Enhancing Language Agent Strategic Reasoning through Self-Play in Adversarial Games

Anonymous ACL submission

Abstract

Language agents often struggle with strategic reasoning in adversarial games. A promising approach is learning from game interactions automatically, but unlike static environments, selecting appropriate opponents in adversarial settings significantly impacts learning—a factor that remains underexplored. We propose **S**tep-level poli**C**y **O**ptimization through **P**lay-**A**nd-**L**earn (**SCO-PAL**), and conduct systematic analysis of opponent selection, finding that self-play is most effective for improving strategic reasoning. With SCO-PAL and self-play, we improve the average win rate from 32.17% (base model) to 50.08%, achieving 54.76% against GPT-4 across six games. The learned skills also generalize to unseen games and broader reasoning tasks, demonstrating unique advantages of LLM-based agents.

1 Introduction

Large Language Model (LLM)–based agents have achieved remarkable success in a variety of language-centric tasks—including web navigation (Zhou et al., 2023; Deng et al., 2024; Putta et al., 2024; Iong et al., 2024), embodied interactions (Wang et al., 2023; Zhu et al., 2023; Lin et al., 2024) and tool invocation (Qin et al., 2023; Shen et al., 2024). Unlike classic RL agents that operate over a small, well-defined set of discrete or continuous controls and clear state transitions, LLM agents must generate high-dimensional natural-language actions conditioned on long, unstructured textual contexts. This fundamentally complicates strategic reasoning in adversarial games, where success hinges on planning over many moves and reacting to an opponent's evolving strategy under sparse, delayed rewards. Benchmarks such as `GTBench` (Duan et al., 2024) (e.g. `Breakthrough`, `Connect Four`, `Nim`) expose these challenges by testing both the depth of planning and the fluency of language-based decision making.

To mitigate the shortcomings in the strategic reasoning of language agents, there are mainly two paths. One is imitating a strong strategy directly inspired by Chen et al. (2023); Zeng et al. (2023). This way is simple and effective, but gathering expert-level data is time-consuming, costly, and challenging to scale. Therefore, studies on another path focus on **play-and learn** paradigm, which enriches data through interaction and learning from the interactions. For example, Song et al. (2024); Xiong et al. (2024); Xi et al. (2024) allow agents to interact autonomously with environments, gathering numerous data for improvement. However, unlike interacting with unchanging conditions in static environments, the agent in adversarial games faces dynamic and diverse opponents, which brings additional challenges. In adversarial scenarios, Wang et al. (2024b); Cheng et al. (2024) enhance the strategic reasoning capabilities of agents with self-play, but they lack analysis of opponent selection in adversarial environments.

Opponent selection has long been recognized as a key factor in adversarial training within the RL community, where methods such as self-play (Silver et al., 2017), opponent sampling (Arulkumaran et al., 2019), and curriculum-based progression (Bengio et al., 2009) are widely used. However, these techniques are designed for structured, low-dimensional action spaces, whereas LLM-based agents operate in open-ended language spaces where each "action" is a natural language utterance. This introduces unique challenges: opponents not only determine win–lose outcomes, but also shape the semantic distribution and reasoning trajectories encountered during learning. Too strong opponents disrupt learning with mismatched strategy distributions; too weak opponents provide limited feedback and cause premature convergence; uneven skill levels create imbalanced experience distributions. Thus, a well-balanced opponent selection strategy is essential for effective training.

To systematically study how opponents influence learning, we propose **S**tep-level poli**C**y **O**ptimization through **P**lay-**A**nd-**L**earn (**SCO-PAL**). SCO-PAL consists of three stages: (1) game interaction with a designated opponent, (2) step-level reward estimation via Monte Carlo methods, and (3) two-stage strategy refinement using Behavioral Cloning followed by Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024).

Using SCO-PAL, we systematically analyze opponent selection and find that self-play is most effective. On six games from GTBench, SCO-PAL with self-play improves the average win rate from 32.17% to 50.08% (a 17.91 percentage point improvement), achieving 54.76% against GPT-4. The learned skills also generalize to unseen games (+8%) and broader reasoning tasks (+4.4% on BBH temporal reasoning), while Chain-of-Thought prompting provides an additional 2.4% gain—confirming unique advantages of LLM-based agents.

Our contributions are: (1) systematic analysis of opponent selection in LLM-based adversarial learning, finding self-play most effective; (2) SCO-PAL, a step-level policy optimization framework for strategic reasoning; (3) comprehensive experiments showing significant improvement (32.17% $\rightarrow$ 50.08%) and generalization to unseen tasks.

2 Related Work

LLM Agents LLMs increasingly power agents for complex tasks (Zhou et al., 2023; Wang et al., 2023; Deng et al., 2024; Putta et al., 2024; Shen et al., 2024). Methods either fine-tune on expert trajectories (Zeng et al., 2023; Chen et al., 2023) or adopt play-and-learn paradigms (Xi et al., 2024; Song et al., 2024; Xiong et al., 2024; Wang et al., 2024a), optimizing via trajectory-level or step-wise feedback. However, most assume static environments; we focus on adversarial games requiring adaptation to dynamic opponents.

Agents in Adversarial Games Strategic games require reasoning and adapting to opponents' behaviors. Benchmarks (Huang et al., 2024; Costarelli et al., 2024; Duan et al., 2024) show LLM agents fall short of human-level strategic reasoning. We use GTBench (Duan et al., 2024) as our testbed. Existing training methods either use successful trajectories (Wang et al., 2024b) or customized discounted rewards (Cheng et al., 2024), but lack systematic study of opponent selection. Our SCO-PAL uses step-level reward estimation and achieves higher win rates with self-play.

3 Preliminary

3.1 Game Modeling

The strategic adversarial games from GTBench can be formalized as an episodic Markov Decision Process (MDP), where two players $\mathcal{P} \in \{\mathcal{P}_1, \mathcal{P}_2\}$ take turns performing actions in a gaming environment. Each game is defined by four elements: the state $\mathcal{S}$, representing the current game configuration; the observation $\mathcal{O}$, the visible part of the state for each player; the action space $\mathcal{A}$; and the transition function $T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$. In state s_i, player $\mathcal{P}_i$ observes o_i and takes an action $a_i \sim \pi_i(\cdot \mid s_i, o_i)$, which transitions the game to state s_{i+1}. The environment returns a reward r when a terminal state is reached. Each player aims to maximize their expected cumulative reward by optimizing their policy π_i. In this work, we focus on improving the policy π_i of the language agent $\mathcal{P}_1$, enabling it to adjust its strategy and achieve higher rewards through interaction.

3.2 Behavioral Cloning

Behavioral cloning (BC) is a supervised learning approach where an agent learns to imitate expert behavior by training on expert-level datasets. Given a dataset $(x, y) \in \mathcal{D}$, the objective of BC is to minimize the following loss:

$$L_{BC}(\pi_\theta; \mathcal{D}) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \pi_\theta(y|x)], \quad (1)$$

where π_θ is the model to be updated.

3.3 Kahneman-Tversky Optimization (KTO)

Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) is an offline reinforcement learning method for preference optimization. Unlike Direct Preference Optimization (DPO) (Rafailov et al., 2024), KTO requires only an output and a binary reward indicating desirability for a given input, eliminating the need for pair-wise preference data. This is advantageous in asymmetric scenarios like werewolf (Ye et al., 2025), where creating pair-wise comparisons is challenging. Given the datasets $(x, y) \in \mathcal{D}$, KTO optimizes the policy π_θ using the following loss:

$$r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \tag{2}$$

$$z_0 = \mathbb{E}_{(x,y)\sim\mathcal{D}}[KL(\pi_\theta(y|x)\|\pi_{ref}(y|x))] \tag{3}$$

$$v(x, y) = \tag{4}$$

$$\begin{cases} \lambda_D \sigma\big(\beta(r_\theta(x, y) - z_0)\big) & \text{if } y \sim y_{\text{desirable}} \mid x \\ \lambda_U \sigma\big(\beta(z_0 - r_\theta(x, y))\big) & \text{if } y \sim y_{\text{undesirable}} \mid x \end{cases} \tag{5}$$

$$L_{KTO}(\pi_\theta; \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\lambda_y - v(x, y)] \tag{6}$$

Here, λ_D and λ_U are hyperparameters representing the losses for desirable and undesirable outcomes, respectively. The parameter λ_y corresponds to λ_D when y is desirable and λ_U when y is undesirable.

4 SCO-PAL

To realize learning from interaction, a proper framework is necessary. In two-player adversarial games, both players engage in strategic reasoning and act to defeat each other, while the environment autonomously determines the game outcome and allocates rewards based on ending conditions. This removes the need for external labeling and facilitates the accumulation of datasets enriched with reward signals. Leveraging this, we design a play-and-learn framework SCO-PAL (Figure 1), using large-scale game interaction to capture strategic optimization directions and improve strategy.

4.1 Game Interaction

The play-and-learn paradigm boosts agents by learning from game interaction with opponents. To realize this, we first conduct interactions between agents in the gaming environments, named Game Interaction (Stage I).

During game interaction, player $\mathcal{P}_i \in \{\mathcal{P}_1, \mathcal{P}_2\}$ performs actions a_t according to policy $\pi_\theta(\cdot \mid s_i, o_i)$ at each time step t. Each action a_t transitions the game from state s_t to s_{t+1}, and the environment concludes the game by determining the ending condition and providing a reward r, indicating a win, loss, or tie. For player $\mathcal{P}_i$, a trajectory τ is denoted as $(s_0, a_1, s_1, a_2, \ldots, a_n, s_n, r)$, where r signifies the outcome reward. Through interaction, we collect a set of trajectories $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ for strategy refinement.

4.2 Step-Wise Reward Estimation

The environment provides only outcome-level rewards, leaving intermediate actions unlabeled. To address this, we design a step-level reward estimation method in Step-Wise Reward Estimation (Stage II) and compare different estimation strategies in § 7.2.

Step-wise reward estimation identifies whether an action is advantageous or disadvantageous, guiding subsequent strategy refinement. We apply Monte Carlo Estimation over interaction trajectories $\mathcal{T}$, estimating the reward $r(s_i, a_i)$ of each $(s_i, a_i) \in \tau$ by its empirical win rate. Specifically, we count total occurrences N_{all} and winning occurrences N_{win} of (s_i, a_i) in $\mathcal{T}$, and compute: $r(s_i, a_i) = \frac{N_{win}}{N_{all}}$.

4.3 Strategy Refinement

After reward estimation, we obtain trajectories with step-level rewards for training. In Strategy Refinement (Stage III), we propose a two-stage strategy improvement method: (1) **Behavioral Cloning** adapts the agent to the game environment using state-action pairs (s_i, a_i) with rewards $r(s_i, a_i)$ above a predefined threshold δ as advantageous actions for training. (2) **KTO Optimization** formulates learning as preference optimization. We label state-action pairs as advantageous (desirable) or disadvantageous (undesirable) based on $r(s_i, a_i)$ and δ, encouraging the agent to prefer advantageous actions while avoiding disadvantageous ones.

In experiments, we find that Behavioral Cloning enables fast adaptation, while KTO Optimization helps the agent effectively distinguish between action qualities, significantly improving win rates in downstream competitions.

5 Analysis on Opponent Selection

5.1 Setup

Game Interaction Opponents To study how opponent strength affects learning, we introduce level-controllable symbolic opponents ranging from random (weak) to MCTS-based (strong). The strength of MCTS is adjusted by varying the maximum number of simulations (`max_simulation`)[1], set to 5, 10, 100, 200, 500, and 1000.

Game Benchmark We select 6 games from `GTBench` (Duan et al., 2024): (1) `Breakthrough`, (2) `Connect Four`, (3) `Tic-Tac-Toe`, (4) `Kuhn Poker`, (5) `Liar's Dice`, and (6) `Nim`[2]. Each game interaction involves 1,000 episodes, and each evaluation is conducted over 100 episodes per opponent.

[1] Details in Appendix B.1.

[2] Game introductions are in Appendix C.1.

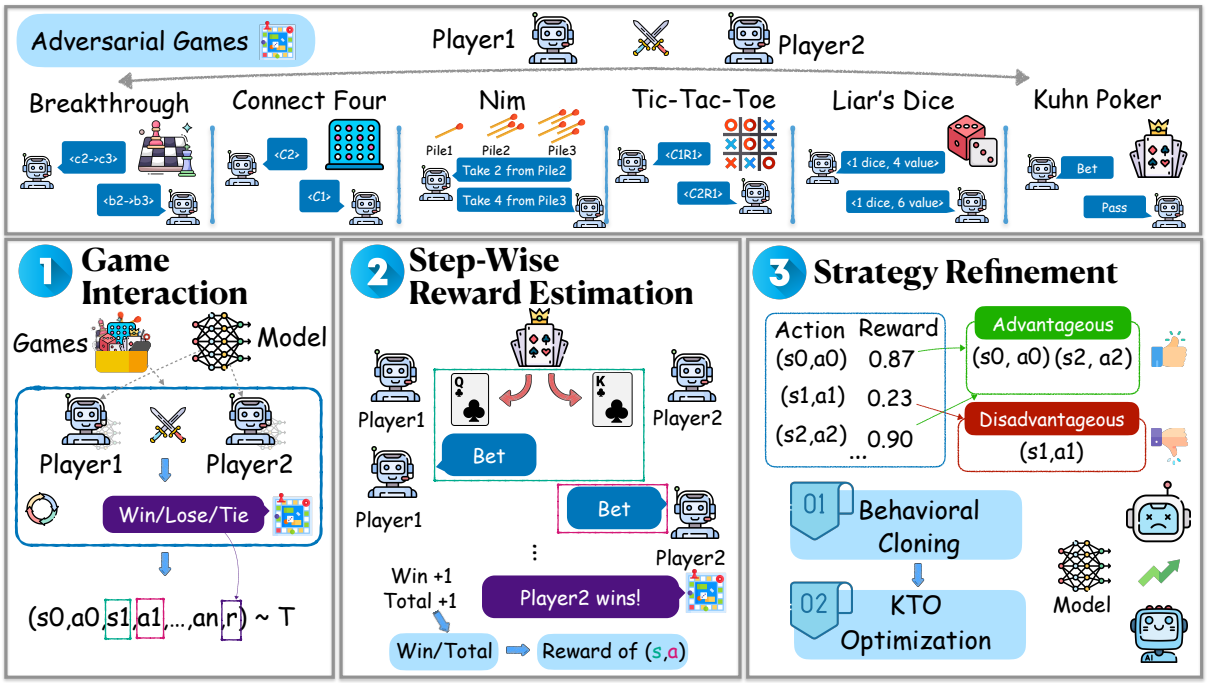


Figure 1: Overview of our play-and-learn method, SCO-PAL. SCO-PAL enables language agents to improve through game interaction via a three-stage process: Stage I Game Interaction assigns an opponent and conducts large-scale gameplay; Stage II Step-Wise Reward Estimation computes step-level rewards; Stage III Strategy Refinement updates the policy via behavioral cloning and KTO Optimization.

To mitigate first-player advantage, players alternate roles and play an equal number of games.

Model Setting We use Qwen2-7B-Instruct (qwe, 2024) as the base model. During game interaction, the temperature is set to 0.7 (see § A.4 for ablation). For evaluation, we follow GTBench (Duan et al., 2024) and set the temperature to 0.2. We also demonstrate the effectiveness of SCO-PAL on other models (*e.g.*, Mistral) in the Appendix A.10.

Parameters The threshold for identifying advantageous actions is 0.5. In Behavioral Cloning, we set the learning rate to 1e-6, batch size to 2, warm-up ratio to 0.1, and train for 5 epochs. For KTO, the batch size is 2, the gradient accumulation steps are 8, the learning rate is 1e-6, and training also lasts 5 epochs. The loss weights λ_D and λ_U are set such that $\lambda_D n_D = \lambda_U n_U$, where n_D and n_U are the numbers of advantageous and disadvantageous samples; the larger weight is set to 1.0, and the smaller scaled proportionally. All experiments are conducted on 8 NVIDIA A100 GPUs (80GB each).

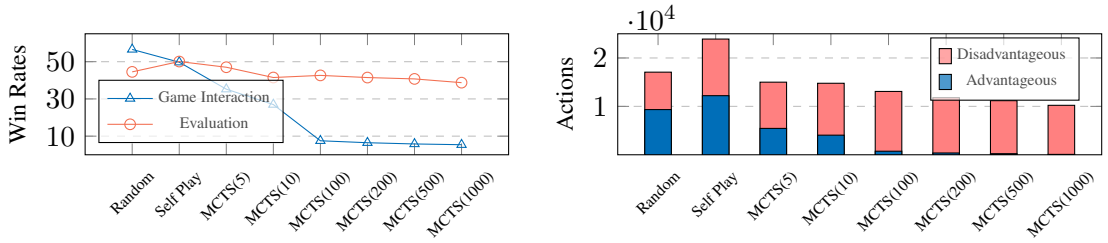
Metrics We evaluate performance using win rate. For each game, we record outcome counts N_{win}, N_{lose}, and N_{tie}, and compute the win rate as: $w = \frac{N_{win}+0.5N_{tie}}{N_{win}+N_{lose}+N_{tie}}$.

Evaluation Opponents We evaluate against four opponents: (1) *Random*, selecting actions at random; (2) *MCTS*, configured with 1 rollout, UCT parameter $c = 2$, and max simulations = 1000; (3) *GPT-3.5* (GPT-3.5-turbo-0125) (OpenAI et al., 2024), an OpenAI language model; (4) *GPT-4* (GPT-4-turbo-2024-04-09) (OpenAI et al., 2024), a more advanced OpenAI model.

5.2 Results

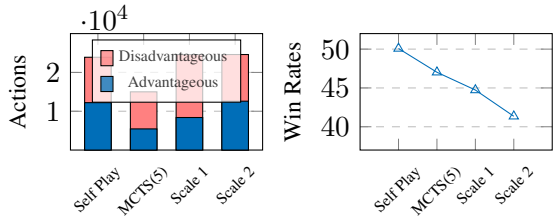
Opponents affect action quantity and distribution. As shown in Figure 2(a), the agent achieves 56.59% win rate against random and nearly 50% in self-play, dropping to 5.40% against strong MCTS as `max_simulation` increases. Figure 2(b) shows that action counts decrease against both weak and strong opponents due to predictable outcomes reducing action diversity. Self-play yields the most balanced advantageous-to-disadvantageous ratio, while this ratio becomes increasingly skewed against stronger opponents.

Self-play is most effective. As shown in Figure 2(a), training with self-play yields the largest performance improvement, outperforming both weak and strong opponents. This mirrors the action count trends: diverse data with balanced advantageous/disadvantageous ratios leads to more



(a) Win rates in game interaction and evaluation. (b) Number of actions collected from different opponents.

Figure 2: Analysis of opponent selection in game interaction. (a) *Game Interaction* shows the base model's win rates against different opponents, while *Evaluation* shows the win rates of the model trained with SCO-PAL against four downstream opponents. (b) displays the number of actions collected when interacting with different opponents. The number in MCTS(·) indicates the `max_simulation` setting.



(a) Number of actions. (b) Win rates after training.

Figure 3: Evaluation of different scaling strategies for balancing advantageous and disadvantageous samples. (a) shows the number of actions obtained under each setting. (b) reports the win rates against four downstream opponents using agents trained on the corresponding data.

effective optimization. Quantitative analysis is in Appendix A.9.

Scaling data does not help. To test whether data quantity compensates for quality, we upsample MCTS(5) data using two strategies (Figure 3(a)): **Scale 1** matches total sample count while preserving the original ratio; **Scale 2** directly aligns advantageous/disadvantageous counts with self-play. As shown in Figure 3(b), both strategies decrease performance—Scale 1 causes overfitting, Scale 2 disrupts optimization. This confirms that data scaling cannot compensate for lack of diversity; balance must be intrinsic to the data.

Combining with strong-opponent imitation. We also explore combining SCO-PAL with imitation learning from strong opponents. Results in Appendix A.1 and A.2 show that imitation can complement SCO-PAL for further improvements.

6 Experiments on SCO-PAL with Self-Play

Analysis in § 5 demonstrates that self-play is the most effective way to conduct play-and-learn in adversarial games. To fully explore the performance of SCO-PAL with self-play, we compare SCO-PAL under self-play with existing play-and-learn methods, both against downstream opponents, and conduct a head-to-head evaluation in this section.

6.1 Setup

Opponents In game interaction, we set the opponent as the language agent itself (self-play setting in § 5). Other settings are the same as these in § 5.

Baselines We select three baselines: (1) Base model: we use Qwen2-7b-Chat as the base model; (2) Behavioral Cloning (BC): fine-tuning the model on successful trajectories from game interactions inspired by Wang et al. (2024b); (3) SPAG (Cheng et al., 2024): that updates the policy by hand-crafted rewards with a discounted coefficient [3].

6.2 Main Results

In Table 1, the base model's average win rate against the four downstream opponents is generally lower, only slightly beating random with a rate of 52.44%. After training, SCO-PAL achieves the most significant improvement, 50.08%, surpassing 43.50% of BC and 41.70% of SPAG, and defeating GPT-4 with an average winning rate of 54.76%. The reason behind the improvement is that both BC and SPAG struggle with learning actions due to inaccurate reward estimation and insufficient optimization methods. For example, in

[3] Detailed introduction of baselines are in Appendix B.2.

Methods	Opponents	Breakthrough	Connect Four	Tic-Tac-Toe	Kuhn Poker	Liar's Dice	Nim	Avg.
Base	GPT-4	7.22	46.46	44.50	20.90	20.31	31.00	28.40
	GPT-3.5	57.65	46.32	55.26	18.75	21.54	53.00	42.09
	Random	66.67	61.62	63.64	39.06	39.66	44.00	52.44
	MCTS	2.74	0.00	6.00	20.90	4.84	0.00	5.75
	AVG	33.57	38.60	42.35	24.90	21.59	32.00	32.17
BC	GPT-4	26.00	32.50	64.00	45.00	57.00	31.00	42.58
	GPT-3.5	52.63	10.10	100.00	45.00	68.69	44.68	53.52
	Random	68.09	66.00	78.00	61.00	83.00	50.00	67.68
	MCTS	3.30	0.00	0.00	45.00	13.00	0.00	10.22
	AVG	37.50	27.15	**60.50**	**49.00**	55.42	31.42	43.50
SPAG	GPT-4	29.87	42.00	33.00	41.00	57.00	39.00	40.31
	GPT-3.5	30.14	29.29	98.00	37.00	64.65	44.00	50.51
	Random	71.43	65.00	75.00	45.45	82.00	57.00	65.98
	MCTS	0.00	0.00	0.00	45.00	15.00	0.00	10.00
	AVG	32.86	34.07	51.50	42.11	54.66	35.00	41.70
SCO-PAL	GPT-4	52.53	62.63	71.43	45.00	60.00	37.00	54.76
	GPT-3.5	60.42	72.00	54.64	45.00	80.00	77.00	64.84
	Random	85.88	76.00	76.02	35.00	89.00	59.00	70.15
	MCTS	2.33	0.00	10.10	45.00	6.00	0.00	10.57
	AVG	**50.29**	**52.66**	53.05	42.50	**58.75**	**43.25**	**50.08**

Table 1: Win rates against four opponents on six games. The best results are **bolded**, and the second best ones are underlined. Statistical significance tests are reported in the Appendix A.6.

BC, successful trajectories might include suboptimal actions. In SPAG, manually specified coefficients can become inaccurate and not general due to changes in environments. SCO-PAL uses more general reward estimates and finer-grained reward signals, using KTO optimization, leading to more comprehensive learning. We also analyze regret in § 7.2, showing that SCO-PAL reduces regret and learns meaningful strategies. We further conduct head-to-head evaluation between methods in Appendix A.11, where SCO-PAL achieves over 50% win rate against all baselines.

6.3 Generalization Capabilities

A key advantage of LLM-based agents is their potential for broad generalization. We evaluate this from two perspectives: transfer to unseen games and transfer to general reasoning tasks.

Zero-shot Transfer to Unseen Games To explore whether strategic reasoning transfers beyond training games, we evaluate SCO-PAL on three unseen games from `GTBench`: `Blind Auction`, `Pig`, and `Prisoner's Dilemma`[4]. As shown in Table 4, SCO-PAL achieves an average improvement of 8% across these games without any task-specific training. The agent demonstrates the ability to make reasonable decisions and adapt strategies based on game states, indicating that SCO-PAL cultivates transferable strategic reasoning skills rather than game-specific heuristics.

[4]Game descriptions are in Appendix C.2.

Category	Benchmark	Base	SCO-PAL	Δ
Math	GSM8K	81.05	80.36	-0.69
Code	MBPP	52.60	54.40	**+1.80**
Complex	BBH (avg.)	56.55	56.24	-0.31
BBH subtasks with notable improvements:				
	temporal_seq	56.80	61.20	**+4.40**
	colored_objects	79.20	83.20	**+4.00**
	object_counting	64.40	66.40	+2.00
	multistep_arith	48.80	50.40	+1.60

Table 2: Generalization to reasoning benchmarks. SCO-PAL maintains performance on math, code, and complex reasoning tasks. Notably, subtasks involving temporal reasoning, state tracking, and multi-step planning show improvements, suggesting strategic game training may enhance structured reasoning capabilities.

Transfer to General Reasoning Benchmarks Beyond games, we examine whether strategic training affects broader reasoning capabilities. As shown in Table 2, SCO-PAL not only preserves but **enhances structured reasoning**: mathematical reasoning remains stable (GSM8K: -0.69%), while code generation improves (MBPP: **+1.80%**). Most notably, BBH subtasks requiring *planning and state tracking*—core skills in strate-

Method	Avg. Win Rate
SCO-PAL (w/o CoT)	50.08%
SCO-PAL (w/ CoT)	**52.48%**
Improvement	+2.40%

Table 3: Effect of Chain-of-Thought (CoT) reasoning. CoT enables the agent to leverage language-based reasoning, leading to improved strategic decisions.

	Auc.	Pig	Pri.	Avg.
Base	**71.00**	88.00	37.00	65.00
SCO-PAL	67.00	**92.00**	**60.00**	**73.00**

Table 4: Evaluation on unseen games. **Auc.** refers to `Blind Auction`; **Pig** refers to `Pig`; **Pri.** refers to `Prisoner's Dilemma`.

gic games—show consistent improvements: temporal reasoning (**+4.40%**), object state tracking (**+4.00%**), and multi-step arithmetic (**+1.60%**). This suggests that adversarial game training cultivates transferable reasoning patterns: planning ahead, maintaining state representations, and reasoning about sequential decisions. These findings demonstrate a unique advantage of LLM-based agents: strategic skills learned through self-play transfer positively to related reasoning domains.

6.4 Effect of Chain-of-Thought Reasoning

A key advantage of LLM-based agents over traditional RL is their ability to perform explicit reasoning through Chain-of-Thought (CoT). To quantify this benefit, we compare SCO-PAL with and without CoT prompting during evaluation. As shown in Table 3, enabling CoT improves the average win rate by 2.4%. This demonstrates that the LLM agent genuinely leverages its language reasoning capabilities for strategic decision-making, rather than merely learning state-to-action mappings like traditional RL approaches. Combined with the generalization results above, these findings highlight unique advantages of LLM-based agents: **interpretable reasoning** through CoT, **zero-shot transfer** to unseen games, and **positive transfer** to related reasoning tasks. We also analyze iterative application of SCO-PAL in Appendix A.3.

7 Ablations

7.1 Ablation of Training Strategies in SCO-PAL

To validate the effectiveness of our two-stage approach, we conduct ablations on both the training procedure and the data used for BC.

Ablation of Training Strategy We compare three training strategies: (1) **Direct KTO**, applying KTO without BC; (2) **Joint Loss**, combining BC and KTO objectives; (3) **Two-Stage (Ours)**, performing BC followed by KTO. As shown in Table 4(a), adding BC improves adaptation to the game environment, with the two-stage strategy achieving a 4.59% performance gain over the joint-loss variant. These results suggest that initializing with BC provides a more stable foundation for subsequent preference optimization.

Ablation of Data for Behavioral Cloning In SCO-PAL, BC is trained on advantageous actions identified from game interaction. We compare two BC data construction methods based on self-play: (1) **Trajectory-Based**, which uses all actions from successful trajectories; (2) **Reward-Based**, which filters actions with estimated rewards exceeding a threshold. As shown in Table 4(b), using only advantageous actions yields better results, as successful trajectories may still contain suboptimal decisions. This highlights the importance of fine-grained filtering in BC data selection.

Modularity of Strategy Optimization A key design principle of SCO-PAL is its **modularity**: the Strategy Refinement stage can flexibly incorporate different preference optimization algorithms. As shown in Figure 4(c), we evaluate four configurations: BC only, BC+DPO, BC+KTO, and BC+GRPO. Adding any preference optimization on top of BC significantly improves performance, confirming the importance of the refinement stage. Among the algorithms, GRPO (Shao et al., 2024) achieves the best performance (54.87%), followed by KTO (50.08%) and DPO (40.52%). KTO outperforms DPO by 9.56 percentage points because it operates at the step level, offering broader coverage of action quality without requiring pairwise comparisons under identical states. While GRPO achieves the highest performance, we use KTO as the default due to its superior efficiency and simplicity. KTO requires significantly fewer computational resources (single trajectory evaluation vs. multiple rollouts for GRPO) and is easier to implement, making it more practical for most scenarios. The 4.79 percentage point gap represents a trade-off between performance and practicality—practitioners with sufficient computational budget may opt for GRPO for maximum perfor-

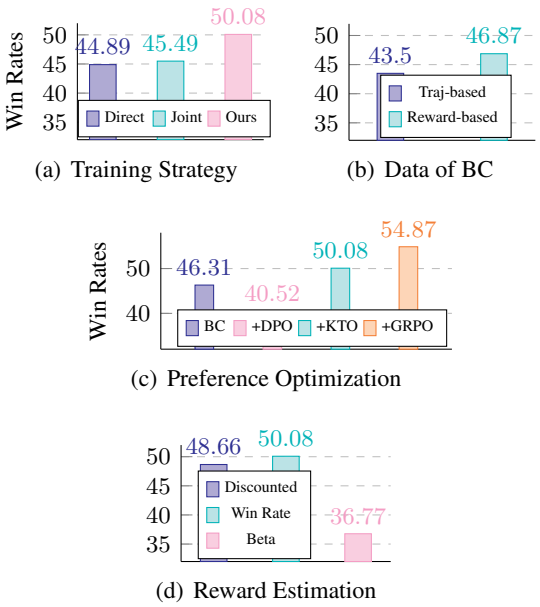


Figure 4: Ablations of SCO-PAL.

mance. This modularity allows SCO-PAL to benefit from advances in preference optimization without architectural changes.

7.2 Ablation of Reward Estimation Methods

Different reward estimation methods produce varying distributions of advantageous and disadvantageous actions, thus influencing SCO-PAL's performance. We compare three approaches—**Discounted Reward**, **Win Rate**, and **Beta**—as detailed in Appendix B.3. As shown in Figure 4(d), the win rate method performs best with a 50.08% win rate, compared to 48.66% for discounted reward and 36.77% for beta. Win rate directly captures success likelihood, aligning well with the game objective and offering a stable, interpretable signal. In contrast, discounted reward requires careful tuning of the discount factor, while the beta method depends on sufficient data and sensitive hyperparameters, reducing practicality. Overall, win rate offers the best trade-off between simplicity, stability, and alignment, making it the most effective choice in SCO-PAL.

Validation of Step-level Rewards To verify that win-rate-based rewards capture true strategic value rather than mere statistical correlation, we validate our step-level reward estimates against Nash equilibrium Q-values for Kuhn Poker, a game with known analytical solutions. As shown in Table 5, the Pearson correlation between empirical win rates and theoretical Q-values is $r = 0.955$ ($p < 0.01$), and Spearman correlation is $\rho = 0.912$ ($p < 0.05$).

State-Action	Nash Q	Win Rate	Samples
J | Bet	-1.00	0.071	11,623
J | Pass	-1.00	0.007	12,421
Q | Bet	-0.50	0.582	11,787
Q | Pass	-0.17	0.458	11,328
K | Bet	+1.17	0.995	11,281
K | Pass	+1.17	0.994	4,763
Pearson $r = 0.955$ ($p < 0.01$)			
Spearman $\rho = 0.912$ ($p < 0.05$)			

Table 5: Validation of step-level rewards against Nash equilibrium Q-values in Kuhn Poker. The high correlation demonstrates that empirical win rates effectively capture true strategic value.

	Nim	Tic-Tac-Toe
Base	0.0564	0.0658
SCO-PAL	**0.0378**	**0.0472**

Table 6: Regret on Nim and Tic-Tac-Toe.

This strong alignment demonstrates that Monte Carlo win rate estimation effectively captures the true strategic value of actions, validating the use of empirical win rates as step-level rewards for policy optimization.

Regret Analysis To further verify that SCO-PAL learns meaningful strategies, we measure regret—the gap between the agent's performance and optimal play—in Nim and Tic-Tac-Toe, two games with known optimal strategies. As shown in Table 6, SCO-PAL consistently reduces regret compared to the base model, indicating that the learned policy moves closer to optimal strategic play. This provides additional evidence that our step-level optimization effectively improves the agent's decision-making quality.

8 Conclusion

In this paper, we focus on improving the strategic reasoning ability of language agents under the play-and-learn paradigm. We design a **S**tep-level poli**C**y **O**ptimization method through **P**lay-**A**nd-**L**earn, SCO-PAL. Using SCO-PAL, we analyze how opponents influence agent learning in adversarial games and find that self-play yields the best results. Incorporating SCO-PAL with self-play, we improve the average win rate from 32.17% (base model) to 50.08%, achieving 54.76% against GPT-4. Our work provides insights into opponent selection in adversarial environments and introduces a new learning framework, offering an effective way to enhance the strategic reasoning of language agents.

Limitations

While SCO-PAL demonstrates strong empirical performance in symbolic adversarial games, it has certain limitations. First, our study is limited to well-defined, turn-based games and does not explore more open-ended or partially observable environments such as negotiation or web-based tasks. Second, the opponent pool is restricted to scripted agents and LLM variants; incorporating human or style-diverse LLM opponents could further enrich strategic diversity but introduces additional challenges. We leave these directions for future work.

Ethical Statement

This work focuses on improving the strategic reasoning abilities of language agents in adversarial games using self-play and interaction-based learning. All experiments are conducted in controlled, simulated environments without real user data or deployment. Our use of existing artifacts is consistent with their intended use, and we follow their license and terms. No sensitive data or human annotations are used. While advances in strategic language modeling may enable more capable agents in competitive settings, we caution against misuse in domains such as manipulation, deception, or adversarial dialogue generation. We encourage future research to incorporate safety constraints and ethical safeguards when deploying such agents in open environments.

References

2024. Qwen2 technical report.

Kai Arulkumaran, Antoine Cully, and Julian Togelius. 2019. Alphastar: an evolutionary computation perspective. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*.

Guillaume Chaslot, Sander Bakkes, Istvan Szita, and Pieter Spronck. 2008. Monte-carlo tree search: A new framework for game ai. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 4, pages 216–217.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. 2024. Self-playing adversarial language game enhances llm reasoning. *arXiv preprint arXiv:2404.10642*.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. *arXiv preprint arXiv:2406.06613*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. 2024. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*.

Iat Long Iong, Xiao Liu, Yuxuan Chen, Hanyu Lai, Shuntian Yao, Pengbo Shen, Hao Yu, Yuxiao Dong, and Jie Tang. 2024. OpenWebAgent: An open toolkit to enable web agents on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 72–81, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, and Xiang Ren. 2024. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*, 36.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

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

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and error: Exploration-based trajectory optimization for llm agents. *arXiv preprint arXiv:2403.02502*.

Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024a. Q*: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*.

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

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024b. Sotopia-π: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, and 1 others. 2024. Agentgym: Evolving large language model-based agents across diverse environments. *arXiv preprint arXiv:2406.04151*.

Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. 2024. Watch every step! llm agent learning via iterative step-level process refinement. *arXiv preprint arXiv:2406.11176*.

Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang, Zhongyu Wei, and Peng Sun. 2025. Multi-agent kto: Reinforcing strategic interactions of large language model in language game. *Preprint*, arXiv:2501.14225.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint arXiv:2310.12823*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, and 1 others. 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

Appendix

Appendix Roadmap.

- **A.1 Imitating Strong Players** studies behavior cloning from expert self-play data (strong-vs-strong), showing that imitation improves win rates but degrades when the skill gap becomes too large.
- **A.2 Combining SCO-PAL and Imitating Strong Players** applies SCO-PAL with self-play on top of the best imitation setting (MCTS(5)) and evaluates against MCTS(1000), demonstrating complementary gains.
- **A.3 Discussion on Iteration** evaluates iterative play-and-learn against historical checkpoints and discusses a slight performance drop in later iterations due to RL overfitting.
- **A.4 Impact of Temperature** analyzes how sampling temperature affects action diversity, game validity, and post-training performance, motivating our choice of temperature.
- **A.5 Impact of Game Balance** evaluates the balancing mechanism that upsamples low-action games and downsamples high-action games to prevent biased learning.
- **A.6 Statistical Significance Tests** reports significance tests versus BC and SPAG, confirming that improvements are statistically significant.
- **A.7 Data Quality Statistics** summarizes action-format validity rates across games and explains how invalid actions are used as negative samples in KTO.
- **A.8 Strategy Analysis: Exploitative vs Nash Equilibrium** compares learned policies with Nash strategies in Kuhn Poker, showing the model learns exploitative behaviors tailored to self-play opponents.
- **A.9 Quantitative Diversity Analysis** reports action diversity and advantageous/disadvantageous ratios under different opponent choices, supporting the opponent-selection analysis.
- **A.10 SCO-PAL on Other Models** evaluates transferability by applying SCO-PAL to Mistral-7B-Instruct-v0.3.
- **A.11 Head-to-Head Evaluation** presents tournament-style direct comparisons across methods on GTBench games.
- **B Method Settings** details experimental configurations, including the MCTS algorithm and hyperparameters, baseline implementations (BC, SPAG), and reward estimation methods (discounted return, win rate, Beta posterior).
- **C Game Introduction** introduces the seen games used for training/evaluation and the unseen games used for generalization tests.
- **D Prompt Design** provides the full system prompt and per-game prompts (game prompt and step prompt) for all environments.

A Supplementary Experiments

A.1 Imitating Strong Players

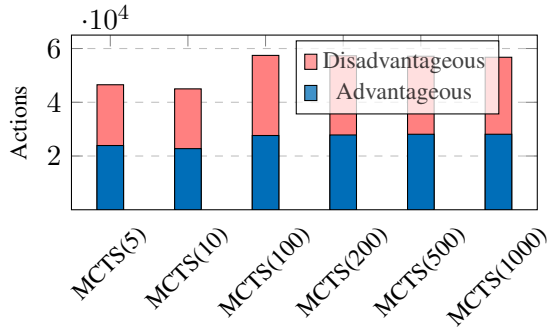
In § 5, we examine how opponents influence performance by introducing stronger ones. Strong opponents tend to adopt better strategies, leading to higher win rates. To further leverage their potential, we explore the effect of imitating strong players.

We first experiment with direct imitation. As shown in § 5, self-play generates diverse, balanced, and high-quality data through interactive gameplay. To better capture strong players' strategies, we adapt SCO-PAL accordingly. Two strong agents of equal strength play against each other, followed by step-wise reward estimation. We then select advantageous samples with win rates above a threshold for BC training. As shown in Figure 5, learning from strong players significantly improves win rates.

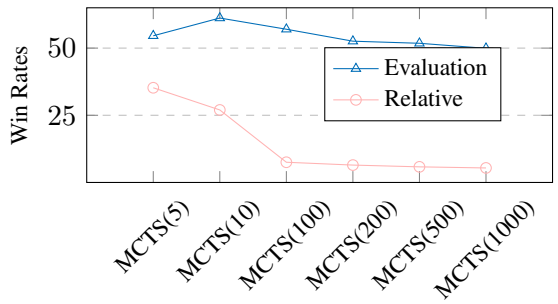
Moreover, we observe that as the opponent's relative strength (compared to the base model) increases, the agent's performance initially improves but later declines. This indicates that while imitating strong players is beneficial, a large skill gap and divergent strategies can impede effective learning.

A.2 Combining SCO-PAL and Imitating Strong Players

In §A.1, we observed that learning from strong players enhances strategic reasoning. Here, we take the best-performing setting—imitating MCTS(5)—and apply SCO-PAL with self-play.



(a) Number of actions.



(b) Win rates: relative and evaluation.

Figure 5: (a) shows the number of actions taken during self-play between strong players of varying levels. In (b), *Relative* denotes the win rate of the base model against different opponents, while *Evaluation* indicates the average win rate of the model trained by imitating strong players, evaluated against four downstream opponents. The number in MCTS(·) specifies the `max_simulation` parameter.

We evaluate the resulting model against a strong opponent, MCTS(1000). As shown in Figure 6, SCO-PAL remains effective even after expert imitation, improving the average win rate against strong opponents by 2.44%. This suggests that imitating strong players can complement SCO-PAL with self-play, leading to further gains.

A.3 Discussion on Iteration

Settings During iteration, the agent plays against historical models in a play-and-learn setup. In the first round, as there is no historical model, the base model (Qwen2-7B-Chat) performs self-play, and we apply SCO-PAL to obtain *Iter1*. In the second round, *Iter1* plays against the base model, is trained to yield *Iter2*, and then *Iter2* plays against *Iter1* to produce *Iter3*.

Results Figure 7 presents the results of iteratively applying SCO-PAL. We observe that iterative self-play with SCO-PAL gradually improves the model's strategic performance. However, in

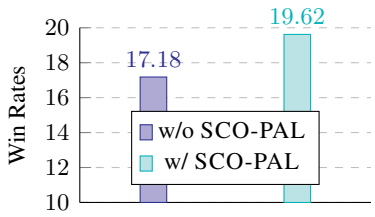


Figure 6: Win rate changes against MCTS(1000) after applying SCO-PAL with self-play to the model trained by imitating MCTS(5).

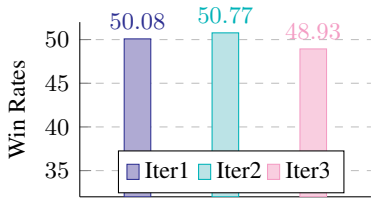


Figure 7: Performance across iterations of SCO-PAL. Each iteration involves self-play and training against previous models.

Iter3, the win rate declines slightly. We attribute this to overfitting introduced by repeated RL training (Zeng et al., 2023; Xiong et al., 2024), which may narrow the model's strategic diversity and reduce effectiveness.

A.4 Impact of Temperature

Temperature is a key parameter in LLM generation, influencing the diversity of outputs. To assess its effect on SCO-PAL, we evaluate five settings: 0.2, 0.5, 0.7, 1.0, and 1.2, and systematically analyze how temperature impacts performance.

Temperature affects action diversity and game validity. Temperature controls the token sampling behavior of language models. Lower temperatures lead to more deterministic outputs, while higher values introduce greater diversity. In our setting, temperature influences the diversity of actions during game interaction by the language agent, which in turn affects the number of actions taken.

As shown in Figure 9(a), action diversity increases with temperature up to a point, peaking between 0.7 and 1.0, then declines. To understand this

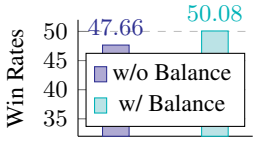
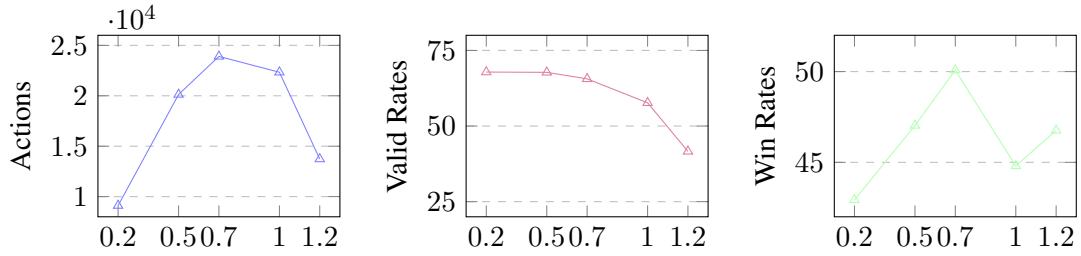


Figure 8: Comparison of win rates with and without the balance mechanism.



(a) Number of actions (b) Game validity rate (c) Post-training win rate

Figure 9: Effect of temperature on (a) number of actions during interaction, (b) validity rate of generated games, and (c) average win rate after training with data from each temperature setting.

	P-Value
BC	0.0017
SPAG	0.0110

Table 7: Significance test for the main experiment.

trend, we examine the rate of valid games—those where the agent's actions are successfully parsed by the environment. At high temperatures, the agent is more likely to generate invalid actions, reducing the number of valid interactions. Figure 9(b) shows that the valid rate peaks between 0.5 and 0.7.

Based on this analysis, we set the temperature to 0.7 during self-play. We also evaluate SCO-PAL under different temperature settings, as shown in Figure 9(c). Results confirm that a temperature of 0.7 yields the highest post-training win rate, indicating that a balance between stability and diversity of valid actions is crucial for optimal performance.

A.5 Impact of Game Balance

Different games have varying action spaces, leading to an unequal number of actions after self-play. To avoid biased learning, we apply data balancing by upsampling actions from games with fewer actions and downsampling those with more, while keeping the total sample size constant. We compare results before and after balancing. As shown in Figure 8, this strategy ensures the language agent learns effectively across different games.

A.6 Statistical Significance Tests

We perform significance tests comparing our method against baselines (BC and SPAG). As shown in Table 7, all p-values are below 0.05, indicating that the improvements are statistically significant.

Game	Total	Valid	Rate
Breakthrough	1,168,008	1,128,700	96.6%
Connect Four	2,559,041	2,558,318	100.0%
Kuhn Poker	1,252,004	1,251,881	100.0%
Liar's Dice	203,873	203,535	99.8%
Nim	1,242,942	1,242,848	100.0%
Tic-Tac-Toe	2,463,757	2,458,078	99.8%
Total	**8,889,625**	**8,843,360**	**99.5%**

Table 8: Data quality statistics across all games. Valid Rate indicates the proportion of correctly formatted actions generated by the base model without SFT warm-up.

Info Set	P(Bet)	Nash	Distance
J (first action)	0.755	0.333	0.421
Q (first action)	0.986	0.000	0.986
K (first action)	0.994	1.000	0.006
Average			**0.471**

Table 9: Comparison of learned policy with Nash equilibrium in Kuhn Poker. P(Bet) denotes the probability of betting at the first action. Distance measures deviation from Nash-optimal play.

A.7 Data Quality Statistics

To address concerns about data quality without SFT warm-up, we analyze the validity rate of collected actions across all games. As shown in Table 8, 99.5% of samples contain valid action formats, demonstrating that the base LLM can effectively generate properly formatted actions without requiring supervised fine-tuning. The small fraction of invalid actions (0.5%) are utilized as negative samples in KTO training, helping the model learn to avoid malformed outputs.

A.8 Strategy Analysis: Exploitative vs Nash Equilibrium

To understand what strategies the model learns through self-play, we analyze the learned policy

Opponent	Samples	Unique	Adv.	Adv/Dis
Self-Play	7.65M	1,721	4.67M	**1.57**
Random	184K	194	85K	0.87
MCTS(5)	629K	360	288K	0.84
MCTS(10)	315K	92	117K	0.59

Table 10: Quantitative analysis of action diversity across different opponents. "Unique" denotes the number of unique actions, "Adv." denotes advantageous actions, and "Adv/Dis" is the ratio of advantageous to disadvantageous actions.

against Nash equilibrium for Kuhn Poker, a game with known analytical solutions. As shown in Table 9, the model correctly learns to bet with King (P=0.994 vs Nash 1.0), but deviates significantly for Queen (P=0.986 vs Nash 0.0) and Jack (P=0.755 vs Nash 0.333). This reveals an important insight: the model learns **exploitative strategies** adapted to the self-play opponent rather than game-theoretically optimal play. Such adaptation explains the strong performance against LLM opponents, as the model exploits predictable patterns in the opponent's behavior.

A.9 Quantitative Diversity Analysis

Table 10 provides quantitative evidence for our opponent selection analysis. Self-play generates the most diverse actions (1,721 unique actions) with the best advantageous/disadvantageous ratio (1.57), indicating balanced learning signals. In contrast, strong opponents (MCTS-10) lead to imbalanced data with ratio 0.59, while weak opponents (Random) produce limited diversity (194 unique actions). This quantitatively validates that self-play provides optimal training signals through both action diversity and sample balance.

A.10 SCO-PAL on Other Models

To test the transferability of SCO-PAL, we apply it to Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). As shown in Table 11, SCO-PAL also improves strategic reasoning in Mistral, demonstrating its versatility across different base models.

	Average Win Rates.
Base	32.18
SCO-PAL	45.80

Table 11: SCO-PAL on Mistral.

	Base	BC	SPAG	SCO-PAL	Avg.
Base	50.00	27.20	24.21	22.97	31.10
BC	72.80	50.00	51.03	45.29	54.78
SPAG	75.79	48.97	50.00	47.31	55.52
SCO-PAL	**77.03**	**54.71**	**52.69**	50.00	**58.61**

Table 12: The vertical axis represents Player 1, the horizontal axis represents Player 2, and the values indicate the win rate of Player 1 against Player 2.

A.11 Head-to-Head Evaluation

We conduct direct battles between methods using six games from `GTBench`, with 100 games played in each match. As shown in Table 12, SCO-PAL achieves a win rate of over 50% compared to all other methods. While BC and SPAG can also outperform the base model, both of them struggle against our SCO-PAL. This indicates that our approach effectively identifies advantageous and disadvantageous actions and uses step-level strategy optimization to achieve consistently high win rates.

B Method Settings

B.1 MCTS Search Algorithm

Monte Carlo Tree Search (MCTS) (Chaslot et al., 2008) is a heuristic search algorithm used for decision-making processes, particularly in game playing. The core principle of MCTS is to build a search tree incrementally and asymmetrically by using random simulations to explore the most promising moves. The algorithm consists of four main steps: selection, expansion, simulation, and backpropagation.

- **Selection:** Starting from the root node, select child nodes based on a policy that balances exploration and exploitation. The most common policy is the Upper Confidence Bound for Trees (UCT), defined as:

$$UCT = \frac{w_i}{n_i} + c \times \sqrt{\frac{\ln N}{n_i}}$$

  where w_i is the number of wins for the node, n_i is the number of times the node has been visited, N is the total number of simulations, and c is the exploration constant.

- **Expansion:** Once a leaf node is reached, expand the tree by adding one or more child nodes corresponding to possible moves. In our setup, the tree expands with each new node representing possible future states.

- **Simulation:** From the new node, perform a rollout, i.e., play out the game randomly until a terminal state is reached. The outcome of this simulation provides an estimate of the node's value.

- **Backpropagation:** Use the result of the simulation to update the information in the nodes along the path from the expanded node back to the root. This involves updating win ratios and visit counts to reflect the newly gathered information.

In our settings, we set `rollout_count`: The number of rollouts performed from each new node is set to 1; `c`: The exploration constant of UCT is set to 2, balancing the trade-off between exploration of new nodes and exploitation of known promising nodes; `max_simulations`: The maximum number of simulations. This parameter controls the total number of simulations that can be performed during decision-making processes. By adjusting `max_simulations`, we can control the level of agents, where a higher number of simulations generally leads to more informed and strategic decisions, enhancing the agent's competitive performance.

B.2 Baselines

BC Inspired by (Wang et al., 2024b), BC involves using the action sequences from winning players as training data to improve the agent's strategy. This method can be broken down into the following steps:

First, we collect a dataset $\mathcal{D}$ comprising trajectories $\mathcal{T}$ from games where the player wins. Each trajectory τ is represented as a sequence of state-action pairs:

$$(s_0, a_1, s_1, a_2, s_2, \ldots, a_n, s_n, r)$$

where r is 1, representing that the agent wins.

The goal is to train the policy π_θ to replicate these successful behaviors. The learning objective is to minimize the Behavioral Cloning loss, which is defined as:

$$L_{BC} = -\mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi_\theta(a|s)]$$

Where π_θ is the policy model that we aim to update.

SPAG (Cheng et al., 2024) The given formula represents the loss function $\mathcal{L}_{\text{SPAG}}(\pi_\theta)$, which is used to optimize the policy π_θ in a strategic game setting. The loss function is composed of several components:

$$\mathcal{L}_{\text{SPAG}}(\pi_\theta) =$$

$$-\frac{1}{2}\mathbb{E}_{\mathcal{T}^1_{\bar{\theta}}}\left[\sum_{t=1}^{T}\frac{\pi_\theta(\mathbf{u}_t|f_1(\mathcal{S}_{t-1}))}{\pi_{\bar{\theta}}(\mathbf{u}_t|f_1(\mathcal{S}_{t-1}))}\hat{A}_t^{\mu_{\bar{\theta}}} - \beta_2 KL[\pi_\theta\|\pi_{\bar{\theta}}]\right]$$

$$-\frac{1}{2}\mathbb{E}_{\mathcal{T}^2_{\bar{\theta}}}\left[\sum_{t=1}^{T}\frac{\pi_\theta(\mathbf{v}_t|f_2(\mathcal{S}'_t))}{\pi_{\bar{\theta}}(\mathbf{v}_t|f_2(\mathcal{S}'_t))}\hat{A}_t^{\nu_{\bar{\theta}}} - \beta_2 KL[\pi_\theta\|\pi_{\bar{\theta}}]\right]$$

where:$\mathbb{E}_{\mathcal{T}^1_{\bar{\theta}}}$ and $\mathbb{E}_{\mathcal{T}^2_{\bar{\theta}}}$ denote the expectations over trajectories for the player1 and player2, respectively. π_θ is the policy being optimized, and $\pi_{\bar{\theta}}$ is a reference policy. $\mathbf{u}_t$ and $\mathbf{v}_t$ are actions taken by player1 and player2 at time t. f_1 and f_2 are the prompts for player1 and player2. $\hat{A}_t^{\mu_{\bar{\theta}}}$ and $\hat{A}_t^{\nu_{\bar{\theta}}}$ are approximations of the advantage functions for player1 and player2. β_2 is a regularization parameter for the Kullback-Leibler divergence $KL[\pi_\theta\|\pi_{\bar{\theta}}]$. In SPAG, they define $\hat{A}_t^{\mu_{\bar{\theta}}} \approx r(s_{t-1}, u_t)$ and $\hat{A}_t^{\nu_{\bar{\theta}}} \approx -r(s'_t, v_t)$.

Given a complete trajectory, the rewards are assigned for each action:

$$r(s_{t-1}, u_t) = \begin{cases} (1-\gamma)\gamma^{T-t}/(1-\gamma^{T+1}), & \text{if player1 wins,} \\ -(1-\gamma)\gamma^{T-t}/(1-\gamma^{T+1}), & \text{if player2 wins,} \\ 0, & \text{if game is tied.} \end{cases}$$

$$r(s'_t, v_t) = \begin{cases} -r(s_{t-1}, u_t), & \text{if attacker wins,} \\ -r(s_{t-1}, u_t), & \text{if defender wins,} \\ 0, & \text{if game is tied.} \end{cases}$$

In our experiments, we set γ to 0.8, the same as the setting in SPAG.

B.3 Reward Estimation Methods

Discounted Reward Discounted reward is used to evaluate the long-term value of actions. The discounted reward R_t from time step t is the sum of all future rewards, each multiplied by a discount factor γ:

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^{T-t} r_T$$

If rewards are given only at the end of the game, then $r_k = 0$ for $k < T$, with a final reward r_T based on the outcome. For a state-action pair (s, a) occurring at time t, the discounted reward is:

$$R(s, a) = \gamma^{(T-t)} \times R_T$$

Calculation Formula:

$$Q(s, a) = \frac{1}{N}\sum_{i=1}^{N} R(s, a)_i$$

N is the total number of occurrences of (s, a). $R(s, a)_i$ is the discounted reward for the i-th occurrence. In our experiments, we set γ to 0.8.

Win Rate Win rate can be used as a simple metric to evaluate the effectiveness of actions. The win rate for a particular state-action pair (s, a) is calculated as the ratio of the number of times the action leads to a win to the total number of times the action is taken. This provides a straightforward measure of how successful an action is in achieving the desired outcome. The winning rate $W(s, a)$ is given by:

$$W(s, a) = \frac{N_{win} + 0.5N_{tie}}{N_{win} + N_{lose} + N_{tie}},$$

N_{win}, N_{tie}, N_{total} are the total number of winning times, tying times, and total times, respectively, of action (s, a).

Beta Using the Beta distribution to estimate rewards is a Bayesian approach that models the uncertainty in the winning rate of state-action pairs (s, a). This method is especially useful in scenarios with small sample sizes or sparse data. The Beta distribution is parameterized by α and β, representing the counts of wins and losses, respectively. The winning rate is estimated using the posterior mean:

$$\hat{p} = \frac{\alpha}{\alpha + \beta} = \frac{\alpha_0 + w}{\alpha_0 + \beta_0 + w + l}$$

where α_0 and β_0 are prior parameters, w is the number of wins, and l is the number of losses. This Bayesian estimate provides a probabilistic view of the winning rate, incorporating prior knowledge and observed data. In our experiments, we set both α_0 and β_0 to 1.

C Game Introduction

C.1 Games

Tic-Tac-Toe Tic-Tac-Toe is a classic two-player game played on a 3x3 grid. Players take turns marking empty squares with their respective symbols—one player uses "X" and the other uses "O". The objective is to be the first to align three of one's symbols horizontally, vertically, or diagonally.

Breakthrough Breakthrough is a modern two-player abstract strategy game. Played on a 6x6, 7x7, or 8x8 grid, each player controls a set of pieces placed on their side of the board. The objective is to move one of your pieces to the opponent's back row. Pieces move similarly to pawns in chess, advancing forward either straight or diagonally when capturing an opponent's piece.

Connect Four Connect Four is a two-player connection game where players choose a color and then take turns dropping colored discs into a vertical grid consisting of seven columns and six rows. The discs fall straight down, occupying the lowest available space within the column. The goal is to be the first to form a continuous line of four of one's own discs vertically, horizontally, or diagonally.

Kuhn Poker Kuhn Poker is a simplified form of poker. Kuhn to illustrate the principles of game theory. The game involves two players and a deck of just three cards: Jack, Queen, and King. Each player antes one chip, receives one card, and there's a single round of betting with limited actions—players can either bet or pass. Despite its simplicity, Kuhn Poker encapsulates fundamental concepts like bluffing, information asymmetry, and mixed-strategy equilibria, making it a valuable educational tool in strategic thinking and economics.

Nim Nim is a mathematical strategy game for two players. The game starts with several piles or heaps of objects, and players take turns removing any number of objects from a single pile. The player who removes the last object wins (or loses, depending on the variation). Nim is notable for its mathematical underpinnings and is often studied in combinatorial game theory. The game has a direct correlation with binary numbers, and mastering Nim involves calculating the "number" or "number sum" to determine winning moves.

Liar's Dice Liar's Dice is a bluffing game for two or more players. Each player starts with a set of dice and a cup to conceal their roll. Players take turns bidding on the total number of a certain face value of dice (e.g., "three sixes") among all players' dice. The next player can either raise the bid or call "liar" to challenge the previous bid. When "liar" is called, all dice are revealed. If the bid is accurate or underbid, the challenger loses a die; if not, the bidder loses a die. The game continues until only one player remains with dice.

C.2 Unseen Games

Prisoner's Dilemma The Prisoner's Dilemma is a thought experiment in game theory involving two rational individuals. Each participant has two choices: cooperate with their partner for mutual benefit or betray them (defect) to seek an individual reward. The dilemma arises because while mutual cooperation leads to a better collective outcome, each participant has a personal incentive to defect, which can lead to a worse overall result if both decide to do so.

Blind Auction A Blind Auction is a commonly used auction type in which all participants submit their bids simultaneously in sealed envelopes. This means no bidder knows the bids of other participants. The highest bidder wins the auction and pays the price they submitted. The key aspect of this auction is that all actions are taken simultaneously, without knowledge of the others' bids.

Pig Pig is a straightforward dice game. Players take turns rolling a single die as many times as they desire during their turn. Each roll's result is added to their running total for that turn. However, if a player rolls a 1, they lose all the points accumulated during that turn, and their turn ends. The objective is to strategically decide when to stop rolling to avoid losing points while accumulating a high score.

D Prompt Design

D.1 System Prompt

```
You are a powerful gaming agent who can make proper decisions to beat the user in
gaming tasks. You are a helpful assistant that strictly follows the user\'s
instructions.
```

D.2 Game Prompt

D.2.1 Tic-Tac-Toe

- Game Prompt

```
Tic Tac Toe is a two-player game played on a grid. Players take turns marking
a space with their respective symbols. The goal is to get 3 of one\'s own
symbols in a row, either horizontally, vertically, or diagonally, before the
opponent does. If all nine squares are filled and no player has three in a row
, the game is a draw. The Tic Tac Toe game is played on a 3 by 3 grid, with
the winning length as 3. Each move is represented by a string consisting of
two parts: the column (C) and the row (R), in that order. For instance, C1R2
means the movement at the position of the first column and the second row of
the grid. You are playing this game with the user (opponent).
```

- Step Prompt

```
Your opponent has finished actions: <OPPONENT_MOVES>. You have finished
actions: <SELF_MOVES>.
```

D.2.2 Breakthrough

- Game Prompt

```
Breakthrough is a two-player game played on a rectangular board. Players take
turns moving their pieces, which can move one space straight or diagonally
forward if the target square is empty. A piece can also move diagonally
forward to capture an opponent's piece. Capturing is optional, and a player
can only capture one piece per turn. The goal is to be the first to reach the
opponent's home row, the farthest row from the player. If all of a player's
pieces are captured, they lose. The game does not allow draws, as pieces can
only move forward or be captured.The Breakthrough board is identified by
columns labeled start from a to c (from left to right) and rows numbered 1 to
8 (from bottom to top). The intersection of a column and a row specifies a
unique square on the board.
```

- Step Prompt

```
The board now looks like: <BOARD_PREVIEW>. Among which, the letter 'b'
represents a black piece, while the letter 'w' represents a white piece. And
the character "." represents vacant space. The numbers in the board are the
indexes of the rows. Your opponent has finished actions: <OPPONENT_MOVES>.You
have finished actions: <SELF_MOVES>.
```

D.2.3 Connect Four

- Game Prompt

```
Connect 4 is a two-player connection board game, where the players choose a
color and then take turns dropping colored discs into a vertically suspended
grid. The pieces fall straight down, occupying the next available space within
 the column. The objective of the game is to be the first to form a horizontal
, vertical, or diagonal line of four of one's own discs. You are a gaming
agent that aims to beat me in Connect 4 games. Each move is represented by a
string consisting of two parts: the column (C) and the row (R), in that order.
 For instance, C1 means the first column.
```

- Step Prompt

```
Your opponent has finished actions: <OPPONENT_MOVES>. You have finished
actions: <SELF_MOVES>.
```

D.2.4 Kuhn Poker

- Game Prompt

```
Kuhn poker is a simple model zero-sum two-player imperfect-information game,
amenable to a complete game-theoretic analysis. In Kuhn poker, the deck
includes only three playing cards: a King (K), a Queen (Q), and a Jack (J).
One card is dealt to each player, and the third is put aside unseen. The
players take turns either <Bet> to match the bet raised by the opponent or <
Pass> to conceds the game. If a player bets, the other player must either call
 the bet by matching it or fold by conceding the game. If both players pass,
the game is over, and the player with the higher-ranking card wins. The card
rankings are as follows: King (K) > Queen (Q) > Jack (J). You are playing Kuhn
 poker with the opponent. The actions are denoted by <Bet> and <Pass>.
```

- Step Prompt

```
In this match, your card is <CARD>. Here are the past moves in this match: <
SELF_MOVES>, <OPPONENT_MOVES>.
```

D.2.5 Nim

- Game Prompt

```
In Nim, a strategic game with a set of four piles containing 1, 3, 5, and 7
matches respectively, players aim to avoid taking the last match. During each
turn, a player may take any number of matches from a single pile, but must
take at least one and cannot exceed the number remaining in that pile. The
objective is to force the opponent to pick up the final match, thereby winning
 the game. The action is presented in <pile:x, take:y>, which means take y
match(es) from the x-th pile.
```

- Step Prompt

```
Currently, the 1st pile has <PILES[0]> match(es), the 2nd pile has <PILES[1]>
match(es), the 3rd pile has <PILES[2]> match(es), 4th pile has <PILES[3]>
match(es).
```

D.2.6 Liar's Dice

- Game Prompt

```
Liar's Dice is a game of bluffing and probability, played with two players and
 each player has 1 dice. During each turn, a player can either bid a higher
quantity of any particular face value or the same quantity of a higher face
value than the previous bid. Each player tries to outbid their opponent
without being caught in a lie. The move in this game is denoted in <x dices, y
 value>, meaning there are at least x dices with face values as y.
```

- Step Prompt

```
Currently, the face value of your dice is <FACE_VALUE>. Last time, the
opponent called <OPPONENT_LAST_ACTION>. You are playing the Liar's Dice with
another opponent. Therefore, there are only two dice in total.
```

D.2.7 Prisoner's Dilemma

- Game Prompt

```
You and your partner are in the Prisoner's Dilemma situation. Specifically, if
 you <Testify> against your partner and your partner remains <Silent>, you
will go free while your partner will get 3 years in prison on the main charge.
 If you remain <Silent> but your partner <Testify> against you, you will serve
 3 years in prison and your partner will be set free. If you and your partner
<Testify> against each other, you and your partner will each serve 2 years. If
 both you and your partner remain <Silent>, you and your partner will each
serve 1 year.
```

- Step Prompt

```
You have been through this situation in the past and here are the decisions
you and your partner made: (In the idx + 1 th round, you decided to <MOVE> and
 your opponent decided to <OPPONENT_MOVE>) * n round.
```

D.2.8 Blind Auction

- Game Prompt

```
A first-price sealed-bid auction (FPSBA) is a common type of auction. It is
also known as the blind auction. In this type of auction, all bidders
simultaneously submit sealed bids so that no bidder knows the bid of any other
 participant. The highest bidder pays the price that was submitted. Each
action is represented by <x> where x refers to the bid.
```

- Step Prompt

```
Now, you are in an auction with an opponent. You want to win the object and at
 the same time, your budget is <VALUATION>. Your bid must be strictly lower
than or equal to <VALUATION>. You shall bid wisely against your opponent. Your
 opponent also has an expected valuation and you do not know it.
```

D.2.9 Pig

- Game Prompt

```
Pig is a fast-paced dice game where players risk accumulating points with each
 roll but risk losing them all if they roll a 1. Each player must decide when
to stop rolling and bank their points, aiming to be the first to reach 100
points. You are playing Pig with the other.
```

- Step Prompt

```
Right now, your current score is <AGENT_CURRENT_SCORE> and your opponent's
current score is <OPPONENT_CURRENT_SCORE>. In this turn, you have earned <
TURN_TOTAL_SCORE> score.
```