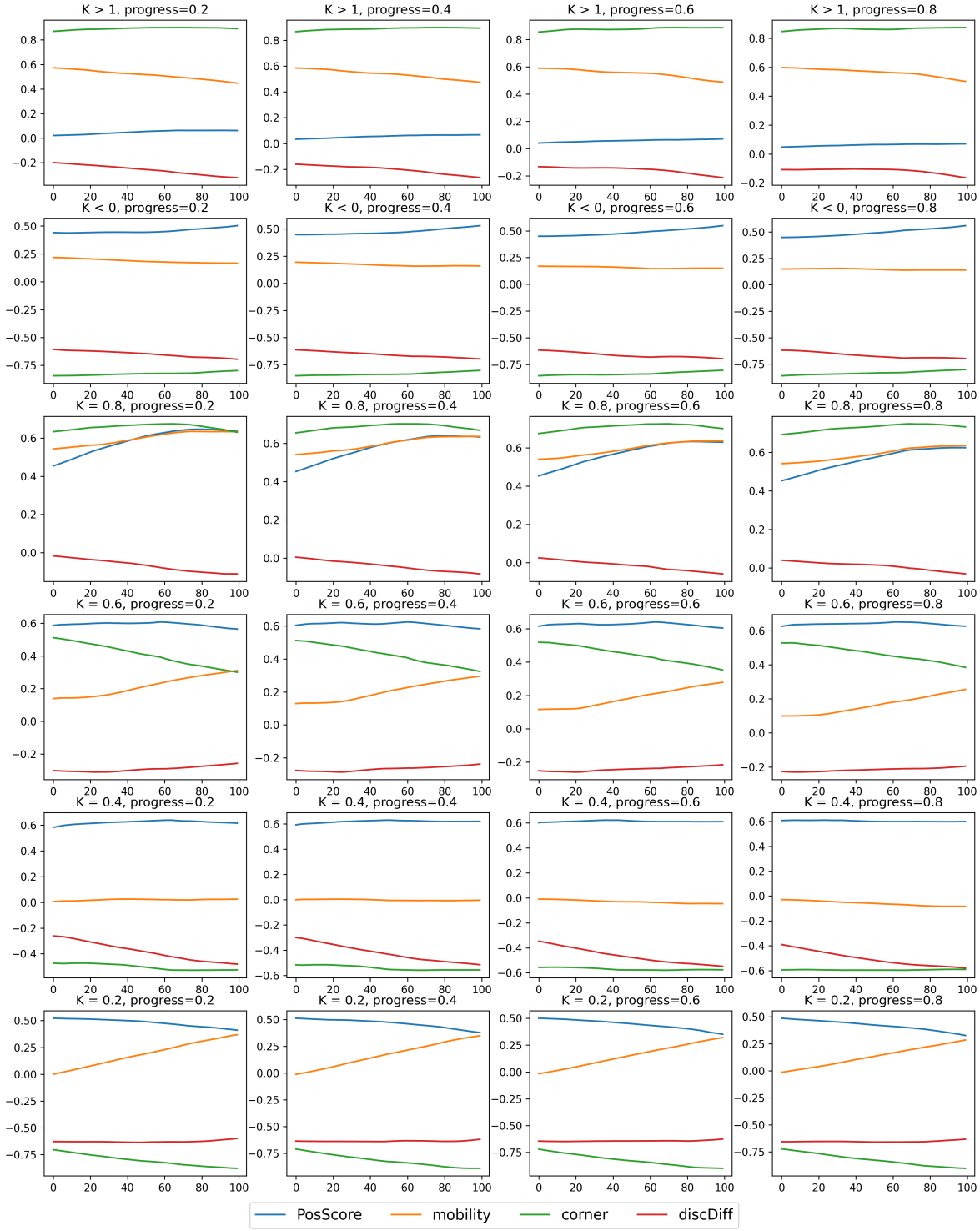


Figure 7: Output weights of the WeightNet with respect to player disc occupancy ρ, on various K and progresses, in layout 1 (Standard 8×8 without obstacles)

well-trained WeightNet will survive; luck is a critical factor for adaptation, especially considering the rather small population size (100) for such a complex and difficult environment.

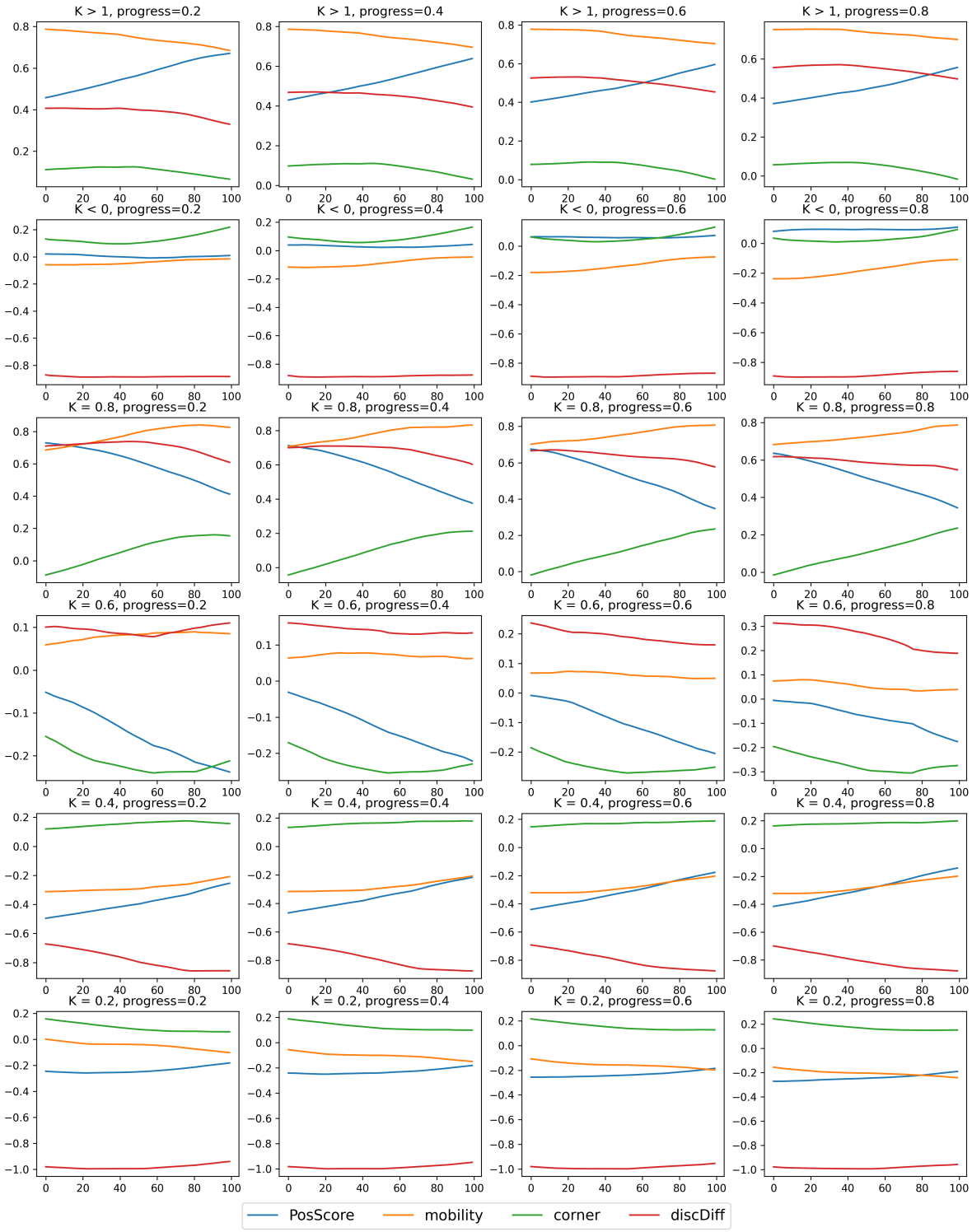


Figure 8: Output weights of the WeightNet with respect to player disc occupancy ρ, on various K and progresses, in layout 2 (Obstacles on corners)