IMPLICIT STRATEGIC OPTIMIZATION: PREDICTION-AWARE LONG-HORIZON LEARNING FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning effective strategies in imperfect-information games requires agents to account for delayed consequences while adapting to partially observed opponent behavior. Although these interactions often contain recurring strategic regimes, the relevant context may become clear only after an action has been taken. Extending the reward horizon alone does not resolve how experience should be assigned when the context predicted before acting was incorrect. To address this problem, we introduce Implicit Strategic Optimization (ISO), a prediction-aware framework for long-horizon learning in large language model (LLM) agents. ISO combines context prediction, delayed-value estimation, and routed policy optimization. Its central mechanism separates the context used for acting from the context used for updating: a predictor guides action selection from pre-action interaction history, while context labels logged after decisions determine which context-conditioned policy components receive updates. A Strategic Reward Model estimates delayed returns from trajectory segments to guide routed group relative policy optimization. We analyze this routing principle in an idealized latent-context repeated game, deriving a contextual regret bound that separates context-specific initialization, prediction errors, and within-context variation. Experiments in No-Limit Texas Hold'em and competitive Pokémon demonstrate improved strategic performance over supervised fine-tuning and reinforcement learning baselines. Ablations support the contributions of both delayed-value estimation and context routing, while evaluations on unseen and hybrid poker opponents show that the return advantage extends beyond the training population. Behavioral analyses further associate the gains with decisions that trade immediate payoff for future strategic value. These findings support prediction-aware context routing as a useful mechanism for long-horizon learning in strategic LLM agents.

1 INTRODUCTION

LLM agents can acquire game-playing skills through supervised fine-tuning and reinforcement learning, yet favorable local outcomes do not necessarily translate into long-horizon strategic return. In No-Limit Texas Hold'em (NLHE), many small wins can be offset by a few large losses; in competitive Pokémon, preserving a weakened team member can prevent a more valuable endgame. These examples illustrate why effective strategic learning requires feedback about the consequences that unfold after a decision. Longer-horizon rewards provide such feedback, but leave a further challenge: assigning experience to the appropriate context-conditioned policy component when the strategic context was uncertain at action time.

We study this challenge through the recurring structure of strategic interaction. Apparent non-stationarity can reflect changes among opponent behaviors or game phases that the agent encounters repeatedly. Before acting, the agent must predict the relevant context from its available history, whereas subsequent interaction may provide clearer evidence. If the update remains attached to the component selected by the original prediction, a mistaken assignment persists despite that later evidence. Retrospective assignment uses the additional evidence to direct learning to the relevant component. When contexts recur, that component can reuse the experience on a later visit, even if it did not generate the original action.

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

To address this assignment problem, we introduce *Implicit Strategic Optimization* (ISO), which combines context prediction, a Strategic Reward Model (SRM), and routed group relative policy optimization (GRPO). A prediction selects the component used for acting; the SRM estimates delayed return; and a retrospective label selects the adapter receiving the route-specific update. The backbone is shared. This is a choice about training-sample assignment, not additional information available to the deployed policy. *Implicit* refers to learning strategic behavior without explicitly solving an equilibrium or constructing a hand-designed plan; the discrete implementation still uses engineered context labels.

The distinction is useful only if later evidence improves assignment enough to offset its costs. Reassigning an action creates a mismatch between its behavior distribution and its updating component. Noisy labels, insufficient behavior support, or interference through shared parameters may eliminate the benefit. We therefore compare ISO with multi-hand GRPO, a recurrent history-only policy, shared-SRM training, equal-capacity context features, and an otherwise identical routed policy that updates the acting component. The last comparison directly tests retrospective assignment: ISO improves poker return by 1.2 BB/100 hands, with a reported paired 95% interval of $[0.2, 2.2]$.

To explain when this separation can be useful, our analysis isolates the routing principle in an idealized repeated game with full loss-vector feedback. Contextual regret separates into the regrets of learners updated on realized-context subsequences and a prediction-mistake cost. An optimistic-learning specialization exposes a tradeoff between context fragmentation, prediction error, and residual variation. This explains the structural opportunity without asserting convergence of the neural optimizer. A learned soft-routing extension additionally tests the information-timing idea without environment-provided discrete labels.

The contributions are: (i) a concrete distinction between predicted-context acting and retrospective adapter updates, including the associated likelihood bookkeeping; (ii) a contextual-regret decomposition and its variation-sensitive specialization; and (iii) controlled evaluations in NLHE and Pokémon, with long-horizon controls, learned contexts, held-out opponents, and value-model shift. The strongest evidence concerns the incremental effect of update assignment under the tested protocols, supported by broader performance and behavioral analyses.

2 Related work

Context inference and modular policies. Recurrent meta-RL uses interaction history for adaptation (Duan et al., 2016); PEARL and VariBAD infer task information for control (Rakelly et al., 2019; Zintgraf et al., 2020). Soft modularization learns task-dependent routes through a shared policy (Yang et al., 2020). Context inference, memory, and modular capacity are therefore established ingredients. ISO examines whether later context evidence should *reassign an update* away from the component that generated the action. The history, feature, and predicted-to-predicted controls test this distinction; we do not claim that feature-conditioned networks lack the capacity to represent the same policy.

Hindsight and privileged supervision. Hindsight experience replay relabels goals (Andrychowicz et al., 2017); hindsight task relabelling extends this idea to sparse-reward meta-RL (Packer et al., 2021). Hindsight policy gradients derive importance-weighted updates for alternative goals (Rauber et al., 2019). Privileged-information training uses richer observations during learning than at deployment (Chen et al., 2020); AHEAD uses subsequent feedback and corrective hints to reweight LLM training signals (Jin et al., 2026). ISO shares this information asymmetry, while retaining environment payoff and assigning updates to strategic-context components. Its labels can depend on the sampled action and continuation, so its single-action ratio is a practical surrogate rather than a trajectory correction. Appendices B.3 and B.5 detail the distinction.

Optimization, strategic agents, and regret. ISO builds on proximal policy optimization (PPO) clipping and GRPO's group-relative scoring (Schulman et al., 2017; Shao et al., 2024). MARSHAL develops turn-level advantage estimation and agent-specific normalization for strategic self-play (Yuan et al., 2026). These mechanisms address the learning signal; ISO studies its context-specific recipient. Counterfactual regret minimization (CFR), Deep CFR, and population-based methods address complementary game-theoretic objectives (Zinkevich et al., 2007; Brown et al., 2019; Lanctot et al., 2017); language agents also combine opponent modeling and search in Pokémon (Karten et al.,

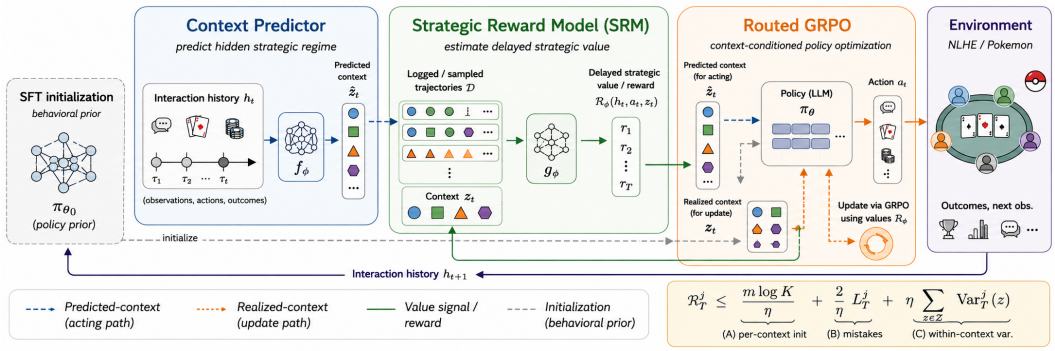


Figure 1: ISO overview. Supervised fine-tuning (SFT) initializes the policy; context prediction guides acting; the SRM estimates delayed value; and realized context assigns updates. The predictor and SRM correspond to q_ψ and R_ϕ in the text. The schematic bound uses $0 < \eta \leq 1$ and the startup-augmented variation defined in Appendix A.4; Equation (9) gives the sharper guarantee.

2025). Our return comparisons concern the evaluated population. The theoretical specialization uses established predictable-sequence and optimistic-learning ideas (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015), applied to realized-context subsequences.

3 Implicit Strategic Optimization

ISO organizes learning around three complementary roles: predicting the context for acting, estimating delayed value, and assigning the resulting update. These roles share a training record that distinguishes information available before a decision from evidence obtained afterward. Figure 1 and Algorithm 1 summarize how this distinction enters data collection, scoring, and policy updating.

3.1 Information Timing and Objective

Let h_t contain only information available before decision t: public state, the agent's private observation, legal actions, and past interaction. A training logger produces a context label $z_t \in \mathcal{Z} = \{1, \ldots, m\}$ after that decision. The label is a proxy for a payoff-relevant regime, not a fully observed game state. It may depend on subsequent observations, and is used for predictor supervision, value estimation, and update routing only. At deployment, the policy uses h_t and its own prediction $\widehat{z}_t$.

The objective is expected cumulative environment payoff over the specified interaction horizon. For a trajectory segment ending at $t + H$ and discount $0 \leq \gamma \leq 1$, the return target is

$$y_t = \sum_{\tau=t}^{t+H} \gamma^{\tau-t} r_\tau. \tag{1}$$

Episode boundaries terminate the sum unless the protocol explicitly continues across hands. Consequently, retaining cross-hand history and optimizing cross-hand return are distinct interventions. Neither a context label nor a strategic-sacrifice statistic replaces environment payoff as the evaluation objective.

3.2 Context Prediction and Delayed-Value Estimation

Given this information boundary, the predictor q_ψ learns to anticipate the logged context from (h_t, z_t) pairs. The SRM R_ϕ provides the complementary training signal by regressing delayed return from a history, an action, and a retrospective context:

$$\widehat{z}_t = \arg\max_z q_\psi(z \mid h_t), \qquad \mathcal{L}_{\text{ctx}} = -\mathbb{E}_{\mathcal{D}_{\text{ctx}}} \log q_\psi(z_t \mid h_t), \tag{2}$$

$$s_t = R_\phi(h_t, a_t, z_t), \qquad \mathcal{L}_{\text{SRM}} = \mathbb{E}_{\mathcal{D}_{\text{SRM}}} \big(R_\phi(h_t, a_t, z_t) - y_t\big)^2. \tag{3}$$

Training segments and validation segments must be disjoint; held-out segments are used to assess prediction and return ranking, not to fit the SRM. The predictor and SRM are fixed within each policy-optimization phase. A later phase may refresh the SRM with recent on-policy data (Section 5.4).

The discrete implementation uses simulator-derived proxies. In NLHE, these summarize aggression, showdown behavior, fold responses, and large-pot behavior across interactions. In Pokémon, they summarize battle phase, revealed opponent information, remaining teams, and recent decisions. These are engineered abstractions. They need not be unique or causally exogenous, and their usefulness depends on being predictable from h_t and informative about delayed returns.

3.3 Predicted-context acting and retrospective updates

Write $\theta = (\omega, \lambda_1, \ldots, \lambda_m)$ for a shared backbone and context-specific low-rank adapter parameters (Hu et al., 2022). The frozen behavior snapshot selects and logs

$$a_t \sim \pi_{\theta_{\text{old}}}(\cdot \mid h_t, \widehat{z}_t), \qquad b_t = \pi_{\theta_{\text{old}}}(a_t \mid h_t, \widehat{z}_t). \tag{4}$$

After collecting the continuation and label z_t, ISO assigns the sample to update route $u_t = z_t$. For candidates $i = 1, \ldots, G$ sampled from the same pre-action history, let $s_t^{(i)} = R_\phi(h_t, a_t^{(i)}, z_t^{(i)})$. Each candidate has its own continuation and resulting label. The group-relative advantage and cross-route likelihood ratio are

$$A_t^{(i)} = \frac{s_t^{(i)} - \bar{s}_t}{\text{std}_j(s_t^{(j)}) + \epsilon_A}, \qquad \rho_t^{(i)}(\theta) = \frac{\pi_\theta(a_t^{(i)} \mid h_t, u_t^{(i)})}{\pi_{\theta_{\text{old}}}(a_t^{(i)} \mid h_t, \widehat{z}_t)}. \tag{5}$$

The behavior denominator is never recomputed under u_t. Using a clipped PPO-style surrogate (Schulman et al., 2017; Shao et al., 2024), we maximize

$$\begin{aligned} \mathcal{J}(\theta) = \mathbb{E}_{t,i}\Big[\min\Big\{ \rho_t^{(i)} A_t^{(i)}, \text{clip}(\rho_t^{(i)}, 1-\epsilon_c, 1+\epsilon_c) A_t^{(i)} \Big\} \\ - \beta\, \text{KL}\Big(\pi_\theta(\cdot \mid h_t, u_t^{(i)}) \,\|\, \pi_{\theta_{\text{old}}}(\cdot \mid h_t, u_t^{(i)}) \Big) \Big]. \end{aligned} \tag{6}$$

Here $\epsilon_A > 0$ stabilizes normalization, $0 < \epsilon_c < 1$ sets the clipping range, and $\beta \geq 0$ weights regularization. Action probabilities refer to the same generated output and token mask under both routes. During a policy update, the logged denominator, labels, and scores are fixed. Gradients update ω and λ_{u_t}; other adapters receive no route-specific gradient, although their policies can change through the shared backbone. For $u_t \neq \widehat{z}_t$, $\rho_t(\theta_{\text{old}})$ need not equal one: clipping can act on route mismatch before any optimizer step.

This is a practical *surrogate*, not an unbiased policy-gradient claim. When $u_t \neq \widehat{z}_t$, the sample is off-policy for the updating component. Clipping and filtering low-support samples trade variance for bias. Moreover, a label derived from future interaction can depend on the action, so a policy-probability ratio alone does not correct all retrospective selection effects or trajectory-distribution shift. The full-information analysis in Section 4 isolates routing and does not remove these implementation limitations.

Update assignment across controls. Both feature conditioning and routing use the same pre-action information; the distinction in Table 1 concerns update assignment rather than policy expressivity. Predicted-to-predicted routing (P→P) and ISO (P→R) share their architecture, acting rule, and SRM, with $u_t = \widehat{z}_t$ and $u_t = z_t$, respectively. Oracle routing receives z_t at action time and serves only as a privileged diagnostic.

3.4 Learned and overlapping contexts

A finite bank can also represent overlapping regimes without allocating a separate full model to each context. In the learned soft-routing extension, a history encoder produces action-time responsibilities $p_t(z)$ over eight adapters; a post-trajectory target encoder produces retrospective update responsibilities $q_t(z)$. Neither encoder uses environment-provided context labels. Acting remains restricted to causal history, whereas the training target may use the completed trajectory. The shared backbone, sparse activation, and load balancing limit active cost and discourage expert collapse. The interpolation experiment in Section 5.3 evaluates this relaxation in a structured opponent mixture; it does not establish unrestricted latent-regime discovery. The analysis in Appendix A.6 treats mixtures of action distributions; the architectural distinction is specified in Appendix A.9.

Table 1: Action and update contexts for the routing controls. The denominator always belongs to the component that generated the sample; z_t is unavailable to ordinary acting policies.

Variant	Acting context	Update route u_t	Denominator route
Predicted-to-predicted (P→P)	$\widehat{z}_t$	$\widehat{z}_t$	$\widehat{z}_t$
ISO (P→R)	$\widehat{z}_t$	z_t	$\widehat{z}_t$
Oracle	z_t (privileged)	z_t	z_t

Algorithm 1 ISO training with a shared policy and routed updates

1: Initialize π_θ by supervised fine-tuning; fit q_ψ and R_ϕ on logged training data.
2: **for** each policy-optimization phase **do**
3: Freeze θ_{old}, q_ψ, and R_ϕ for data collection and scoring.
4: For each history h_t, predict $\widehat{z}_t$ and sample candidate actions under that route.
5: Store $(h_t, a_t^{(i)}, \widehat{z}_t, \log b_t^{(i)})$ under the actual behavior route.
6: Collect subsequent interaction; construct $z_t^{(i)}$ and the delayed-return target.
7: Score each $(h_t, a_t^{(i)}, z_t^{(i)})$ with the SRM; compute within-history advantages.
8: Assign $u_t^{(i)} = z_t^{(i)}$; re-evaluate the same output under this update route.
9: Apply the support filter and maximize Equation (6), updating shared and selected parameters.
10: Validate on separate trajectories; optionally refresh the SRM between phases.
11: **end for**

4 Prediction-aware contextual regret

The preceding algorithm motivates a simpler analysis that isolates the effect of routing from neural optimization and value estimation. We consider an idealized repeated game, fixing a player and suppressing its index. At round t, a context $Z_t \in \mathcal{Z}$ and opponents' play induce a loss vector $\ell_t \in [-1, 1]^K$. The player receives a prediction $\widehat{Z}_t$, acts, and then observes Z_t and the *full* vector ℓ_t. This is stronger feedback than the sampled trajectories available to a neural policy.

A routed learner maintains a loss-minimizing distribution $x_{t,z} \in \Delta_K$ for each context. It plays $w_t = x_{t,\widehat{Z}_t}$, updates the learner for Z_t with ℓ_t even on prediction mistakes, and leaves the other learners unchanged. For $\mathcal{F}_z = \{t : Z_t = z\}$, define

$$u_z^\star \in \arg\min_{u \in \Delta_K} \sum_{t \in \mathcal{F}_z} \langle u, \ell_t \rangle, \qquad R_T^{\text{ctx}} = \sum_{t=1}^{T} \langle w_t - u_{Z_t}^\star, \ell_t \rangle, \qquad L_T = \sum_{t=1}^{T} \mathbf{1}\{\widehat{Z}_t \neq Z_t\}. \tag{7}$$

The comparator may change across realized contexts, but is fixed within each context.

Proposition 4.1 (Routing decomposition)**.** *Suppose each base learner, updated on every round in its realized-context subsequence, satisfies* $\sum_{t \in \mathcal{F}_z} \langle x_{t,z} - u, \ell_t \rangle \leq B_z(T)$ *for every* $u \in \Delta_K$*. Then*

$$R_T^{\text{ctx}} \leq \sum_{z \in \mathcal{Z}_T} B_z(T) + 2L_T, \tag{8}$$

where $\mathcal{Z}_T$ *is the set of contexts visited by time* T*.*

Proof idea. Add and subtract the realized-context distribution x_{t,Z_t}. Its subsequence regrets form the first term; the remaining inner product is zero on a correct prediction and at most two otherwise. The complete argument and base-learner construction appear in Appendix A.

Optimistic entropic mirror descent. Each base learner uses optimistic entropic mirror descent, initialized uniformly and using the preceding loss on its context subsequence as its prediction (zero on the first visit). Let $t_z(r)$ enumerate $\mathcal{F}_z$, $s_T = |\mathcal{Z}_T|$, and $V_z^{(2)} = \sum_{r=2}^{|\mathcal{F}_z|} \|\ell_{t_z(r)} - \ell_{t_z(r-1)}\|_\infty^2$. The predictable-sequence guarantee (Rakhlin & Sridharan, 2013) and Proposition 4.1 give

$$R_T^{\text{ctx}} \leq \underbrace{s_T\left(\frac{\log K}{\eta} + \frac{\eta}{2}\right)}_{\text{context-specific initialization}} + \underbrace{2L_T}_{\text{context prediction}} + \underbrace{\frac{\eta}{2}\sum_z V_z^{(2)}}_{\text{within-context variation}}. \tag{9}$$

Appendix A gives the exact updates and a self-contained proof. An in-hindsight balancing choice yields $R_T^{\mathrm{ctx}} \le 2L_T + \sqrt{2s_T \log K \,(s_T + \sum_z V_z^{(2)})}$. For fixed m, K, choosing η on the order of $T^{-1/2}$ and using $\sum_z V_z^{(2)} = O(T)$ gives $R_T^{\mathrm{ctx}} = O(\sqrt{T}) + 2L_T$. Thus sublinear prediction mistakes suffice for sublinear regret; lower within-context variation can tighten the bound further. The guarantee applies to the stated idealized learner, not neural GRPO.

Effect of recurring contexts. For a single variation-adaptive learner, the bound depends on chronological loss variation. In Equation (9), variation is instead measured along context-specific subsequences, with additional initialization and prediction terms. For alternating contexts with losses $(0, 1)$ and $(1, 0)$, chronological variation is linear in T and within-context variation is zero. Accurate predictions allow each learner to reuse information from earlier visits. This example illustrates favorable structure for routing; it does not provide a lower bound for non-routed algorithms, since sufficiently expressive contextual or recurrent learners can also exploit that structure.

The comparators also differ: a static learner is assessed against one distribution, while Equation (7) competes with one per context. Smaller regret bounds against different comparators do not establish empirical dominance. ISO should help when regimes recur, prediction is informative, and the reduction in residual variation offsets fragmentation. Its neural realization shares parameters, uses estimated returns, and observes sampled feedback; the theorem establishes neither its convergence nor a Nash-equilibrium guarantee.

5 Experiments

We evaluate whether the routing principle translates into improved strategic performance in NLHE and competitive Pokémon. The controlled comparisons first separate retrospective update assignment from longer reward horizons, memory, and delayed-value estimation. We then examine learned contexts and unseen opponents, followed by diagnostics of prediction reliability, SRM distribution shift, and behavior.

5.1 Experimental setup and evaluation metrics

Environments. The primary benchmark is heterogeneous-table No-Limit Texas Hold'em (NLHE), with 100-big-blind (BB) starting stacks, blinds of $0.5/1$ BB, and common legal betting templates across methods. Competitive Pokémon uses the Gen 1 OU ruleset on Pokémon Showdown. The learned policies use a Qwen2.5-0.5B backbone initialized by supervised fine-tuning (SFT). The controlled poker results aggregate five random seeds and eight seating permutations; the broader reference evaluation uses 10,000 hands per configuration. The controlled comparisons share opponent draws and paired deal/seating schedules. Checkpoints are selected with a common held-out validation protocol rather than final test return.

Controlled long-horizon comparisons. Multi-Round GRPO (MR-GRPO) aggregates discounted payoff over 20 consecutive hands against the same opponent population, retaining causal cross-hand history. Each hand still resets to the same stack/blind rules. History-only recurrent GRPO tests memory; shared SRM-GRPO tests delayed-value modeling without routing; the context-feature control uses the predicted context without retrospective route assignment; and P→P keeps the routed architecture but updates the acting route. The matched study holds initialization, backbone, observations, action interface, interaction and optimizer-step budgets, evaluation, and total/active parameter budgets fixed. Oracle routing receives the retrospective context at action time and is evaluated only as a diagnostic.

Metrics. Poker long-term return is $\mathrm{LTR} = (100/N) \sum_{n=1}^{N} \Delta c_n / b$, where Δc_n is net chip profit and b is the big blind. This measures payoff magnitude; it does not by itself establish cross-hand planning. In Pokémon, we report battle win rate and Elo, together with strategic sacrifice rate (SSR), a diagnostic of locally costly actions associated with positive later team value. SSR is not a training target or a causal estimate. CFR-style methods optimize equilibrium-related robustness, which is a different objective from realized return against a fixed heterogeneous population. We therefore use them as reference points rather than evidence of game-theoretic dominance.

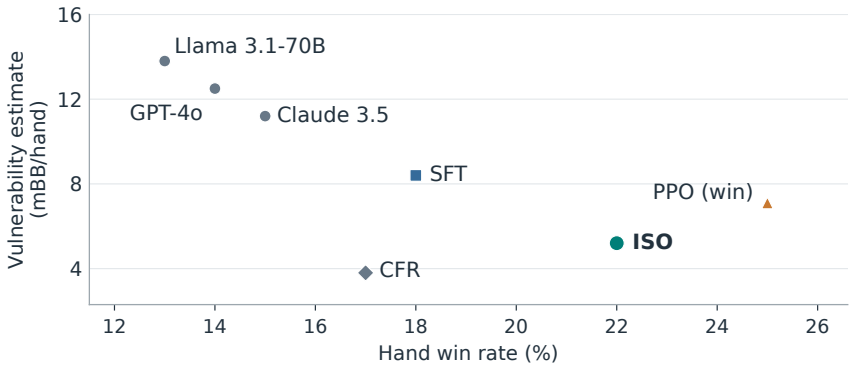


Figure 2: Win rate and robustness in the broader poker reference evaluation. Vulnerability is the protocol-specific estimate (mBB/hand), not certified NashConv. PPO uses win reward. Corresponding returns are in Table 10; neither axis alone measures long-horizon return.

Table 2: Controlled NLHE comparison. LTR is in BB/100 hands. Δ denotes ISO minus the row, with reported paired 95% confidence intervals where available. All rows belong to this matched protocol; values from the broader reference evaluation in Table 10 are kept separate.

Method	LTR ↑	Δ LTR [95% CI]
Per-hand profit GRPO	11.9	$+3.7$
MR-GRPO (20-hand return)	12.5	$+3.1$ $[1.8, 4.4]$
History-only recurrent GRPO	13.6	$+2.0$ $[0.7, 3.3]$
Shared SRM-GRPO, no routing	13.2	$+2.4$ $[1.0, 3.8]$
Equal-capacity context feature	13.3	$+2.3$ $[1.0, 3.6]$
Predicted-to-predicted routing (P→P)	14.4	$+1.2$ $[0.2, 2.2]$
ISO (P→R)	**15.6**	—
Oracle routing (privileged context)	16.6	-1.0

Figure 2 shows that hand win rate and estimated vulnerability rank methods differently in the reference protocol; the controlled routing comparison below evaluates return under a separate matched protocol.

5.2 Long-horizon performance and update routing

In Table 2, ISO improves return by 1.2 BB/100 over the otherwise identical P→P variant; the reported paired interval excludes zero. This comparison isolates retrospective update assignment within the routed architecture. The other controls show gains from a longer reward horizon, recurrent history, and SRM-based training. The context-feature control performs similarly to shared SRM-GRPO, while P→P improves further. These complementary comparisons support contributions from both the learning pipeline and retrospective assignment, but do not decompose the total gain additively or attribute all improvement over per-hand GRPO to routing.

Competitive Pokémon evaluation. In Pokémon, the full-episode GRPO baseline already uses feedback spanning the initial observation through the terminal outcome. ISO improves win rate and Elo over this baseline, accompanied by higher SSR (Table 3). The comparison evaluates the value and routing mechanism beyond access to terminal reward. The two domains provide evidence across game structures, without testing zero-shot policy transfer. These differences remain descriptive because confidence intervals were not reported.

Figure 3 further shows incremental gains across successive policy versions, with smaller gains against the immediate predecessor at later iterations.

Table 3: Competitive Pokémon under Gen 1 OU. The GRPO baseline already optimizes a complete multi-turn battle. SSR is a behavioral diagnostic, reported as a percentage.

Policy	Elo ↑	Win rate (%) ↑	SSR (%)
Base (no fine-tuning)	1510	40.0	6.1
SFT	1668	61.0	10.4
Full-episode GRPO	1705	64.0	12.1
ISO	**1750**	**68.0**	**15.2**

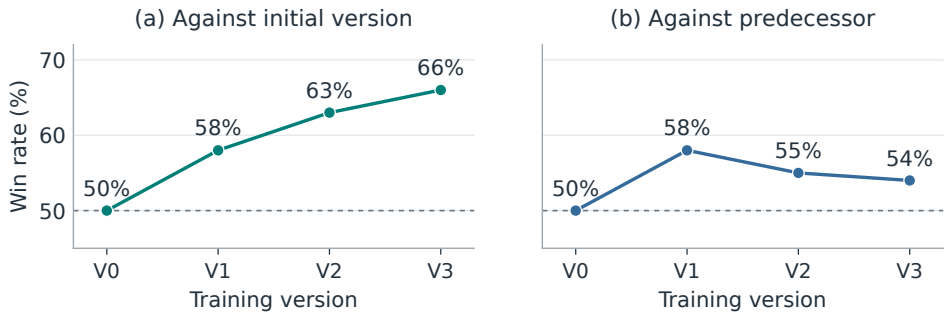


Figure 3: Iterative improvement in Pokémon. Left: win rate against the initial version. Right: win rate against the immediate predecessor; V0 marks the 50% reference. Labels give win rates, and the dotted lines mark 50%. These version-to-version comparisons use different opponents from Table 3.

5.3 Learned contexts and unseen opponents

To reduce dependence on a handcrafted discrete taxonomy, the learned extension in Section 3.4 is trained against continuously mixed opponents and evaluated at held-out mixture weights and on an unseen strategy family. Table 4 compares history-only learning, latent features, hard routing, and soft routing. Soft ISO improves over the latent-feature control on both splits, while remaining below an oracle supplied with the continuous context. The observed maximum/minimum expert-load ratio is 1.35; this is a utilization diagnostic, not a proof against representation collapse.

The discrete-context evaluation provides a complementary test on unseen and hybrid opponent populations (Table 13). ISO retains a return advantage even though all methods degrade on these splits. Together, the two evaluations suggest that useful context structure can extend beyond the training population, while leaving generalization to arbitrary new games and unconstrained regime discovery open.

5.4 Prediction reliability and delayed-value shift

Causal-history audit. Predictor inputs must end at the action boundary. In the poker audit, macro-F1 is 0.79 with the original feature set, 0.78 after removing identified post-decision fields, 0.76 under strict action-boundary truncation, and 0.69 on an opponent/session-disjoint split. Strict truncation is the relevant information constraint. These are predictor diagnostics, not a reported end-to-end rerun of every policy under each audit condition; a small change in F1 alone cannot certify the absence of leakage. Table 15 gives the available predictive and calibration metrics. Prediction corruption reduces ISO's return, but the non-routed and oracle results also change across corruption settings (Table 21); that sweep characterizes joint sensitivity rather than isolating a causal effect of routing error alone.

SRM distribution shift. As the policy changes, a frozen SRM trained on SFT trajectories loses ranking quality: held-out Spearman correlation declines from 0.46 initially to 0.34 at the latest reported checkpoint. Refreshing on a mixture of equal amounts of original and recent on-policy data recovers it to 0.42. The corresponding end-to-end comparison yields 15.6 BB/100 with a frozen SRM, 16.1 with refresh, and 16.0 with uncertainty-gated SRM/Monte Carlo mixing. These are modest

Table 4: Learned contexts in controlled opponent mixtures. No environment-provided context labels are used by the learned variants. LTR is in BB/100 hands; uncertainty intervals were not reported for this study.

Method	Held-out interpolation ↑	Unseen family ↑
History-only GRPO	11.4	9.2
Latent context as a feature	12.0	9.8
Unsupervised hard routing	12.5	10.2
Learned soft-routing ISO	**13.6**	**11.5**
Oracle continuous context	14.5	12.5

descriptive gains without reported confidence intervals. Refreshing mitigates, but does not eliminate, distribution shift; it also adds data collection, model fitting, and monitoring costs (Table 17).

Behavioral evidence. To connect the aggregate improvements to decisions during play, we examine the broader poker evaluation by betting street and pot size. ISO obtains larger contributions on the turn and river and positive returns in larger pots (Tables 18 and 19). ISO also takes more actions judged locally costly by a shallow evaluator, with a larger fraction associated with positive later returns (Table 20). These analyses are consistent with the delayed-value interpretation. Since actions, pot sizes, and later outcomes are selected jointly by play, they do not identify the counterfactual effect of any individual sacrifice.

6 Discussion and Limitations

The empirical results and the idealized analysis identify complementary roles for delayed-value estimation and context routing. The SRM determines how an observed action is scored, while routing determines which component learns from that score. This distinction explains the role of the matched controls: extending the reward horizon or supplying a predicted-context feature does not implement retrospective assignment. Conversely, assigning updates retrospectively still depends on the quality of the value signal. The observed advantage over the otherwise identical predicted-to-predicted variant supports the specific assignment mechanism, while the broader comparisons evaluate the combined framework.

ISO's advantage depends on recurring, payoff-relevant contexts that can be predicted from causal history. The discrete version requires domain-specific label construction; the learned version has only been tested on structured opponent mixtures. Weakly identifiable, rapidly changing, or overly fragmented contexts may erase the benefit. Shared parameters can interfere across routes, and retrospective labels can be action-dependent. Cross-route clipping and support filtering do not guarantee unbiased learning. An offline SRM can become miscalibrated as the policy changes, while refreshing it introduces additional complexity.

The regret analysis assumes full loss-vector feedback and base learners with stated guarantees; those conditions are not established for the neural implementation. The experiments cover two adversarial game protocols, with limited uncertainty reporting for the learned-context and Pokémon studies. Detailed compute-efficiency comparisons remain open: parameter and interaction matching do not establish equal end-to-end cost. These limits constrain claims of broader applicability.

7 Conclusion

We introduced ISO, a prediction-aware framework that combines context prediction, delayed-value estimation, and routed policy optimization for strategic LLM agents. Its central mechanism uses the context predicted before a decision to select actions and later context evidence to assign updates. The contextual-regret analysis explains how recurring structure can be useful when prediction error and within-context variation are controlled. Experiments in NLHE and competitive Pokémon, together with matched routing controls and behavioral analyses, support the value of this mechanism in the evaluated settings. Extending it to less structured contexts and maintaining reliable value estimates under stronger distribution shift remain directions for future work.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting, editing and translation, literature search, refinement of the optimization description and mathematical proofs, interpretation of author-supplied results, and preparation of figures and LaTeX. The reported empirical results originate from the authors' experiments, and the authors take responsibility for the final manuscript and artifacts.

ETHICS STATEMENT

The experiments study simulated adversarial games and do not involve human-subject experiments or real-money play. Strategic agents could nevertheless be misused in gambling, manipulation, or other interactions affecting people. These evaluations do not establish deployment readiness in human-facing or high-stakes settings. Any reuse should respect the licenses of the underlying models, game software, and data, and should distinguish research evaluation from operational use.

REPRODUCIBILITY STATEMENT

The paper and appendices specify the method, theoretical assumptions and proofs, and evaluation protocols. The accompanying source provides the reported numerical results and editable plotting code to support verification of the analyses and reconstruction of the figures. Artifact scope and statistical conventions are described in Appendix B.9.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*, 2017. URL https://arxiv.org/abs/1707.01495.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 793–802, 2019. URL https://proceedings.mlr.press/v97/brown19b.html.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 66–75, 2020. URL https://proceedings.mlr.press/v100/chen20a.html.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016. URL https://arxiv.org/abs/1611.02779.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiaolong Jin, Dingmin Wang, Vijay Lingam, and Varun Kumar. AHEAD: Adaptive hindsight with environment-augmented distillation for agentic RL. *arXiv preprint arXiv:2608.24114*, 2026. URL https://arxiv.org/abs/2608.24114.

Seth Karten, Andy Luu Nguyen, and Chi Jin. PokéChamp: an expert-level minimax language agent. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29205–29222, 2025. URL https://proceedings.mlr.press/v267/karten25a.html.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1711.00832.

Charles Packer, Pieter Abbeel, and Joseph E. Gonzalez. Hindsight task relabelling: Experience replay for sparse reward meta-RL. *arXiv preprint arXiv:2112.00901*, 2021. URL https://arxiv.org/abs/2112.00901.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340, 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 993–1019, 2013. URL https://proceedings.mlr.press/v30/Rakhlin13.html.

Paulo Rauber, Avinash Ummadisingu, Filipe Mutz, and Jürgen Schmidhuber. Hindsight policy gradients. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.06006.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1507.00407.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/32cfdce9631d8c7906e8e9d6e68b514b-Abstract.html.

Huining Yuan, Zelai Xu, Zheyue Tan, Xiangmin Yi, Mo Guang, Kaiwen Long, Haojia Hui, Boxun Li, Xinlei Chen, Bo Zhao, Xiao-Ping Zhang, Chao Yu, and Yu Wang. MARSHAL: Incentivizing multi-agent reasoning via self-play with strategic LLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.15414.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/08d98638c6fcd194a4b1e6992063e944-Abstract.html.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.08348.

Appendix guide. Appendix A gives proofs and analytic examples; Appendix B specifies learning and reporting details; Appendix C collects quantitative results; Appendix D describes the interface and metric conventions; Appendices E and F provide auxiliary probes and training diagnostics; Appendices G and H analyze representative game trajectories.

A Theoretical details and proofs

A.1 Notation and levels of description

Table 5: Key notation. Lower-case empirical labels and upper-case theoretical contexts have different assumptions; the same timing principle connects them.

Symbol	Meaning
h_t, a_t	Pre-action interaction history and generated action
$\widehat{z}_t, z_t, u_t$	Predicted context, retrospective training label, and selected update route
q_ψ, R_ϕ	Context predictor and Strategic Reward Model (SRM)
y_t, s_t	Observed delayed-return target and the SRM's estimate
$b_t, \rho_t(\theta)$	Logged behavior probability and cross-route likelihood ratio
ω, λ_z	Shared backbone parameters and the adapter for route z
p_t, q_t	Action-time and retrospective responsibilities in the soft-routing extension
$\mathcal{Z}, K$	Finite context set and number of actions in the idealized game
$Z_t, \widehat{Z}_t$	Realized and predicted theoretical context
ℓ_t	Full loss vector, bounded coordinatewise in $[-1, 1]$
$x_{t,z}, w_t$	Current distribution of learner z and the distribution used to act
$\mathcal{F}_z, s_T$	Visits to context z and number of visited contexts
$u_z^\star$	Best fixed distribution on the realized-context subsequence
$L_T, V_z^{(2)}$	Prediction-mistake count and squared within-context loss variation
η	Learning rate of the idealized online learner

A.2 Repeated-game interpretation and assumptions

Fix a player j with K actions. The joint mixed strategy is $w^j \otimes w^{-j}$. A latent context Z_t and the opponents' strategy induce a vector

$$\ell_t^j[k] = \mathbb{E}_{a^{-j} \sim w_t^{-j}} c^j(a_k^j, a^{-j}, Z_t), \qquad c^j(w_t^j \otimes w_t^{-j}, Z_t) = \langle w_t^j, \ell_t^j \rangle.$$

If a linear-in-context representation $c^j(a, Z) = \langle \phi^j(a), Z \rangle$ is used, then $\ell_t^j = \Phi^j(w_t^{-j})^\top Z_t$, with column k of Φ^j equal to $\mathbb{E}_{a^{-j}} \phi^j(a_k^j, a^{-j})$. Linearity in context is convenient but unnecessary for the routing decomposition.

The analysis assumes finite action and context sets, a context selected before the action, bounded loss vectors, and full loss-vector feedback after the round. The learner for each realized context is updated on every visit, including visits on which another learner generated the action. The regret inequalities below hold pathwise for the induced loss sequence. They do not assert that the sequence would stay unchanged if the player followed a counterfactual policy that altered future opponent behavior. Empirical labels constructed after a decision are proxies and need not satisfy the same exogeneity interpretation.

A.3 Exact decomposition and the mistake cost

For arbitrary fixed comparators $u_z \in \Delta_K$,

$$\begin{aligned} \sum_t \langle w_t - u_{Z_t}, \ell_t \rangle &= \sum_t \langle x_{t,Z_t} - u_{Z_t}, \ell_t \rangle + \sum_t \langle x_{t,\widehat{Z}_t} - x_{t,Z_t}, \ell_t \rangle \\ &= \sum_z \sum_{t \in \mathcal{F}_z} \langle x_{t,z} - u_z, \ell_t \rangle + \sum_t \langle x_{t,\widehat{Z}_t} - x_{t,Z_t}, \ell_t \rangle. \end{aligned} \tag{10}$$

The second term vanishes for $\widehat{Z}_t = Z_t$. Otherwise, Hölder's inequality bounds it by $\|x_{t,\widehat{Z}_t} - x_{t,Z_t}\|_1 \|\ell_t\|_\infty \le 2$. Choosing $u_z = u_z^\star$ proves Proposition 4.1. The base-learner guarantee is applied to the *whole* subsequence $\mathcal{F}_z$, never to only the correctly predicted rounds. A guarantee for a full sequence does not automatically hold for an arbitrary selected subset of that sequence.

The constant 2 cannot be improved using only the given boundedness assumptions: on a mistaken round, take the two route distributions to be $(1, 0)$ and $(0, 1)$ and the loss vector to be $(1, -1)$. The difference is exactly 2. This demonstrates sharpness of the per-round algebraic bound, not a minimax lower bound for all contextual algorithms.

A.4 Optimistic entropic mirror descent on each subsequence

We give a concrete instantiation following the predictable-sequence approach of Rakhlin & Sridharan (2013). For one context, index its visits by $r = 1, \dots, n$ and write its losses as g_r. Let v_1 be uniform on K actions, $m_1 = 0$, and $m_r = g_{r-1}$ for $r > 1$. Use

$$x_r = \arg\min_{x \in \Delta_K} \{\eta \langle m_r, x\rangle + \mathrm{KL}(x \| v_r)\}, \tag{11}$$

$$v_{r+1} = \arg\min_{x \in \Delta_K} \{\eta \langle g_r, x\rangle + \mathrm{KL}(x \| v_r)\}. \tag{12}$$

Both updates are exponentiated-weight updates and remain in the simplex interior when initialized uniformly. In the routed algorithm, x_r is the current distribution $x_{t,z}$ for this context. Even if it is not selected to act, Equation (12) is performed when this context is realized. Each context has its own clock; periods spent in other contexts do not advance it.

Lemma A.1 (Per-context predictable-sequence bound)**.** *For every $u \in \Delta_K$ and $\eta > 0$, the updates above satisfy*

$$\sum_{r=1}^{n} \langle x_r - u, g_r\rangle \le \frac{\mathrm{KL}(u\|v_1)}{\eta} + \frac{\eta}{2}\sum_{r=1}^{n} \|g_r - m_r\|_\infty^2 \le \frac{\log K}{\eta} + \frac{\eta}{2}\left(1 + \sum_{r=2}^{n} \|g_r - g_{r-1}\|_\infty^2\right). \tag{13}$$

Proof. The Bregman three-point inequality applied to Equation (12) yields

$$\eta\langle g_r, v_{r+1} - u\rangle \le \mathrm{KL}(u\|v_r) - \mathrm{KL}(u\|v_{r+1}) - \mathrm{KL}(v_{r+1}\|v_r).$$

Applying the same inequality to Equation (11), with comparator v_{r+1}, yields

$$\eta\langle m_r, x_r - v_{r+1}\rangle \le \mathrm{KL}(v_{r+1}\|v_r) - \mathrm{KL}(v_{r+1}\|x_r) - \mathrm{KL}(x_r\|v_r).$$

Add the two displays and the term $\eta\langle g_r - m_r, x_r - v_{r+1}\rangle$. Pinsker's inequality gives $\mathrm{KL}(v_{r+1}\|x_r) \ge \frac{1}{2}\|v_{r+1} - x_r\|_1^2$. Combining Hölder's and Young's inequalities,

$$\eta\langle g_r - m_r, x_r - v_{r+1}\rangle - \mathrm{KL}(v_{r+1}\|x_r) \le \frac{\eta^2}{2}\|g_r - m_r\|_\infty^2.$$

Drop the nonpositive term $-\mathrm{KL}(x_r\|v_r)$, divide by η, and sum. The remaining KL terms telescope; the final negative term may be dropped. Finally, $\mathrm{KL}(u\|v_1) \le \log K$ and $\|g_1\|_\infty \le 1$ establish the stated specialization. □

Summing Equation (13) over the s_T visited contexts and applying Proposition 4.1 proves Equation (9). The first-visit $\eta s_T/2$ term is retained explicitly. With $Q_T = s_T + \sum_z V_z^{(2)}$, minimizing $s_T \log K/\eta + \eta Q_T/2$ gives $\eta^\star = \sqrt{2 s_T \log K / Q_T}$ and $R_T^{\mathrm{ctx}} \le 2L_T + \sqrt{2 s_T \log K\, Q_T}$ for $K \ge 2$ and $s_T > 0$. This stepsize uses aggregate quantities in hindsight; the fixed-η result is the online guarantee proved here. Selecting or adapting stepsizes in advance is separate from the routing argument. Since $\|g_r - g_{r-1}\|_\infty \le 2$, the squared variation is also bounded by twice the unsquared variation. Thus the same interpretation applies to the latter, with the first-visit term preserved.

Relation to the overview schematic. The bound displayed in Figure 1 is a looser form of Equation (9). Restore the player index j and define the startup-augmented within-context variation

$$\mathrm{Var}_T^j(z) = \frac{1}{2}\left(\mathbf{1}\{\mathcal{F}_z \ne \varnothing\} + V_z^{(2)}\right),$$

with $V_z^{(2)} = 0$ for an unvisited context. For a bank of m contexts and $0 < \eta \leq 1$, Equation (9) implies

$$R_T^{\mathrm{ctx}} \leq \frac{m \log K}{\eta} + \frac{2}{\eta} L_T + \eta \sum_{z \in \mathcal{Z}} \mathrm{Var}_T^j(z). \tag{14}$$

Indeed, $s_T \leq m$, $2L_T \leq 2L_T/\eta$, and the startup-augmented variation absorbs exactly the first-visit term and half the squared within-context variation. The main bound retains the sharper $2L_T$ mistake cost and makes initialization explicit. The schematic therefore illustrates the same decomposition under this convention, without adding a convergence claim for the neural optimizer.

A.5 Comparison with non-routed guarantees

Let $V_{\mathrm{global}}^{(2)} = \sum_{t \geq 2} \|\ell_t - \ell_{t-1}\|_\infty^2$. A single optimistic learner of the same form has a bound

$$\frac{\log K}{\eta} + \frac{\eta}{2}\left(1 + V_{\mathrm{global}}^{(2)}\right)$$

against a single static comparator. The routed guarantee pays for multiple initializations and context prediction, but can replace global variation by the sum of within-context variations. There is no general inequality ordering these two variation quantities: the subsequences can suppress cross-context changes or expose fluctuations that are less visible chronologically.

For the alternating example in the main text, the best context-specific comparator incurs zero loss, while the best static comparator incurs $T/2$ for even T. This quantifies a difference in the achievable comparator value as well as a difference in variation. It does not prove that a context-feature neural policy cannot learn the same mapping. The matched experiments, rather than this bound, compare such architectures.

The theory also cannot compare realized return directly with a certified equilibrium solver. External regret under a realized sequence and exploitability against a best response are distinct objects. A no-regret argument for independent idealized learners does not transfer automatically to adapters that share a changing backbone.

A.6 Soft responsibilities

The finite-context decomposition has a useful soft analogue. Let $p_t, q_t \in \Delta_m$ denote action-time and retrospective responsibilities, and suppose the played distribution is $w_t = \sum_z p_t(z) x_{t,z}$. Assess it against $\sum_z q_t(z) u_z$. If each weighted base learner obeys

$$\sum_t q_t(z) \langle x_{t,z} - u_z, \ell_t \rangle \leq B_z^q(T),$$

then

$$\sum_t \left\langle w_t - \sum_z q_t(z) u_z, \ell_t \right\rangle \leq \sum_z B_z^q(T) + \sum_t \|p_t - q_t\|_1. \tag{15}$$

Indeed, add and subtract $\sum_z q_t(z) x_{t,z}$. The weighted base regrets form the first term, and the remaining term is $\sum_{t,z} (p_t(z) - q_t(z)) \langle x_{t,z}, \ell_t \rangle$, bounded by the responsibility discrepancy because each inner product lies in $[-1, 1]$. One-hot responsibilities recover the $2L_T$ mistake bound.

This result describes a finite bank with overlapping assignments. It does not certify the quality of learned q_t, establish identification of latent regimes, or prove a bound for an infinite set of independently parameterized experts. Neural adapter interpolation need not equal a mixture of action distributions, either. The proposition therefore supplies a conceptual extension, with its mixture and weighted-regret assumptions stated explicitly, rather than a guarantee for the learned neural extension.

A.7 A two-round example of update assignment

Consider two contexts A, B and two actions, with uniform initial learners. On round one, the realized context is A, the prediction is B, and the loss vector is $(0, 1)$. Both acting distributions are initially uniform, so the expected round-one loss is $1/2$ under either update rule. The distinction appears after feedback.

With realized-context updating, learner A receives $(0, 1)$. Its entropic anchor becomes proportional to $(1, e^{-\eta})$, and its next optimistic prediction is $(0, 1)$. If round two again has context A and is predicted correctly, Equation (11) therefore selects

$$x_{2,A} = \frac{(1, e^{-2\eta})}{1 + e^{-2\eta}}, \qquad \langle x_{2,A}, (0, 1) \rangle = \frac{1}{1 + e^{2\eta}} < \frac{1}{2}.$$

If the first loss instead updates the acting learner B, learner A remains uniform and incurs expected loss $1/2$ on round two. Retrospective assignment has preserved information for the next visit to A, despite the earlier acting error.

This is an analytic example, not an additional benchmark experiment or a dominance proof. Other context sequences can favor different assignments. It relies on observing the full loss vector even when A did not act; that is precisely the feedback condition that the neural implementation does not satisfy. It also shows why mistake count can be conservative: the initial prediction is wrong, yet its immediate mismatch cost is zero because the two distributions coincide.

A.8 Prediction accuracy versus decision-relevant mismatch

The exact decomposition retains more information than the mistake count. Let

$$D_T = \sum_t \frac{\max_k \ell_t[k] - \min_k \ell_t[k]}{2} \|x_{t,\widehat{Z}_t} - x_{t,Z_t}\|_1.$$

Because the two distributions have equal total mass, subtracting the midpoint of the largest and smallest losses does not change their inner product with the loss vector. Hölder's inequality then gives

$$R_T^{\mathrm{ctx}} \le \sum_{z \in \mathcal{Z}_T} B_z(T) + D_T, \qquad 0 \le D_T \le 2L_T.$$

Thus a mistaken label matters less when its two routes prescribe similar behavior or when all actions have similar loss. This refinement is a consequence of the same algebra, not a guarantee based solely on measured classification accuracy. In the experiments, macro-F1, return, and route utilization consequently remain distinct diagnostics.

A.9 Scope of the idealization

Table 6: Correspondence between the analyzed learner and neural ISO. The routing timing is shared; the assumptions needed for the regret guarantee are not established for neural training.

Aspect	Analyzed learner	Neural implementation
Context	Selected before the action, revealed later	Retrospective proxy may depend on the action and continuation
Feedback	Full loss vector on every realized-context visit	Sampled continuations and estimated delayed returns
Parameters	Separate base learners with regret guarantees	Shared backbone and route-specific adapters
Update	Exact stated online-learning rule	Clipped and filtered likelihood surrogate
Comparator	One fixed action distribution per context	Empirical return against the tested opponents
Soft routing	Convex mixture of action distributions	Adapter interpolation need not equal a policy mixture

The guarantee explains why recurring contexts can replace chronological variation with within-context variation when its assumptions hold. It does not certify the SRM, identify a useful label taxonomy, or account for changes to inactive routes through a shared backbone. The controlled experiments test the practical update assignment under these additional sources of approximation. This separation is also why the theoretical comparator is not used to rank the neural policy against CFR or a best-response solver.

B PROTOCOL AND IMPLEMENTATION DETAILS

B.1 INFORMATION AVAILABLE AT EACH STAGE

A strategic context is an abstraction of recurrent interaction, not a revelation of the full hidden game state. The hard-routing experiments use simulator-derived labels to supervise a predictor; the learned-context experiment removes those environment-provided labels. Both versions preserve the same timing distinction.

Table 7: Information boundary for a training sample. A retrospective quantity may be used for training without becoming an action-time observation.

Stage	Available quantities	Use
Before the action	Public state, own private observation, legal actions, past interaction	Construct h_t, predict $\widehat{z}_t$, choose the behavior route
Action generation	Sampled output and its likelihood under the frozen behavior route	Execute a_t and store $\log b_t$
After the continuation	Subsequent observations, realized payoff, retrospective label	Construct y_t, supervise the predictor/SRM, assign update route u_t
Policy update	Logged $h_t, a_t, \log b_t, z_t$ and SRM score	Re-evaluate under u_t and optimize the routed surrogate
Deployment	Causal history and predicted context	Select an action without retrospective labels or future returns

The poker proxies summarize aggression, showdown tendencies, responses to raises, and large-pot behavior. Battle proxies summarize phase, revealed information, remaining team composition, and recent opponent choices. These summaries can be informative even when an individual decision remains ambiguous. They are also domain dependent: an accurate classifier of a poor abstraction need not improve control. A post-decision label is therefore a training target whose usefulness must be evaluated, not ground-truth access to an opponent's intent.

The predictor audit distinguishes removing specifically identified post-decision fields from imposing a strict timestamp boundary on all inputs. The latter is the stronger test. Separating opponents and sessions across splits further reduces memorization of recurrent identities and near-duplicate histories. The audit results are reported in Table 15; they do not by themselves determine the end-to-end return of retraining every variant with the strictest split.

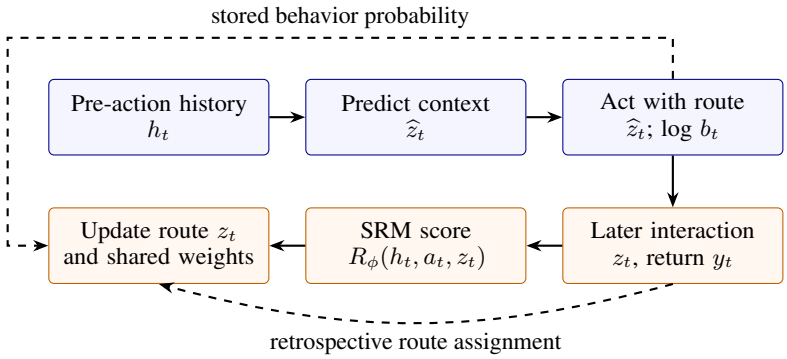


Figure 4: Detailed information flow. Blue nodes use action-time information; orange nodes use later training feedback. The behavior likelihood is logged before assigning the update route. The shared backbone receives all policy gradients; only the adapter selected for updating receives the route-specific gradient.

B.2 Routed optimization and behavior support

The route selected for generating an action and the route selected for updating its likelihood are separate fields in the training record. The denominator in Equation (5) must remain the probability that generated the action. Replacing it with the old probability under the retrospective route would erase the behavior mismatch rather than account for it.

For a generated token sequence $a = (a_1, \ldots, a_J)$, the action log-probability can be represented as

$$\log \pi_\theta(a \mid h, z) = \sum_{j=1}^{J} \log \pi_\theta(a_j \mid h, z, a_{<j}). \tag{16}$$

The numerator and denominator must refer to the same generated output and token mask. Prompt tokens and padding cannot enter one side but not the other. If an implementation instead uses a token-level surrogate, it must specify that choice explicitly; token-level clipping is not generally identical to clipping a sequence-level ratio.

The group normalization in Equation (5) compares candidate actions from one pre-action history. Each candidate's score uses its own continuation and retrospective label. The numerical stabilizer ϵ_A handles groups with nearly equal scores. In the zero-variance case, subtracting the group mean gives no preference signal. This normalization cannot replace held-out validation of the SRM: a confidently incorrect ranking can still drive a large update.

The described training procedure excludes low-support samples and clips the remaining ratios. These operations control instability at the cost of changing the update objective. The behavior log-probabilities should be retained even for excluded samples so that the exclusion rate can be audited. The numerical support threshold and clipping coefficient are implementation choices that affect this bias–variance tradeoff.

Shared and route-specific parameters. All contexts share the transformer backbone. Context-indexed low-rank adapters add route-specific trainable parameters. Hard routing activates one adapter for acting and assigns each sample to one adapter for updating; the assignments need not coincide. The shared backbone receives gradients from all samples. Thus, the neural implementation is not a collection of statistically independent learners, even though the idealized analysis uses such a collection to isolate the routing principle.

The matched feature and history controls use the reported equal total/active parameter budget. This matching addresses capacity as an alternative explanation of return differences; it is not a measured wall-clock or accelerator-memory equivalence. The strongest mechanism control is P→P, which keeps the same routed parameterization and changes only the update route.

B.3 Cross-route likelihood ratios and retrospective selection

The denominator in Equation (5) records the actual behavior probability. Using the retrospective route in that denominator would no longer describe the distribution that generated the action. Correct behavior bookkeeping is necessary, but does not by itself transform the update into an unbiased gradient of deployment return.

To see the remaining distinction, fix a causal history h and its acting route. Let $b(a \mid h)$ be the behavior policy and $\kappa(dy, z \mid h, a)$ the joint probability kernel of the retrospective label and continuation return under the data-collection process. Assume $b(a \mid h) > 0$ for every action in the sum. For a fixed integrable score f,

$$\begin{aligned} &\mathbb{E}_{a\sim b,\,(z,y)\sim\kappa}\left[\frac{\pi_\theta(a \mid h, z)}{b(a \mid h)} f(a, z, y)\right] \\ &\quad = \sum_{a,z} \int \pi_\theta(a \mid h, z) f(a, z, y)\, \kappa(dy, z \mid h, a). \end{aligned} \tag{17}$$

If z depends on a, the weights on the right need not sum to one. For an elementary illustration, take two actions, a label equal to the chosen action, $b(a \mid h) = 1/2$, and $\pi_\theta(a \mid h, z) = \mathbf{1}\{a = z\}$. The total weight is two. This is an algebraic example showing why the expression is not generally the

expectation under an ordinary deployment policy; it is not an experiment or a claim that the empirical labels have this particular form.

If the label law is action-independent, the total weight instead normalizes, but the continuation still follows the recorded rollout process. Changing all later policy choices requires additional distributional accounting. Group-relative advantages, clipping, and support filtering further distinguish Equation (6) from an unbiased return gradient. Hindsight policy gradients derive trajectory corrections under their goal-conditioned assumptions (Rauber et al., 2019); those assumptions cannot simply be imported for an action-dependent strategic label.

Initial cross-route ratios and clipping. For ordinary same-route updates, the ratio equals one at $\theta = \theta_{\text{old}}$. In a cross-route update its initial value is already

$$\rho_0 = \frac{\pi_{\theta_{\text{old}}}(a \mid h, z)}{\pi_{\theta_{\text{old}}}(a \mid h, \widehat{z})}.$$

For a fixed advantage A, the clipped term is $A \min\{\rho, 1+\epsilon_c\}$ when $A \geq 0$, and $A \max\{\rho, 1-\epsilon_c\}$ when $A < 0$. Some cross-route samples can therefore lie on a flat part of this term before the first optimizer step. This behavior is part of the practical surrogate; it is not removed by accurately logging the denominator. The KL term and the support filter impose additional constraints.

Shared gradients and route-specific gradients. For a sample assigned to u, only λ_u enters its routed likelihood and same-route KL term, whereas the shared parameter ω enters every route. Updating ω can change the behavior of components whose adapters are inactive for that sample. The statement that another adapter receives no gradient therefore does not imply that its complete policy is unchanged. This distinction is essential when relating the practical implementation to the independent learners in Table 6.

Diagnostic implications. The relevant optimization diagnostics include the distribution of cross-route ratios, the fraction of samples filtered or clipped, and the sample counts assigned to each route. These quantities explain why predictor accuracy alone is insufficient to characterize a routed optimizer. We do not infer their values from the reported return differences. The P→P control remains the direct empirical test of the chosen assignment rule under the matched architecture and reward model.

B.4 Interpretation of complementary ablations

Table 8: Questions addressed by the available controls. Each comparison targets a different alternative explanation, so its numerical effect should be interpreted within its own protocol.

Control	Question	Relevant distinction
No SFT initialization	Does starting behavior matter?	Initialization versus later optimization
Win-based reward without SRM	Is delayed-value supervision useful?	Outcome frequency versus delayed payoff
SFT + SRM without GRPO	Is the optimization stage useful?	Scoring versus updating a policy
Multi-hand GRPO	Is a longer reward horizon sufficient?	Extended return versus context routing
History-only recurrent control	Is memory sufficient?	Causal history versus routed parameters
Shared SRM-GRPO	Is the SRM alone sufficient?	Same value model with shared updates
Equal-capacity context feature	Is context merely a useful input?	Feature conditioning versus routed updates
Predicted-to-predicted routing	Does update assignment matter?	Same routed structure, different update route

Together these controls make the claim more specific: prediction-aware routing is evaluated in addition to history, delayed returns, reward modeling, and capacity. None of the individual control results implies that the other components are unnecessary. The full method combines them, while the matched comparisons identify what remains after selected alternatives are held fixed.

B.5 Relationship to hindsight and modular learning

The use of information available only after an action is established in reinforcement learning. The question for ISO is how that information assigns experience to a context-conditioned policy component. Table 9 separates this question from context inference, reward relabelling, and advantage construction. These distinctions locate the method; they do not establish a performance ordering over the cited approaches.

Table 9: Related mechanisms and the specific question studied by ISO. Descriptions concern the cited methods, not additional implemented baselines.

Mechanism	Role of context or later information	Relation to this study
PEARL / VariBAD (Rakelly et al., 2019; Zintgraf et al., 2020)	Infer latent task information for adaptation and control	Establish the value of context inference; ISO asks which component receives an update after more evidence arrives
Soft modularization (Yang et al., 2020)	Select task-dependent computation within a shared policy	Establishes modular capacity; the P→P control holds routed architecture fixed
Hindsight task relabelling (Packer et al., 2021)	Reinterpret experience under a hindsight task and its reward	ISO retains the environment payoff and changes context-specific sample assignment
Hindsight policy gradients (Rauber et al., 2019)	Correct trajectory reuse under alternative goals	ISO's action-dependent labels do not automatically satisfy those correction assumptions
AHEAD (Jin et al., 2026)	Use later feedback and hints to reweight token advantages	Uses privileged supervision to adjust a signal; ISO assigns the sample to a retrospective component
MARSHAL (Yuan et al., 2026)	Estimate turn-level advantages with agent-specific normalization in self-play	Addresses multi-turn signals; ISO tests update assignment under a common SRM

The matched P→P versus P→R comparison isolates a change within the ISO architecture. It does not substitute for implementing each related system under a shared game protocol. In particular, differences in rulesets, opponent populations, model sizes, and training budgets preclude ranking those systems using their published headline results.

B.6 Return horizons and evaluation populations

Poker uses 100-BB initial stacks and $0.5/1$-BB blinds. The same legal betting templates are used across the matched variants. Per-hand GRPO uses a hand's payoff. MR-GRPO aggregates discounted payoff over blocks of 20 hands while retaining causal cross-hand history, with the stack reset rules unchanged. A history-conditioned observation alone does not imply that optimization accounts for subsequent hands.

The controlled study uses the same SFT initialization, Qwen2.5-0.5B backbone, causal observations, interaction count, optimizer-step budget, opponent draws, seat permutations, and evaluator. Five seeds and eight seating permutations are reported, with paired deal and seating schedules across variants. These are paired sources of environmental variation, not a claim that distinct policies visit identical action trajectories. The reference evaluation uses 10,000 hands per configuration.

In Pokémon, the ruleset is Gen 1 OU. A training episode is a complete battle from the initial observation to the terminal battle outcome. The full-episode GRPO baseline already receives the eventual battle outcome. Elo and win rate describe performance under the reported opponent pool; they are not directly comparable to ratings from another generation, matchmaking population, or benchmark. In particular, the Gen 9 OU setting of Karten et al. (2025) is related work, not a matched numerical baseline.

Unseen and hybrid opponents. The discrete-context generalization study separates the seen population, previously unseen opponents, and hybrid opponents that combine strategic patterns. The learned-context study instead tests held-out interpolation weights in a controlled mixture and an unseen strategy family. These are different evaluations and their absolute returns should not be pooled. Both test whether useful structure extends beyond the training population, with narrower scope than arbitrary open-world adaptation.

B.7 Learning overlapping contexts

The soft-routing extension uses eight adapter components. Its action-time encoder maps causal history to responsibilities $p_t \in \Delta_8$; the target encoder maps post-trajectory information to responsibilities $q_t \in \Delta_8$. The learned variants do not receive environment-provided context labels. Sparse activation and load balancing are used to limit active cost and discourage concentration on one component.

The weighted-mixture result in Appendix A.6 concerns distributions over actions. Mixing adapter outputs inside a nonlinear neural network need not equal a convex mixture of the resulting policies. Consequently, that result describes a routing principle rather than a theorem for the specific neural soft-routing implementation.

B.8 Value-model validation and refresh

The initial SRM fits delayed-return targets from logged training segments. Its held-out evaluation uses rank correlation and top-decile precision, which assess whether higher scores identify higher-return actions. These diagnostics are distinct from predictor NLL and expected calibration error, which apply to context probabilities.

The distribution-shift experiment compares a frozen SRM with a refreshed SRM that mixes original and recent on-policy trajectories equally. The reported monitoring threshold for held-out Spearman correlation is 0.38. Uncertainty-gated mixing uses Monte Carlo returns for uncertain samples. Table 17 indexes evaluations by their order to compare ranking quality across successive policy distributions.

B.9 Statistical interpretation and reporting

The main controlled comparison reports paired 95% confidence intervals for ISO-minus-baseline differences. Pairing uses common deals and seating schedules, with results aggregated across five training seeds and eight seating permutations. The number of evaluated hands should not be confused with the number of independent training runs. The main claims use the reported paired intervals without adding unreported significance tests.

The broader reference study reports marginal return intervals, whereas the matched study reports intervals on paired differences. These quantify different objects. The learned-context, SRM-refresh, and Pokémon tables report point estimates without uncertainty intervals; their comparisons are descriptive. No uncertainty bars are inferred from point estimates.

The implementation uses a VERL-based training framework with a poker rollout and data interface. The principal reproducibility requirements are the information boundary in Table 7, behavior likelihoods logged under the acting route, retrospective assignment in Equation (5), and consistent history and reward definitions across controls. The accompanying manuscript source provides aggregate results and editable plotting code for eight reconstructed experimental figures and Figure 7, with source annotations identified and original training traces retained. This package supports reconstruction of the reported figures; complete training-run artifacts are not included.

C ADDITIONAL EMPIRICAL RESULTS

This appendix retains the broader reference evaluation and the diagnostic studies separately from the matched experiment in Table 2. In particular, the reference ISO return is 15.8 BB/100 and the matched ISO return is 15.6 BB/100. These are distinct protocol outcomes, not interchangeable estimates of a single run.

C.1 BROADER POKER REFERENCE EVALUATION

Table 10: Broader NLHE reference evaluation. LTR is in BB/100 hands with reported marginal 95% intervals. Vulnerability is a protocol-specific estimate in mBB/hand, not certified NashConv; its best-response class, budget, and abstraction are insufficiently specified for an equilibrium claim. The external LLMs and CFR methods are reference systems with different capacities and objectives.

Method	LTR [95% CI] ↑	Win (%)	Vulnerability ↓
SFT	3.1 [2.0, 4.2]	18.0	8.4
PPO (win reward)	5.6 [4.3, 6.9]	25.0	7.1
PPO (profit reward)	8.7 [7.2, 10.1]	23.4	6.8
GRPO (profit reward)	11.9 [10.5, 13.2]	22.6	6.0
Deep CFR	9.6 [8.3, 10.8]	20.8	4.4
CFR	1.2 [0.4, 2.0]	17.0	3.8
GPT-4o	−8.2 [−9.8, −6.5]	14.0	12.5
Claude 3.5	−5.4 [−6.8, −4.0]	15.0	11.2
Llama 3.1-70B	−6.8 [−8.1, −5.4]	13.0	13.8
ISO	**15.8** [14.2, 17.3]	22.0	5.2

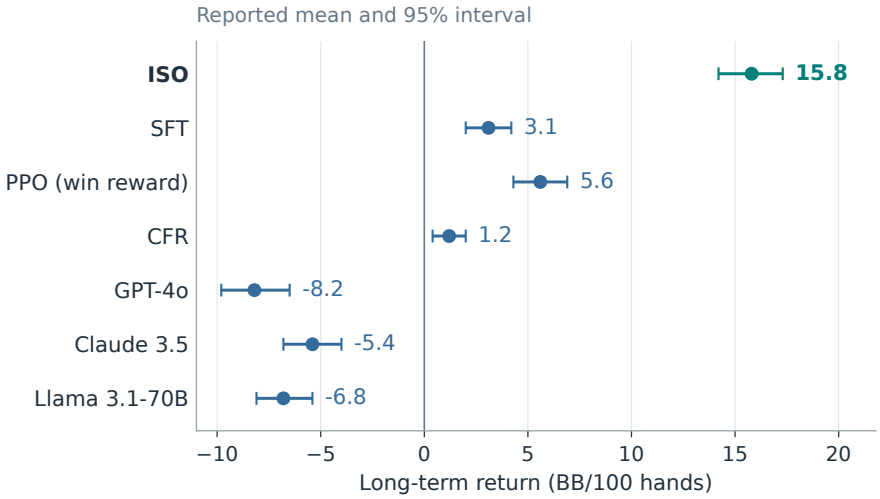


Figure 5: Long-term return for the seven reference systems displayed in the original comparison. Points and horizontal bars show the means and reported marginal 95% intervals from Table 10, in BB/100 hands. PPO uses the win-based reward. The complete reference table additionally includes PPO with profit reward, GRPO with profit reward, and Deep CFR.

The highest hand win rate and the highest return belong to different methods. This illustrates why chip return is the primary poker objective: frequent small wins need not compensate for occasional large losses. ISO's larger population return does not supersede the lower vulnerability estimates of CFR and Deep CFR.

C.2 Reference dispersion and pipeline ablations

Table 11: Reported dispersion in the broader poker reference evaluation, with five seeds and eight seating permutations for each method. Standard deviations and marginal intervals describe the reference protocol; paired difference intervals for the matched protocol appear in the main text.

Method	Mean LTR	Reported std.	Reported 95% interval
PPO (win reward)	5.6	1.4	$[4.3, 6.9]$
PPO (profit reward)	8.7	1.6	$[7.2, 10.1]$
GRPO (profit reward)	11.9	1.5	$[10.5, 13.2]$
Deep CFR	9.6	1.3	$[8.3, 10.8]$
ISO	15.8	1.7	$[14.2, 17.3]$

The reported dispersion is appreciable relative to some between-method differences, as expected in poker. Multiple seating permutations expose each policy to different interaction positions, but they do not create additional independently trained seeds. We retain the reported intervals instead of deriving new ones from the hand count or treating every action as an independent replicate. The main controlled comparison uses the reported paired differences to assess the routing and reward controls.

Table 12: Pipeline ablations in the reference poker evaluation. Removing SFT changes initialization; removing SRM changes the training signal; removing GRPO changes the optimization stage. These are complementary ablations, not individually matched estimates of the routing effect.

Variant	LTR (BB/100) ↑	Win rate (%)
ISO	15.8	22.0
Without SRM (win-based reward)	6.8	21.0
Without GRPO (SFT + SRM)	9.2	19.0
Without SFT initialization	-2.1	12.0

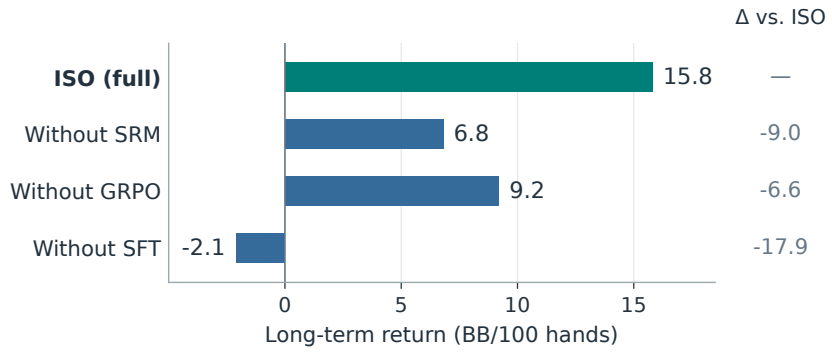


Figure 6: Pipeline ablation figure. The four returns are 15.8 for ISO, 6.8 without SRM, 9.2 without GRPO, and -2.1 without SFT, in BB/100 hands. The corresponding win rates and variant definitions are given in Table 12.

The three component removals alter different parts of the learning procedure. Removing SRM replaces the delayed-value signal with a win-based reward. Removing GRPO leaves the SFT-plus-SRM variant without the full policy-optimization stage. Removing SFT changes the policy initialization before subsequent learning. Relative to ISO, the respective return reductions are 9.0, 6.6, and 17.9 BB/100, computed directly from the reported values.

These results support the joint use of a useful policy prior, a delayed-value signal, and policy optimization. They do not decompose the total improvement into additive component effects: each removal changes the conditions under which the remaining components operate. In particular, the no-SRM comparison changes the reward objective, whereas the shared-SRM control in the main text keeps the learned reward and removes context-specific routing. The latter is the appropriate comparison for asking whether reward shaping alone explains the result.

Similarly, the no-SFT result does not imply that supervised training itself solves strategic learning. SFT provides a starting behavior distribution; the full method must still improve the choices made within that distribution and the continuations induced by those choices. The reference SFT return of 3.1 and full ISO return of 15.8 describe this distinction at the endpoint level.

C.3 Generalization across opponent populations

Table 13: Discrete-context poker evaluation on seen, unseen, and hybrid opponent populations. Entries are LTR in BB/100 hands.

Method	Seen ↑	Unseen ↑	Hybrid ↑
SFT	3.1	1.0	0.4
PPO (profit)	8.7	5.4	4.8
GRPO (profit)	11.9	8.2	7.1
Deep CFR	9.6	7.9	6.8
ISO	**15.8**	**12.6**	**11.3**

ISO retains a return advantage over these references, while its own return falls as the evaluation population changes. The results support transfer within the tested poker populations. They do not show that the predictor discovers arbitrary opponent types or that adapters trained in one game transfer to another game.

Table 14: Additional vulnerability and return-shift detail for the populations in Table 13. Vulnerability is the same protocol-specific estimate as in Table 10, in mBB/hand. The final column is hybrid LTR minus seen LTR, in BB/100.

Method	Seen vulnerability	Unseen	Hybrid	LTR change
SFT	8.4	9.6	10.1	−2.7
PPO (profit)	6.8	7.9	8.2	−3.9
GRPO (profit)	6.0	7.0	7.4	−4.8
Deep CFR	4.4	4.9	5.3	−2.8
ISO	5.2	5.9	6.3	−4.5

ISO retains the highest return among these methods on the tested unseen and hybrid populations, while its vulnerability estimate increases. Deep CFR retains the smallest vulnerability estimate in this subset. A small absolute return drop alone is not evidence of the best generalization: SFT drops less partly because its seen-population return starts much lower. Both the final return and the change from the seen population are relevant.

C.4 Context prediction and information-boundary diagnostics

Table 15: Reported context-prediction metrics. NLL is negative log-likelihood; ECE is expected calibration error (both lower is better). Dashes indicate quantities not reported for that condition. The input-audit rows are separate conditions, not additional metrics of one identical evaluation.

Evaluation condition	Accuracy	Macro-F1	NLL	ECE
Poker, original seen-population evaluation	0.83	0.79	0.53	0.06
Poker, identified post-decision fields removed	—	0.78	—	—
Poker, strict action-boundary truncation	—	0.76	—	—
Poker, opponent/session-disjoint evaluation	—	0.69	—	—
Poker, permuted-label control	—	0.33	—	—
Pokémon, seen opponents	0.80	0.75	0.60	0.07
Pokémon, unseen opponents	0.71	0.65	0.79	0.12

The permuted-label result is a negative control, not a theoretically determined chance level without the class proportions. The performance reduction on disjoint opponents and sessions is larger than the reduction from removing the identified post-decision fields. This motivates auditing both time boundaries and distributional separation. End-to-end return must still be checked under the intended deployment boundary; predictor quality alone cannot establish absence of downstream leakage.

C.5 Return estimation and policy-induced distribution shift

Table 16: Held-out delayed-return ranking in the initial SRM evaluation. Top-10% precision is reported as a percentage.

Scoring model	Spearman ρ	Kendall τ	Top-10% precision
Strategic Reward Model	0.46	0.33	71
Win-rate critic	0.18	0.12	55
Random scores	0.00	0.00	10

Table 17: SRM ranking under evolving policies (top) and end-to-end mitigation results (bottom). Snapshots are listed in their reported order; the axis describes successive policy distributions. Refresh uses an equal mixture of original and recent on-policy trajectories.

Policy snapshot	Frozen SRM ρ	Refreshed SRM ρ
Initial	0.46	0.46
First evaluated policy snapshot	0.43	0.46
Second evaluated policy snapshot	0.39	0.44
Third evaluated policy snapshot	0.34	0.42
Mitigation	**LTR (BB/100)**	
Frozen SRM	15.6	
Periodic SRM refresh	16.1	
Uncertainty-gated SRM/Monte Carlo mixing	16.0	

Refreshing improves ranking on later-policy data without restoring the initial correlation completely. The return differences are small relative to the overall ISO gain and lack reported uncertainty intervals. They support refresh as a mitigation for estimator drift, without establishing equal computational cost or equivalence between the two mitigation strategies.

C.6 BEHAVIORAL DIAGNOSTICS

Table 18: Reported decomposition of reference LTR by poker street, in BB/100 hands. Each row sums to its corresponding total return. These are accounting contributions, not causal effects of arriving at a particular street.

Method	Pre-flop	Flop	Turn	River	Total
ISO	2.1	3.4	4.6	5.7	15.8
SFT	1.0	1.4	0.5	0.2	3.1
PPO (win reward)	2.4	2.1	0.8	0.3	5.6
GPT-4o	-1.0	-2.0	-2.4	-2.8	-8.2

Table 19: Reported conditional expected value by pot-size bin, in BB/hand. The bins select different subsets of play and cannot be interpreted as randomized interventions.

Method	< 5 BB	5–20 BB	20–50 BB	> 50 BB
ISO	0.010	0.018	0.030	0.060
SFT	0.006	0.008	0.004	-0.005
PPO (win reward)	0.014	0.010	-0.006	-0.030
GPT-4o	-0.005	-0.012	-0.020	-0.040

Table 20: Poker sacrifice diagnostics. NIE denotes actions judged to have negative immediate expectation by the shallow evaluator; LH+ is the reported fraction of those actions followed by positive long-horizon value. The final column is the reported associated LTR contribution, not a counterfactual causal gain.

Method	NIE (%)	LH+ (%)	Associated LTR (BB/100)
ISO	12.8	61.0	$+4.9$
SFT	4.1	48.0	$+0.7$
PPO (win reward)	5.6	44.5	$+0.4$
CFR	2.2	50.5	$+0.2$

Taken together, the street, pot-size, and sacrifice diagnostics associate ISO's advantage with later and higher-payoff decisions. Their interpretation depends on the shallow evaluator and the policy-induced state distribution. More sacrifice is not inherently better: an agent could sacrifice frequently while decreasing total return. Environment payoff remains the primary outcome.

C.7 PREDICTION-CORRUPTION SENSITIVITY

Table 21: Reference corruption sweep. p is the reported corruption level. All returns are in BB/100 hands. Because both the non-routed baseline and privileged oracle change with p, this table is a joint sensitivity study rather than a controlled intervention on ISO's predictor alone.

p	ISO LTR	GRPO LTR	Oracle LTR	Random-route LTR	ISO vulnerability
0.0	15.8	11.9	16.9	3.4	5.2
0.1	13.9	11.3	16.1	3.2	5.6
0.3	10.4	9.8	14.2	2.8	6.8
0.5	6.7	7.3	11.0	2.5	8.9
0.7	2.1	4.6	7.0	2.1	12.5

ISO's advantage over GRPO disappears at high corruption levels. This is consistent with the prediction-error cost in the idealized decomposition and with cross-route support problems in the

neural optimizer. It is not a numerical verification of the regret theorem, whose feedback model, losses, and comparator differ from these experiments.

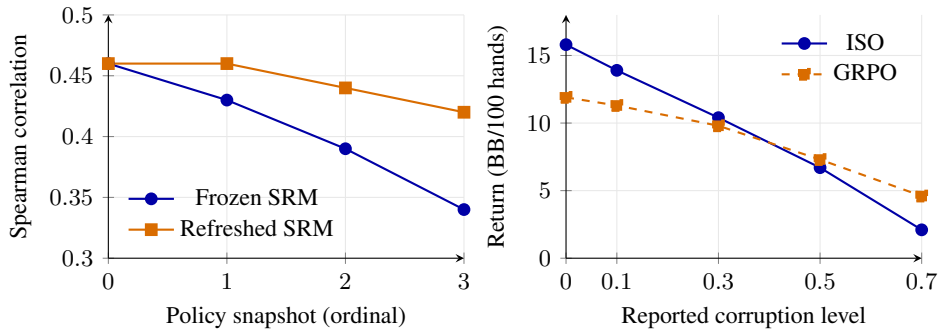


Figure 7: Aggregate diagnostic results replotted from Tables 17 and 21. Left: refreshing mitigates the decline in SRM ranking as the policy changes. Right: ISO's return advantage erodes under the reported corruption settings. The x-axis on the left is an ordinal evaluation index, not a training-step or compute scale. No uncertainty bands are inferred from the point estimates.

D Environment, Observation Interface, and Evaluation Units

This section describes the environment interface and the observation and measurement conventions used to interpret the poker and battle evaluations. The protocol guide in Table 23 distinguishes the matched controls, broader reference results, auxiliary capability probes, and selected trajectories.

D.1 Poker Environment and Agent Gateway

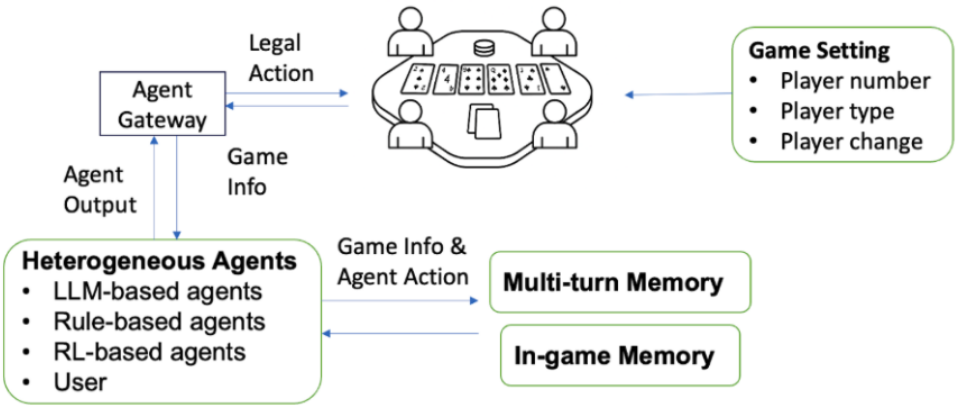


Figure 8: Poker environment and agent interface. The agent gateway connects heterogeneous agents to the game engine, while in-game and multi-turn memory expose different history horizons. Game settings determine the player configuration. The user node denotes an interface capability; the reported experiments concern simulated agent interactions.

Game state and legal actions. The engine advances through pre-flop, flop, turn, and river, enforces legal betting actions, and resolves terminal chip payoffs. The agent receives its own hole cards, the public board, position, pot and stack information, and the available action set. The gateway translates the selected action into the engine's representation. This separation matters because a language-model output must satisfy both a format constraint and the current game's action constraints. A syntactically valid action is not automatically a legal action at every state.

The action interface includes folding, checking or calling, and prescribed betting templates such as half-pot, pot-sized, and all-in actions when legal. The matched comparisons retain a common legal-action interface. Consequently, differences in action vocabulary are not the intended explanation for the ISO improvement. The separate raw-model probe in Appendix E.1 explicitly measures parsing failures, since these can alter realized behavior through the fallback rule.

Within-hand and cross-hand memory. In-game memory records the current hand's public action sequence. Multi-turn memory retains permitted information from earlier interaction, allowing the model to condition on recurrent opponent behavior. These histories serve different roles: current-hand information supports board- and range-dependent decisions, whereas cross-hand information can inform a prediction of strategic regime. Memory extends the observation horizon; the training return determines which future consequences enter optimization. Supplying a long history to a one-hand reward optimizer does not by itself make its objective a multi-hand objective.

D.2 Observation timing and retrospective supervision

The environment can log more information after an action than an agent could have known before taking it. ISO uses this asymmetry for learning. At action time, the predictor receives only the causal history h_t. Later feedback may determine the retrospective context label and delayed-return target. These later quantities supervise the update without being inserted into the earlier observation.

Table 22: Roles of the environment fields. The admissibility of a field depends on when it becomes observable, rather than only on its name.

Field family	Action-time role	Later training role
Own cards and public board	Describe the current decision	Preserve the sampled decision context
Pot, stacks, and legal actions	Constrain executable choices	Interpret action likelihood and payoff
Observed action history	Inform opponent-context prediction	Construct history-conditioned samples
Permitted previous-hand records	Describe recurring behavior	Support context supervision and analysis
Subsequent actions and outcomes	Not yet available for this decision	Form continuation targets and labels
Opponent private information	Use only if already revealed by the rules	Keep separate from causal predictor inputs

For example, a showdown tendency computed from previous completed hands can be part of causal history. The current hand's future showdown cannot be used in the same way before it occurs. Similarly, a response-to-raise summary may use previously observed responses, but the response to the action currently being selected belongs to the continuation. This distinction makes the timestamp boundary in Table 7 operational: a retrospective label is a permissible training target, not an additional deployment observation.

Figure 4 gives the learning information flow corresponding to the environment interface in Figure 8.

D.3 Experimental scales and metric interpretation

The manuscript contains several complementary evaluations with different opponents, outcome scales, and comparison units. Table 23 provides a reading guide. An absolute return from one row should be compared with its controls in that row, rather than used to rank an agent from a different protocol.

Table 23: Protocol and metric guide for the additional figures. Reference and matched results use the same return unit but remain separate evaluations.

Evaluation	Reported scale	Interpretation
Matched poker controls	BB/100 hands; ISO 15.6	Mechanism comparison with shared budgets and initialization
Broader poker references	BB/100 hands; ISO 15.8	Return across a wider set of systems and objectives
Auxiliary context diagnostic	BB/100 hands; ISO 6.3	Feature versus route comparison and accuracy sensitivity
Raw-model heads-up probe	Original return columns; see Appendix E.1	Short capability and interface diagnostic
Raw-model six-player probe	BB/h on the original axes	Model- and position-dependent return diagnostic
Training reward trace	Average reward; normalized step index	Optimization trajectory, not held-out poker LTR
Battle evaluation	Win rate and opponent-specific rating	Outcome against the stated battle opponent or pool
Selected battle cases	Turns, counts, and HP descriptors	Descriptive summaries of three individual victories

For the primary poker evaluation, if g_i is the net chip change in hand i divided by the big blind, the return convention is

$$\widehat{\mathrm{LTR}} = \frac{100}{N}\sum_{i=1}^{N} g_i \qquad \text{(BB/100 hands).} \tag{18}$$

Win rate instead measures the frequency of winning outcomes and does not weight those outcomes by their chip magnitude. Vulnerability uses milli-big-blinds per hand (mBB/hand): one mBB is 10^{-3} BB. Its unit is a payoff scale, but its best-response objective differs from population return. The street-wise quantities in Table 18 are reported accounting contributions to LTR. The pot-size quantities in Table 19 are conditional returns within different subsets of play. Neither is an independent additional return to be added to the primary score.

The raw-model probes retain the supplied numerical records. The original heads-up figure labels return as BB/h whereas its accompanying source table labels the return columns as BB/100. The reconstructed heads-up panel therefore uses "reported return" and states the unresolved source-unit discrepancy explicitly in Appendix E.1; we do not rescale those numbers or pool them with the primary evaluation. The six-player plot retains its original BB/h convention. These probes describe interface behavior and position sensitivity; the main quantitative claims use the separately specified matched evaluation.

Population return and equilibrium vulnerability. For a fully specified finite game with utilities U_j and strategy profile π, the standard NashConv quantity is

$$\mathrm{NashConv}(\pi) = \sum_j \left[\max_{\pi'_j} U_j(\pi'_j, \pi_{-j}) - U_j(\pi_j, \pi_{-j})\right].$$

Each summand measures the gain from a unilateral best response; the sum vanishes at a Nash equilibrium. In two-player zero-sum games, regret minimization provides a route to equilibrium approximation (Zinkevich et al., 2007). That guarantee is distinct from the expected payoff of one agent against a specified distribution of opponents. An agent can earn more against that population while leaving larger gains available to an adversarial response.

The paper's vulnerability estimate does not specify a complete best-response certificate, and the primary poker setting has multiple players. We therefore preserve the reported values as protocol-specific diagnostics. Neither the units nor the presence of a CFR reference establishes exact NashConv, a two-player exploitability convention, or an equilibrium guarantee for neural ISO.

Battle observations. The battle experiments use the Gen 1 OU ruleset. The agent's history contains the available moves and switches, its remaining team, revealed opponent information, and recent interaction. Terminal battle outcome is the principal external success signal. The three selected cases in Appendix G examine how local HP changes, status actions, and switching can precede a later win; their descriptors use different units from both poker returns and aggregate battle win rate.

E AUXILIARY POKER CAPABILITY AND INTERFACE PROBES

E.1 HEADS-UP RAW-MODEL EVALUATION AGAINST SLUMBOT

Each model in this capability probe plays 100 hands against Slumbot. A constrained parser extracts its action; failed extraction invokes the `CHECK_CALL` fallback in the interface. The probe therefore measures the behavior of the complete model–parser–fallback system. It is distinct from the trained ISO comparison.

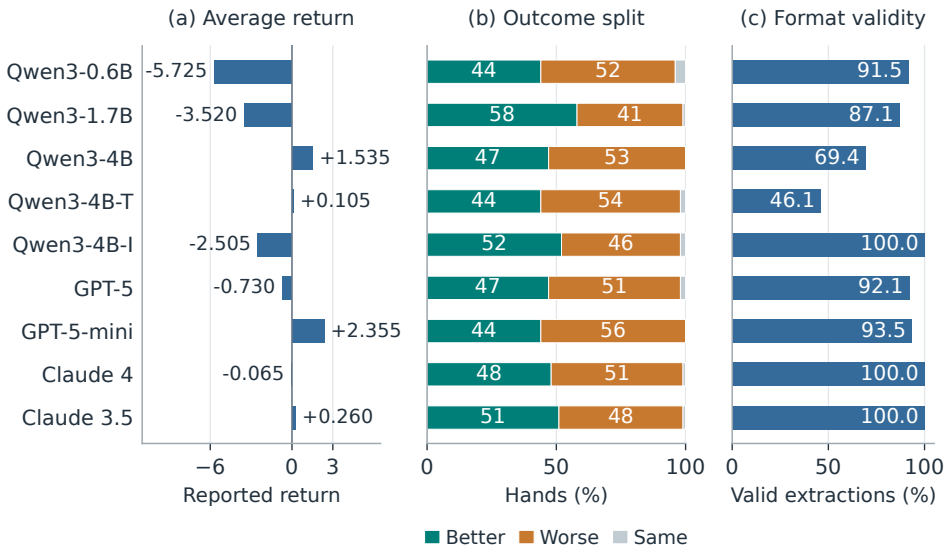


Figure 9: Heads-up reference evaluation. Panel (a) plots the average-return column in Table 24; panel (b) gives its better/worse/same outcome distribution; panel (c) gives successful action extraction as a percentage of parsing attempts. T and I abbreviate Thinking and Instruct. The original figure labels return as BB/h whereas its source table uses BB/100. Panel (a) retains the table values on their reported scale without choosing a conversion; these values are not pooled with primary LTR results.

Table 24: Numerical return and parsing records for the heads-up probe. Return columns preserve the original table values; the unit-label discrepancy in the original sources is described above. The relative column is labeled "vs. Slumbot" in the source and is distinct from the average-return column plotted in panel (a). Format percentages are computed from the displayed success/failure counts.

Model	Average return	Relative column	Parse S/F	Format (%)
Qwen3-0.6B	-5.725	-1.795	280/26	91.5
Qwen3-1.7B	-3.520	$+2.785$	249/37	87.1
Qwen3-4B	$+1.535$	$+1.360$	179/79	69.4
Qwen3-4B-Thinking	$+0.105$	-0.905	107/125	46.1
Qwen3-4B-Instruct	-2.505	$+2.135$	277/0	100.0
GPT-5	-0.730	$+0.120$	281/24	92.1
GPT-5-mini	$+2.355$	$+2.175$	257/18	93.5
Claude 4	-0.065	$+1.010$	145/0	100.0
Claude 3.5	$+0.260$	$+0.280$	186/0	100.0

Action-level denominator. The format statistic is $100S/(S+F)$, where S and F count successful and failed parsing attempts. A hand can contain multiple decisions, so the sum of these counts need not equal the 100-hand evaluation length. For example, Qwen3-0.6B has 306 parsing attempts and a 91.5% extraction rate. The action-level rate and the hand-level outcome distribution are therefore different summaries.

Table 25: Better/worse/same outcome percentages for the heads-up probe. Each row sums to 100%. These retain the source comparison categories and are not substituted for the primary benchmark's hand win rate.

Model	Better (%)	Worse (%)	Same (%)
Qwen3-0.6B	44	52	4
Qwen3-1.7B	58	41	1
Qwen3-4B	47	53	0
Qwen3-4B-Thinking	44	54	2
Qwen3-4B-Instruct	52	46	2
GPT-5	47	51	2
GPT-5-mini	44	56	0
Claude 4	48	51	1
Claude 3.5	51	48	1

Format validity and strategic return. Formatting reliability and strategic return are distinct. Qwen3-4B-Instruct has no recorded parsing failures but a negative average return, while some models with lower extraction rates have a positive average return in this short probe. Reliable action extraction is necessary for interpreting the intended policy, but it does not establish profitable strategic behavior. Conversely, when extraction fails often, the fallback contributes materially to the effective policy and can obscure the model's intended decisions.

The modest hand count and the unit-label discrepancy limit numerical comparisons with the primary experiment. The relevant methodological lesson is to standardize the gateway, track fallback frequency, and evaluate environment payoff separately from output validity. ISO's principal return comparisons use the controlled training and evaluation conditions, rather than this raw-model ranking.

E.2 SIX-PLAYER MODEL AND POSITION DIAGNOSTIC

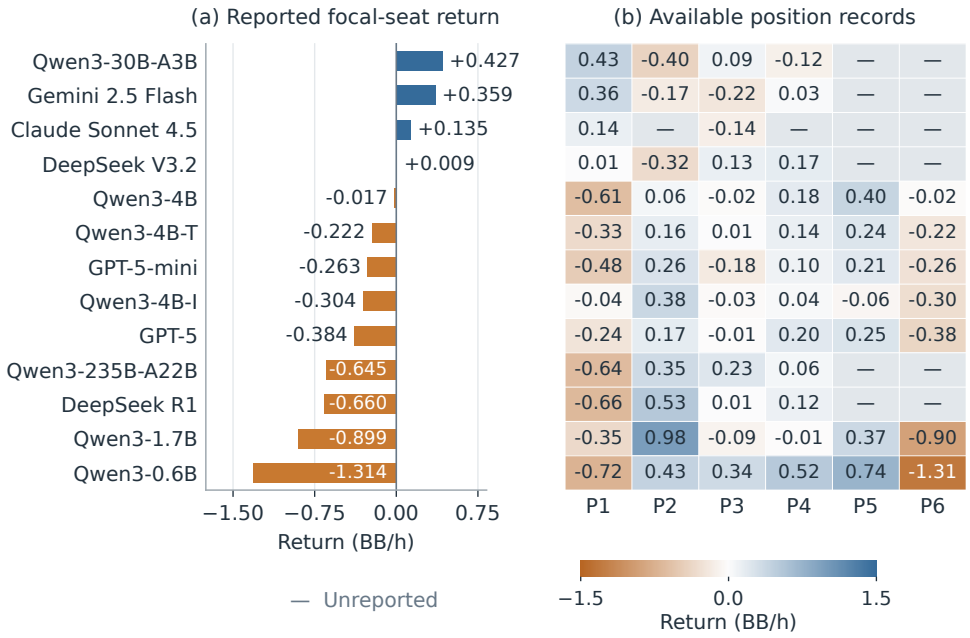


Figure 10: Six-player reference diagnostic. Left: the reported focal-seat model comparison. Right: the available model-by-position return records, aligned to the same model order. The original BB/h convention is retained. Heatmap values reproduce the printed two-decimal annotations; gray cells marked by a dash are unreported, not zero. T and I abbreviate Thinking and Instruct. This is a raw-model and seating diagnostic, separate from the trained ISO comparison.

The heatmap illustrates why a single seat is an incomplete summary of multi-player behavior. For example, the Qwen3-0.6B row contains both positive and negative entries across positions. Position changes what has already been observed, which players remain to act, and how a policy interacts with the other agents. The observed return is consequently a property of a policy within an interaction configuration, rather than a model attribute independent of opponents and seating.

Table 26: Return values annotated in the left panel of Figure 10, preserving the original BB/h axis convention. Model names follow the original figure. These values are not averages reconstructed from the heatmap.

Model	Annotated return
Qwen3-30B-A3B	+0.427
Gemini 2.5 Flash	+0.359
Claude Sonnet 4.5	+0.135
DeepSeek V3.2	+0.009
Qwen3-4B	−0.017
Qwen3-4B-Thinking	−0.222
GPT-5-mini	−0.263
Qwen3-4B-Instruct	−0.304
GPT-5	−0.384
Qwen3-235B-A22B	−0.645
DeepSeek R1	−0.660
Qwen3-1.7B	−0.899
Qwen3-0.6B	−1.314

The heatmap has incomplete position coverage for several models. We do not fill missing cells, infer all-position means, or identify every left-panel result with a common heatmap column. The diagnostic motivates repeated seating permutations in the primary protocol, where the scheduling is shared across matched variants. It does not establish a universal ordering of the raw models or an ISO training effect.

F Training dynamics and context-prediction diagnostics

F.1 Training-step evolution of return

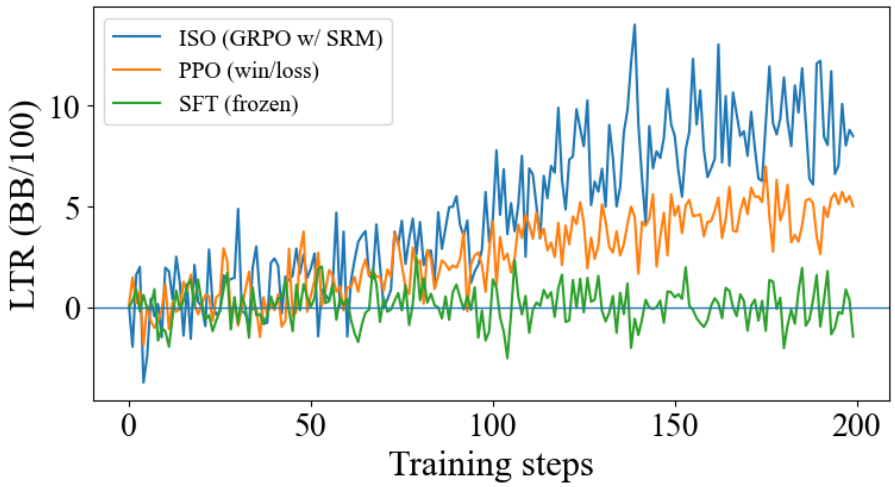


Figure 11: Poker training-return trace for ISO, PPO, and frozen SFT over the displayed training-step range. The plotted trajectories are noisy, and the figure contains no seed-level confidence bands. It is a training diagnostic, separate from the final reference evaluation in Table 10.

The ISO trajectory increases over the displayed range and separates from PPO at later steps, while the frozen-SFT diagnostic remains comparatively flat. This temporal view complements the endpoint tables: it shows how the displayed policies evolve during optimization rather than presenting only their final scores. The local oscillations are substantial, so an upward overall tendency should not be read as monotonic improvement at every update.

The frozen-SFT level in this trace differs from its return in the final reference table. We retain the original diagnostic as its own series and do not calibrate or shift it to match a held-out endpoint. The step axis also does not establish equal elapsed training time: a step can involve different rollout or scoring costs across algorithms. The matched experiment's explicitly shared interaction and optimizer-step budgets provide the relevant budget comparison.

Interpretive limits of the training trace. The trace documents an improvement that develops during the displayed run. A single trajectory does not characterize convergence times across independent runs or establish optimizer convergence. Final-performance evidence is provided separately by the seed-level reference summaries and paired matched controls.

Relation to the contextual analysis. The theoretical comparator is defined for an idealized repeated game with context-indexed learners and specified feedback. A noisy neural training curve is not a regret curve against that comparator. The link between the analysis and this figure is conceptual: learning can benefit from recurring structure without requiring successive policy scores to be monotone. The theorem and the empirical optimization trace are evaluated under their respective assumptions and measurement definitions.

F.2 TRAINING REWARD AND LOCAL VARIABILITY

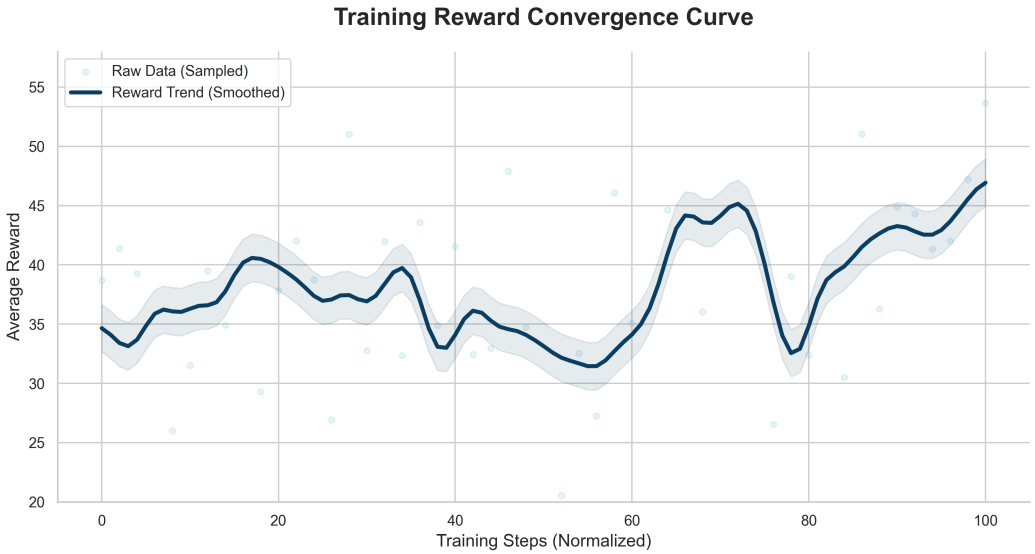


Figure 12: Training-reward diagnostic. Pale markers show sampled rewards, the dark line shows the supplied smoothed trend, and the shaded region is retained from the original plot. The axis is average training reward against normalized training steps. The shading is not treated as a 95% confidence interval because its estimator is not specified.

This diagnostic makes the variability of the optimization signal visible. The smoothed trend is higher near the end than at the beginning, but it also contains pronounced intermediate declines. Raw sampled rewards vary around that trend. The figure is therefore useful for observing the evolution of the training signal, rather than for establishing uninterrupted progress or a formal stability guarantee.

The vertical scale is distinct from BB/100 hands. It must not be compared numerically with the return trajectories or with the final poker LTR. Likewise, the normalized step axis gives an ordering within this run, not the number of environment hands, accelerator hours, or policy updates in another experiment. The original samples and smoothing are retained without inferring additional raw measurements.

Reward trends and estimator validity. Policy optimization changes the distribution on which the SRM is queried, so increasing model scores can coincide with declining accuracy in ranking realized outcomes. Environment return and held-out SRM ranking provide complementary checks on this distribution shift. In Table 17, frozen-SRM correlation decreases across successive policy distributions, and refresh mitigates the decrease. The reward trace alone therefore does not establish continued estimator validity.

Connection to refresh. The reported refresh procedure mixes original and recent on-policy segments equally. This preserves some coverage of the initial behavior distribution while incorporating states encountered by the improved policy. The refresh and uncertainty-gated results in Table 17 are end-to-end evaluations; they are not additional curves inferred from this figure. Their point-estimate improvements should be interpreted at the precision reported in the table.

F.3 Prediction accuracy and the auxiliary routing diagnostic

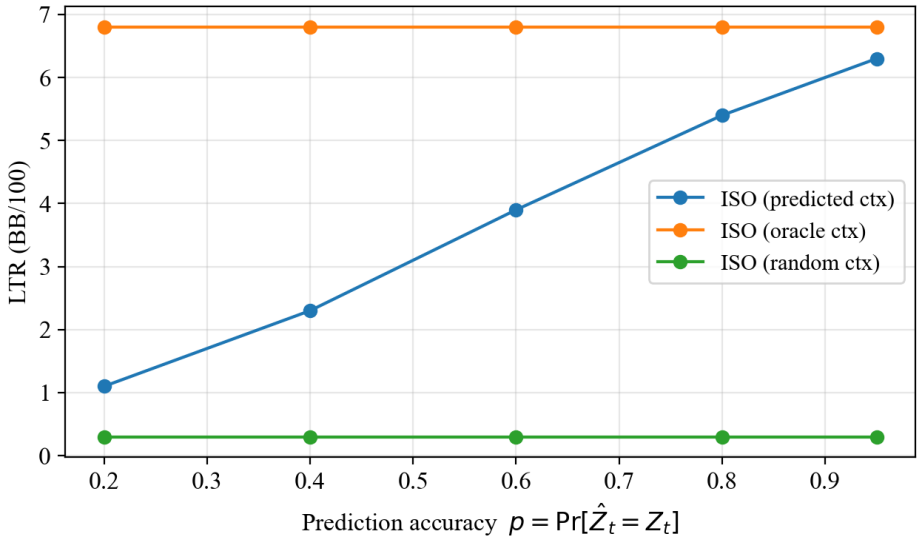


Figure 13: Context-prediction accuracy diagnostic. The predicted-context curve improves toward the oracle-context reference as prediction becomes more accurate; random-context routing remains low. Accuracy on this horizontal axis is distinct from the corruption-level parameter in Table 21. This diagnostic uses a different return scale from the broader reference and matched evaluations.

Table 27: Context-conditioned comparison associated with the auxiliary routing diagnostic. LTR is in BB/100; the final column retains the reported vulnerability estimates. These values are distinct from the later matched controls in the main text.

Method	Use of context	LTR	Vulnerability estimate
GRPO	None	2.7	7.6
GRPO + predicted feature	Shared-policy input	3.1	7.1
ISO-GRPO	Predicted routing	6.3	4.9
ISO-GRPO, oracle	Oracle routing	6.8	4.7

Adding a predicted feature improves return modestly in this diagnostic, while the routed variant obtains a larger increase. The 0.5 BB/100 difference between the reported predicted and oracle results indicates residual headroom under this particular oracle comparator. The oracle is an empirical reference supplied with privileged context information; it is not a universal upper bound on all policies in the game.

The figure's horizontal variable is prediction accuracy, $p_{\text{acc}} = \Pr(\widehat{Z}_t = Z_t)$. Higher values indicate more accurate predictions. By contrast, the p in Table 21 denotes corruption level, where increasing corruption makes the reported setting more difficult. Although both diagnostics address the usefulness of predictive information, one should not identify their axes by $p_{\text{acc}} = 1 - p$ without a specified corruption model and class distribution. Their different return levels also preclude merging their points into a single response curve.

Interpretation of random routing. The low random-context curve shows that splitting updates among components does not by itself supply useful strategic structure. An informative assignment is needed to align histories with useful specialization. The plot nevertheless does not isolate every effect of prediction quality: action selection, data collection, and later updates can all change when

the acting route changes. The controlled P→P versus P→R comparison provides the more direct test of update assignment.

Relation to the error term in the theory. In the idealized analysis, erroneous context predictions incur an additive cost. The empirical rise with accuracy is compatible with that motivation. The figure does not estimate the theorem's loss variation, comparator regret, or learning-rate constants, so it should be read as a mechanism diagnostic rather than a numerical test of the bound. The continuous-context experiments further test whether useful routing persists when contexts overlap and are learned rather than supplied as a fixed taxonomy.

G Additional battle evaluation and long-horizon case analysis

G.1 Training stages against a fixed GPT-4o opponent

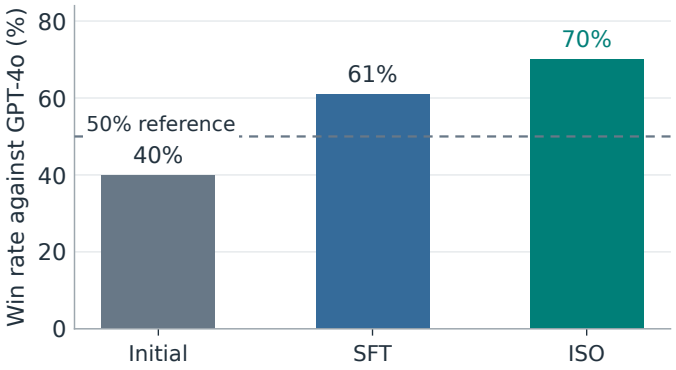


Figure 14: Pokémon training-stage evaluation against GPT-4o: 40% for the initial model, 61% after SFT, and 70% after ISO. The dashed line marks a 50% reference level, not an additional measured random-policy result.

Holding the named opponent fixed makes this figure a direct description of improvement across the displayed stages. The increase from the initial model to SFT is 21 percentage points, and the subsequent increase from SFT to ISO is 9 percentage points. These are percentage-point differences in win rate, not relative percentage increases. The figure includes no uncertainty intervals or additional baseline curve beyond the displayed reference line.

The 70% result is specific to the GPT-4o opponent comparison. It should not replace the 68% ISO result in the main battle table, which reports the separately specified evaluation, or be pooled with the successive-version head-to-head rates in the original main-text figure. A fixed-opponent test, a population evaluation, and a predecessor comparison answer related but different questions about progress.

The improvement after SFT indicates that supervised behavior is a useful starting point. The additional gain after ISO is consistent with an optimization benefit beyond that starting point. The full-episode GRPO control in the main text is particularly relevant because it already receives a terminal battle outcome; delayed terminal feedback is not unique to ISO. The central empirical question is whether the value and routing mechanism improves learning given such feedback.

G.2 Summary of three selected battle trajectories

The following cases illustrate three different ways a policy can trade a local resource for a later opportunity: removing an important opposing threat, using status to shape a later endgame, and accepting damage during switches to preserve favorable team interaction. All three are selected

victories. They are used to explain observed behavior, while aggregate evaluation supplies the performance evidence.

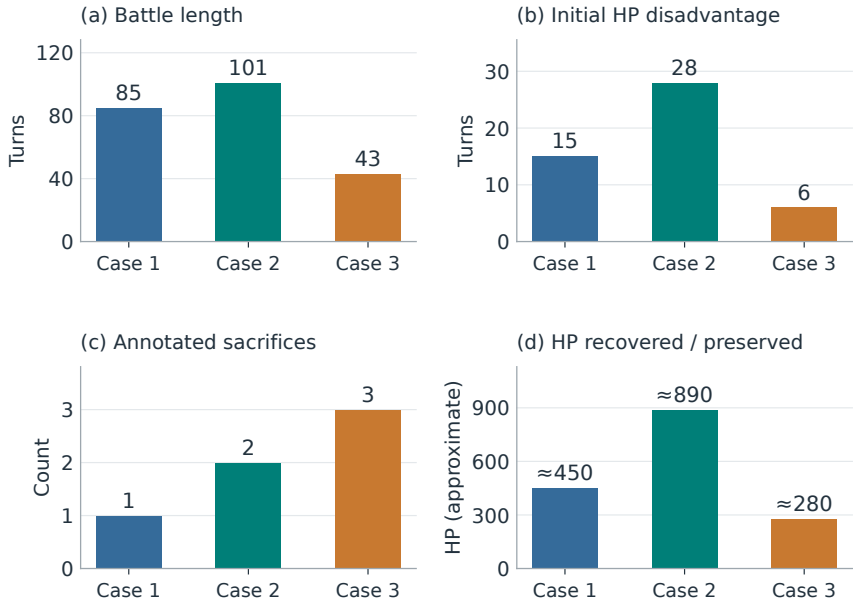


Figure 15: Descriptive comparison of the three Pokémon cases. Panels show battle length, the duration of the initial HP disadvantage, the annotated strategic-sacrifice count, and approximate HP recovered or preserved before victory. These are case descriptors, not estimates over the full battle population.

Table 28: Numerical summary accompanying Figure 15. HP recovered or preserved is the approximate aggregate descriptor reported for each case, not a single team member's maximum HP or a counterfactual causal gain.

Descriptor	Case 1	Case 2	Case 3
Battle length (turns)	85	101	43
Initial disadvantage duration (turns)	15	28	6
Annotated strategic sacrifices	1	2	3
Approx. recovered/preserved HP	450	890	280
Final remaining team members	3–0	1–0	1–0

G.3 Case 1: spending a resource to remove a later threat

The first case centers on the opposing Snorlax as an important late-game threat. The recorded sequence includes switching and recovery before a later exchange in which the agent loses Tauros and the opponent loses Snorlax. The battle eventually ends in a win after 85 turns, with three of the agent's team members remaining. The case descriptor records a 15-turn initial disadvantage phase and one strategic sacrifice.

Local and continuation views. A local resource view penalizes losing Tauros. A continuation view also considers what options remain to the opposing team after Snorlax is removed and how those options match against the agent's surviving team. The relevant strategic object is the resulting team interaction, rather than the identity or immediate HP of one member in isolation. This explains how the same exchange can look costly locally yet be compatible with a favorable continuation.

Observed outcomes and causal limitations. The selected record contains the exchange and a later victory. It does not contain a matched replay in which Tauros is preserved under identical future randomness. The case therefore illustrates a resource-trading pattern; it does not establish that the exchange was uniquely optimal or quantify its causal contribution to the final win. The approximate recovered/preserved HP descriptor summarizes the case and should not be attributed entirely to that one exchange.

G.4 Case 2: status pressure before a delayed endgame

The second case uses a weakened Exeggutor during an extended interaction with the opponent's defensive core. Selected observations include Stun Spore against Chansey at turn 27, later Psychic pressure at turn 55, and the eventual loss of Exeggutor by turn 78. The full battle ends in a win after 101 turns, with one team member remaining. This is the longest of the three examples and has the longest recorded initial disadvantage phase, at 28 turns.

Delayed effects of status actions. A status action can alter the value of later moves without dealing immediate damage. An evaluation based only on the next HP change can consequently misrank actions that affect later speed, action availability, or matchup pressure. In the reported narrative, Exeggutor's role extends beyond its immediate survival: it helps shape an endgame that is completed by the remaining team.

The case makes the temporal distinction concrete. The enabling action occurs well before the terminal outcome, and intervening switches and recovery separate the two. SRM is intended to provide a learning signal based on such continuations rather than only on local damage. The example does not reveal the SRM's internal reasoning; it identifies the kind of delayed outcome that its trajectory target is meant to value.

G.5 Case 3: switching while accepting an HP disadvantage

The third case contains a sequence of switches involving Golem, Moltres, and Snorlax during early interaction with the opposing Tauros. Several team members take damage along the way. The recorded opposing Tauros HP falls to 39% at turn 7 and 9% at turn 8. The eventual victory occurs at turn 43 with one team member remaining. The summary records three strategic sacrifices and a six-turn initial disadvantage phase.

Team-level allocation. The relevant choice is not simply whether to avoid damage at the current turn. It is also which team member should absorb a risk, which matchups remain available afterward, and whether the resulting position constrains an opposing threat. A policy may distribute losses across several members to retain a useful continuation. Such behavior can be coherent even when the sum of immediate HP changes is unfavorable.

The selected sequence is consistent with this team-level interpretation, but it does not prove that every switch was the best response to the hidden opponent state. The predictor acts before the opponent's next choice is observed. Retrospective knowledge of how the line unfolded must therefore be kept separate from the information used to select the switches.

G.6 Relationship between case studies and aggregate evaluation

Table 29: Interpretive comparison of the three battle cases. The continuation role is an explanation of the recorded behavior, not an independently estimated treatment effect.

Case	Local resource cost	Possible continuation role	Observed endpoint
Threat removal	Loss of a valuable attacker	Remove an opposing late-game threat	Win after 85 turns
Status setup	Damage and eventual loss of Exeggutor	Reshape the endgame for teammates	Win after 101 turns
Counter-switching	Damage across several members	Preserve matchups and pressure Tauros	Win after 43 turns

The examples show why a single local score, such as immediate HP preservation, does not fully describe strategic value. They also show why sacrifice count alone is not an objective: sacrificing more can be useful or harmful depending on the continuation. The selected cases contain only successful outcomes and do not estimate how often the same patterns fail. Aggregate battle win rates, the full-episode GRPO control, and the broader behavioral diagnostics are needed to assess whether the training method improves performance beyond these illustrations.

Across cases, the mechanism-level interpretation is the same as in poker: a local action changes the distribution of future interaction, and the useful learning signal depends on that continuation. The domains differ in their observations and payoffs. ISO shares a prediction-aware learning structure across them; these examples do not imply that a poker context taxonomy or its trained adapters can be transferred unchanged to battles.

H Extended poker decision cases and their interpretation

This section reproduces the essential cards, observed action sequences, and policy choices from three representative poker decision examples. The examples are abbreviated decision snapshots, not counterfactual evaluations. Their raw interface uses 500-chip starting stacks and blinds of 1/2 chips, corresponding to a 250-BB starting stack; this differs from the primary 100-BB evaluation. We therefore use them to illustrate decisions without pooling their chip amounts with the benchmark returns.

H.1 Case 1: declining an unfavorable initial investment

Table 30: Selected fields from the pre-flop decision example. Card letters use the original interface convention: D for diamonds and S for spades.

Field	Recorded value
Private cards	`7D, 2S`
Board	Empty; pre-flop
Table	Six players; hero position 3; button position 0
Prior action history	Empty in the supplied snapshot
Pot	3 chips from the posted blinds
ISO action	`FOLD`
Baseline action	`RAISE_POT`

The contrast is between declining a weak starting hand and taking an aggressive action that could win the blinds immediately. The ISO choice avoids committing additional chips in the recorded situation. The baseline choice pursues initiative despite the weak private cards and the players still able to act.

This is intentionally a simple example. Folding this hand in the stated situation is not by itself evidence of a novel long-horizon capability. Its role is to expose a difference in the policies' willingness to pursue immediate initiative. The broader return and pot-size results are needed to determine whether the resulting selectivity is beneficial over many decisions.

In a return-based interpretation, the foregone chance to win a small pot must be weighed against the distribution of future losses after entering the hand. This is a statement about the decision objective, not a measured action-value difference for this particular snapshot. No counterfactual return for the baseline raise is reported in the case record.

H.2 Case 2: delaying aggression with a very strong hand

Table 31: Selected fields from the strong-hand flop example. The action token `RAISE_POT` is retained from the interface; its executable betting action is resolved by the gateway for the current state.

Field	Recorded value
Private cards	`AS, AD`
Board	`AH, 7C, 2D`
Pre-flop history	Hero raises; the big blind calls; other players fold
Current public action	The big blind checks on the flop
ISO action	`CHECK`
Baseline action	`RAISE_POT`

The agent holds three aces using its private pair and the public ace. The selected ISO action forgoes immediate pot growth. A plausible continuation-based interpretation is that checking can preserve later action from weaker holdings or bluffs. The baseline instead uses a pot-sized aggressive action to build value immediately.

The action comparison depends on the opponent model. Against some response patterns, immediate betting may have higher value; against others, a delayed line may retain more profitable future interaction. The snapshot does not report an opponent-range distribution or an independently evaluated action-value table. We therefore describe the chosen line and its possible continuation role without declaring checking universally optimal on this board.

This case also clarifies why strategic optimization should not be equated with maximizing immediate activity. A passive action at one state can be part of a policy that seeks greater later return. Conversely, passivity alone does not demonstrate planning. The interpretation becomes informative when considered with the population-level evidence and the actual continuation, rather than inferred solely from the action's name.

H.3 Case 3: revising the value of a strong private pair

Table 32: Selected fields from the paired-board decision example. The history is summarized as action order without reconstructing a complete chip ledger from the abbreviated snapshot.

Field	Recorded value
Private cards	`JH, JD`
Board	`KC, KH, 4S`
Pre-flop history	An opponent raises; hero calls on the button
Flop history	Opponent makes a half-pot action; hero raises; opponent uses `ALL_IN`
ISO action	`FOLD`
Baseline action	`CHECK_CALL`

Pocket jacks are a strong starting hand, but the public board and subsequent aggression change the relevant decision. ISO folds after the observed re-escalation, while the baseline continues. A

continuation-based explanation conditions on how often the opponent takes this line with strong holdings versus bluffs. The private pair's pre-flop strength is not sufficient to settle that question.

The opponent's private cards are not part of the recorded pre-action observation. Consequently, the case does not establish that the opponent held a king or that calling had negligible equity. The choice depends on the inferred range and the cost of continuing. The example is useful because it shows the policy responding to accumulated public evidence rather than treating the initial hand category as a fixed prescription.

Prior investment is also distinct from the value of the next decision. Chips already committed cannot be recovered simply by continuing, and the comparison concerns future costs and benefits from the current state. This is a general interpretation of the decision; the abbreviated record does not supply a numerical counterfactual estimate for either action.

H.4 Connecting qualitative decisions with quantitative evidence

Table 33: How the poker cases relate to the paper's aggregate diagnostics. The final column identifies complementary evidence rather than a causal estimate for the individual action.

Case	Observed policy contrast	Strategic question	Complementary evidence
Weak initial hand	Fold versus raise	Is immediate initiative worth later exposure?	Return versus hand win rate
Strong flop hand	Check versus aggressive action	Can deferring extraction preserve future value?	Later-street return accounting
Paired-board escalation	Fold versus call	How should new opponent evidence change commitment?	Conditional return in larger pots

The three cases describe different decisions rather than a universal bias toward folding or checking. The first declines entry, the second retains a strong hand without immediate aggression, and the third abandons a hand after informative public action. Their common feature is that the value of the action depends on possible later interaction, not only on an immediate win indicator or a fixed hand-strength label.

The aggregate diagnostics provide complementary measurements. Street-wise return associates ISO's advantage with later decisions; pot-size conditioning examines outcomes in higher-payoff subsets; the sacrifice analysis relates locally unfavorable assessments to later return. Each has a different unit of analysis and is subject to policy-induced selection. Taken together with matched optimization controls, they support the interpretation in the main text without converting these selected snapshots into additional benchmark trials.

Finally, these examples explain behavior at the level of observable decisions. They do not establish that the model executes an explicit search algorithm, represents an opponent's private state correctly, or produces a faithful verbal account of its internal computation. The claim is about learned strategic performance and its association with continuation-sensitive behavior, which is the level supported by the experiments and analysis.