# Who is a Better Player: LLM against LLM

**Anonymous Author(s)**
Affiliation
Address
email

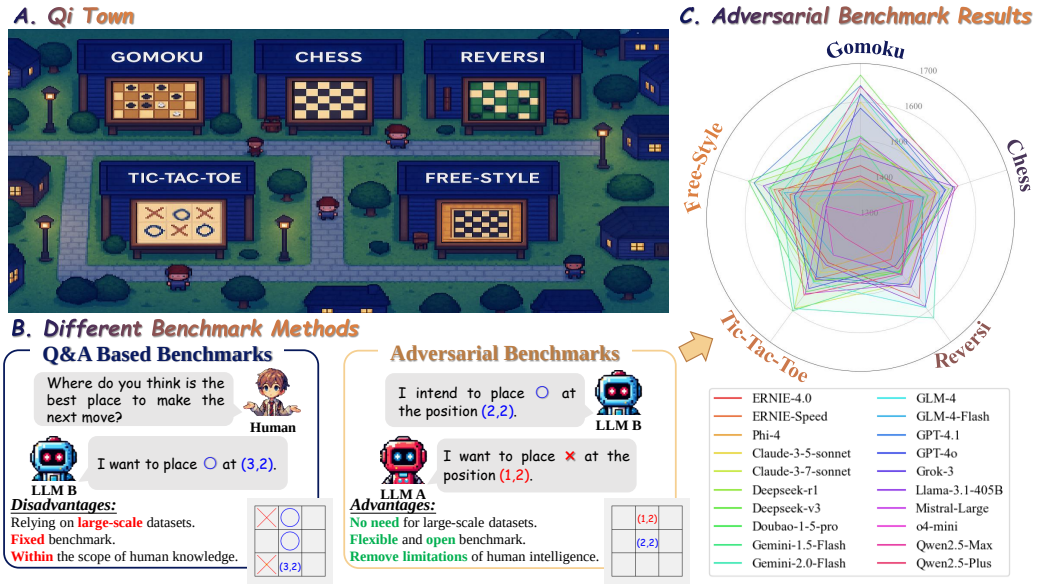



Figure 1: The implementation, advantages and results of our adversarial benchmark. Subfigure A shows the Qi Town proposed in this paper, which provides a basic platform for adversarial benchmarks; Subfigure B compares the differences between Q&A-based benchmark schemes and our adversarial benchmark; Subfigure C shows the performance measurement results of different LLMs.

## Abstract

Adversarial board games, as a paradigmatic domain of strategic reasoning and intelligence, have long served as both a popular competitive activity and a benchmark for evaluating artificial intelligence (AI) systems. Building on this foundation, we propose an adversarial benchmarking framework to assess the comprehensive performance of Large Language Models (LLMs) through board games competition, compensating the limitation of data dependency of the mainstream Question-and-Answer (Q&A) based benchmark method. We introduce **Qi Town**, a specialized evaluation platform that supports 5 widely played games and involves 20 LLM-driven players. The platform employs both the Elo rating system and a novel Performance Loop Graph (PLG) to quantitatively evaluate the technical capabilities of LLMs, while also capturing Positive Sentiment Score (PSS) throughout gameplay to assess mental fitness. The evaluation is structured as a round-robin tournament, enabling systematic comparison across players. Experimental results indicate that, despite technical differences, most LLMs remain optimistic about winning and losing, demonstrating greater adaptability to high-stress adversarial environments than humans. On the other hand, the complex relationship between cyclic wins and losses in PLGs exposes the instability of LLMs' skill play during games, warranting further explanation and exploration.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

# 1 Introduction

Board games have long been regarded as intellectually demanding competitions that assess a broad spectrum of cognitive and affective skills, including memory, logical reasoning, calculation, strategic planning, risk anticipation, and emotional regulation. Their enduring popularity worldwide stems not only from their competitive nature but also from their capacity to holistically evaluate human intelligence. With the rapid advancement of artificial intelligence (AI), these traditionally human-versus-human games have increasingly become arenas for human-computer interaction.

Recently, the emergence of large language models (LLMs) has positioned them as novel participants in such games, opening up new opportunities for comprehensive LLM evaluation. Unlike conventional assessment methods that typically isolate specific cognitive capabilities, strategic board games offer a richer, multifaceted environment to evaluate LLMs' integrated competencies. These include game memory, move reasoning and analysis, opponent behavior prediction, and emotional stability in dynamic, adversarial settings. To harness these properties, we propose an adversarial evaluation paradigm—LLM against LLM—as a new framework for LLM benchmarking. As shown in Fig. 1, in contrast to standard question-and-answer (Q&A)-based benchmarks that rely on static, human-curated corpora, adversarial benchmark enables continuous, self-generated assessment content through real-time interactions between models. Overall, the proposed paradigm offers several notable advantages: 1) **Dynamic benchmarking** is achieved through diverse and evolving game states; 2) **Environment awareness** is considered through the inclusion of high-stress adversarial conditions that challenge LLM adaptability; 3) **Safety and reliability** are enhanced by a dynamic evaluation mechanism.

In this work, we construct Qi Town as shown in Fig. 1, undertaking a comprehensive investigation of LLM performance in adversarial board game settings. Specifically, we design experiments across five games: Gomoku, Chess, Reversi, Tic-Tac-Toe—representing traditional rule-based games—and a novel Free-Style game mode. In the Free-Style setup, LLMs collaboratively determine the rules before initiating play, thus testing their autonomy and negotiation capabilities. Across all games, we employ the Elo rating system to assess competitive skill, and we introduce a Performance Loop Graph (PLG) to visualize LLM behaviors across multiple matches. To evaluate emotional aspects, we record sentiment changes during gameplay and compute a Positive Sentiment Score (PSS) to capture LLMs' affective responses. The principal contributions of this work are summarized as follows:

- **Adversarial Benchmarking Platform:** We present Qi Town, a novel evaluation framework supporting 20 LLMs engaged in board game competition. This platform introduces an adversarial benchmarking paradigm that surpasses the limitations of static Q&A-based evaluations and reduces reliance on human-generated data.
- **Diverse Game Set and Multidimensional Metrics:** The benchmark includes both fixed-rule games and a Free-Style mode that allows models to co-construct game rules. We evaluate performance across both technical (Elo ratings, PLG) and psychological (PSS) dimensions, providing a multidimensional view of LLM capabilities.
- **Comprehensive Experimental Analysis:** The experimental results reveal limitations in some LLMs' reasoning, decision-making capabilities, and possible risk of affective tendencies toward negative emotion. However, most LLMs can adapt to high-stress competition with optimism. Besides, the cyclic win–loss patterns observed in PLGs offer novel insights into the structural stability and relative performance dynamics of LLMs in competitions.

# 2 Related Work

## 2.1 Q&A-based Benchmarks for LLM

Benchmarks play a critical role in both validating and advancing the capabilities of LLMs. Currently, the predominant evaluation paradigm is Q&A-based, which involves constructing large-scale evaluation corpora and assessing LLMs' performance based on their responses. Typically, benchmarks like C-Eval [1], MMLU [2], and Big-Bench [3] are designed to assess knowledge comprehension and memorization, while others such as TNEWS [4], IFLYTEK [4], ChID [4], and CMRC2018 [5] focus on evaluating natural language processing skills. This Q&A-based approach has also been extended to multimodal LLMs (MLLMs), where benchmarks such as MME and MMBench assess perceptual and reasoning abilities, MER-Bench [6] and MEMO-Bench [7] emphasize affective comprehension,

Q-Bench [8], Q-Bench-Video [9], and A-Bench [10] evaluate visual quality understanding, and VSI-Bench [11] targets visuospatial perception.

Despite the diversity and depth of these benchmarks, Q&A-based evaluation schemes face several inherent limitations. Constructing large-scale, high-quality benchmarks is resource-intensive, requiring substantial time, cost, and manual annotation. Furthermore, because ground truth labels are constrained by human cognition, the scope of such evaluations is fundamentally limited to human-level intelligence. As LLMs continue to evolve and begin to exhibit behaviors and competencies that potentially exceed those of human evaluators, these constraints are becoming increasingly apparent. This highlights the urgent need to explore novel benchmark paradigms that move beyond traditional Q&A frameworks to more fully capture the emerging capabilities of advanced LLMs.

### 2.2 AI in Games

Over the past decade, the advent of reinforcement learning (RL) has catalyzed significant breakthroughs in the application of AI to gaming. In competitive game environments, RL-based agents such as AlphaStar [12] and OpenAI Five [13] have demonstrated superhuman performance in complex real-time strategy games like StarCraft II and Dota 2, respectively, by leveraging layered model architectures and self-play training paradigms. In the domain of board games, AlphaGo [14] marked a transformative milestone by integrating Monte Carlo Tree Search (MCTS) with dual deep neural networks to defeat the world champion in Go in 2016. Its successor, AlphaGo Zero [15], further advanced this approach by relying entirely on self-play without human supervision.

In recent years, LLMs have emerged as promising agents in various game-related tasks. For example, AI Town, proposed by Park $et\ al.$ [16], introduced a virtual community of 25 LLM-based agents, showcasing initial explorations of LLMs in interactive, multi-agent environments. Wang $et\ al.$ [17] examined the gaming capabilities of MLLMs using puzzles such as Sudoku and Minesweeper, as well as tasks involving search algorithms, though their work did not fully exploit the competitive and strategic nature of multi-agent gameplay. Huang $et\ al.$ [18] investigated the decision-making capabilities of LLMs through game-theoretic scenarios, such as the El Farol bar problem [19], successfully integrating game theory into LLM evaluation. However, their analysis was limited to static, single-play games. To address these limitations, we propose a novel evaluation paradigm based on dynamic, strategy-rich board games. Specifically, we introduce a competitive setting wherein LLMs directly confront one another in a variety of board games characterized by explicit rules and complex decision spaces. This LLM against LLM framework is designed to comprehensively assess both the technical and psychological capabilities of state-of-the-art LLMs in adversarial scenarios.

## 3 Qi Town

### 3.1 Overview

Qi Town, as illustrated in Fig. 2, is a virtual community specifically designed to facilitate competitive board games among LLMs. From a scheduling perspective, Qi Town is equipped to manage round-robin tournaments, enabling flexible organization of adversarial matchups. The core functionality of Qi Town lies in its game process control module, which orchestrates matches between LLMs in accordance with predefined game rules. During gameplay, the system continuously updates the game status and alternately prompts each LLM with the current game context. Based on these prompts, each player generates a response comprising three elements: the movement, emotion and a brief analysis. The move updates the game status and is logged, while the emotional and explanatory components are retained for post-game analysis. Besides, Qi Town provides automated visualization and storage of game sequences. This process iterates until a win or draw is reached, after which the system proceeds to the next scheduled match. For evaluation, Qi Town continuously tracks and updates each LLM's Elo score and associated performance metrics, ensuring a dynamic and quantitative assessment of both technical proficiency and emotional changes across games.

### 3.2 Game Types

**Fixed-Rule Games** are a class of deterministic, turn-based environments governed by fixed rules and clear win conditions, offering ideal testbeds for evaluating strategic reasoning and decision-making. This study considers four representative fixed-rule games of increasing complexity. Among all,

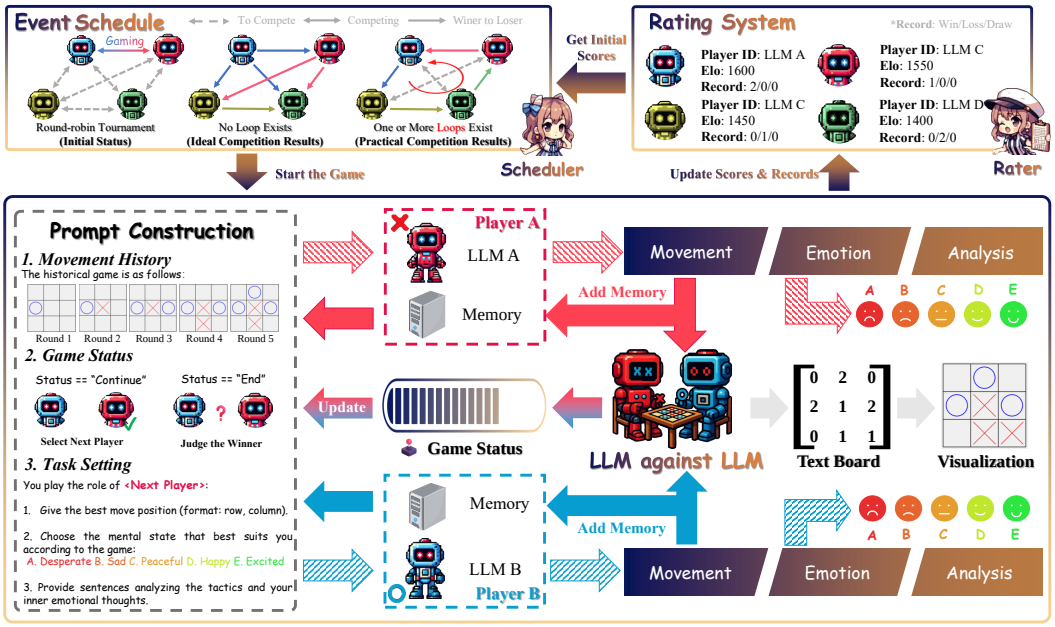



Figure 2: The framework of Qi Town, which consists of three parts: tournament scheduling, LLM against LLM and rating system.

Tic-Tac-Toe is a simple two-player game played on a 3×3 grid, where players alternate marking empty cells with "X" or "O" to align three symbols in a row. Gomoku expands this structure to a 15×15 board, requiring players to align exactly five consecutive stones of the same color. Reversi, conducted on an 8×8 grid, involves flipping the opponent's discs by flanking them, aiming to control the majority of the board by the end of the game. Chess is a globally standardized strategy game played on an 8×8 board with 16 uniquely functioning pieces per player, requiring both deep tactical calculation and long-term planning to achieve checkmate under complex movement and game-state rules. These games span a spectrum of spatial, combinatorial, and cognitive demands, making them effective platforms for analyzing algorithmic strategies and LLMs' behavior in rule-bound settings.

**Free-Style** is an open-ended board game paradigm designed to assess negotiation and rule-making abilities in LLM against LLM settings. Played on a 5×5 board, the game does not begin with predefined rules. Instead, the two participating LLMs engage in an iterative negotiation process to co-construct a mutually agreed-upon rule set. Once finalized, gameplay proceeds according to these self-defined rules until a winner is determined. Free-Style removes human-imposed constraints from the evaluation process, offering a unique opportunity to study the rule formation, adaptation, and interactive negotiation capabilities of LLMs in a controlled adversarial context.

### 3.3 Players: Large Language Models

Distinct from prior studies on competitive board games, which typically involve human players or search-based algorithms as opponents, this work introduces a novel evaluation paradigm by setting both competing entities as LLMs. To ensure the diversity and representativeness of the participants, we included LLMs developed by twelve different organizations, selecting the recently released versions from each to maintain experimental timeliness. In total, 20 LLMs are chosen for participation in the chess tournament, including: Claude-3-7-sonnet [20], Claude-3-5-sonnet [21], Deepseek-v3 [22], Deepseek-r1 [23], Doubao-1.5-pro [24], ERNIE-Speed [25], ERNIE-4.0 [26], Llama-3.1-405 [27], Mistral-Large [28], Phi-4 [29], Grok-3 [30], Gemini-1.5-Flash [31], Gemini-2.0-Flash [32], Qwen2.5-Plus [33], Qwen2.5-Max [33], GPT-4.1 [34], GPT-4o [35], and o4-mini [36]. All LLMs are accessed via their official Application Programming Interfaces (APIs) to ensure consistency in interaction and scheduling. As illustrated in Fig. 2, the gameplay interface is purely textual, with no visual processing involved; instead, board status is represented and interpreted using a standardized algebraic notation system, through which all move decisions are made.

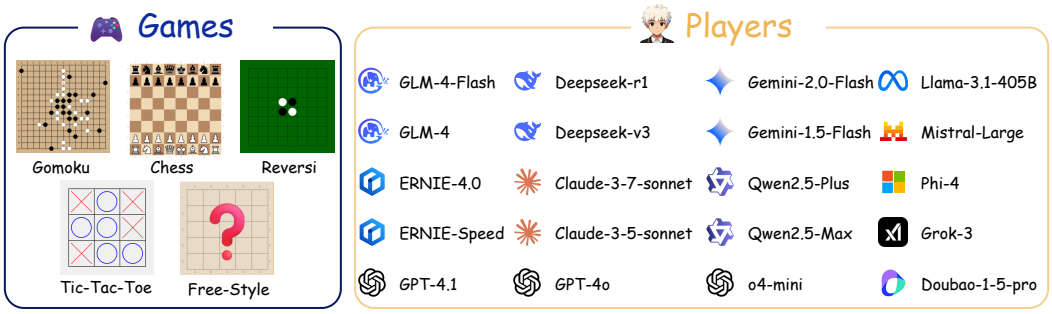



Figure 3: Visualization of various board games and a list of players.

### 3.4 Game Visualization

While coordinate and algebraic notations can fully represent the progression of a game, they are not well-suited for further analysis due to their lack of visual intuitiveness. To facilitate a clearer understanding of the moves made by both sides' LLMs, we implement a visual representation of each move using Pygame [37] and Python-chess [38], as illustrated in Fig. 3. Additional visualizations, including examples from newly introduced Free-Style game types, are presented in Sec. A.

### 3.5 Elo Rating System

The Elo rating system [39], is a widely adopted statistical method for evaluating player performance across various competitive domains. Originally designed for chess, it has since been applied to other disciplines such as Chinese Chess, Go, football, basketball, and eSports, and is now regarded as a standard for quantifying relative skill levels. In this study, we employ the Elo scoring system to evaluate the overall performance of 20 players across a series of chess matchmaking tournaments. The core assumption of the Elo model is that a player's skill level follows a normal distribution centered around their mean performance. The probability density functions for two competing players $X$ and $Y$ with average performance scores $U_x$ and $U_y$ can be described as:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\delta} \exp\left(\frac{-(x-U_x)^2}{2\delta^2}\right), \\ f(y) &= \frac{1}{\sqrt{2\pi}\delta} \exp\left(\frac{-(y-U_y)^2}{2\delta^2}\right), \end{aligned} \tag{1}$$

where $\delta$ denotes the standard deviation, reflecting the same stability of each player's performance. Based on this assumption, the expected probability of player $X$ defeating player $Y$, given their performance scores are random variables $x$ and $y$ respectively, is computed as:

$$P(D) = P(x-y) = \frac{1}{2} + \int_0^D \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx, \sigma = \sqrt{2}\delta, \tag{2}$$

where $D$ denotes the difference in performance scores between the two competitors. Since the integral form in Eq. (2) is computationally intensive, an approximation is commonly used in practice, derived through curve fitting using the least squares method. This yields the following sigmoid-like function:

$$P(D) = \frac{1}{1+10^{-\frac{D}{400}}}. \tag{3}$$

It is worth stating that the constants in Eq. (3) are empirically derived to closely match the theoretical distribution in Eq. (2), providing a computationally efficient alternative for estimating the expected win probability $P(D)$. Finally, given the Elo scores of players $X$ and $Y$ before $n$-th ($n \in N^+$) match, denoted as $ES_X(n-1)$ and $ES_Y(n-1)$, their updated scores after the match are calculated as:

$$\begin{aligned} ES_X(n) &= ES_X(n-1) + K(W_X - P(D)), \\ ES_Y(n) &= ES_Y(n-1) + K(W_Y - P(D)), \end{aligned} \tag{4}$$

where $K$ is a scaling factor that controls the sensitivity of the rating updates, and $W_X$, $W_Y$ represent the actual match outcomes. Specifically, a win yields $W = 1$, a loss $W = 0$ and a draw $W = 0.5$ for both players. This update process is repeated for each match in the tournament series, ultimately yielding the final Elo scores for all 20 players.

### 3.6 Performance Loop Graph

While the Elo rating system provides an aggregate measure of each player's overall performance, it does not capture the relational dynamics between individual competitors. To address this limitation, we introduce a Performance Loop Graph (PLG), a directed graph-based visualization that encodes the win–loss relationships among players. Formally, the PLG is defined as:

$$G = \{(V, E)|v \in V, e \in E\}, \tag{5}$$

where each player is represented as a vertex $v$, and every directed edge $e$ represents a match outcome. Specifically, an edge is directed from the winner to the loser, indicating the result of a head-to-head match. In cases of a draw, no edge is added between the corresponding nodes. The set $V$ thus comprises all participating players, and $E$ contains all directed edges formed based on match outcomes. This graph structure allows for the application of graph-theoretic metrics to gain deeper insights into player performance. For each vertex $v$, the out-degree $d^+(v)$ corresponds to the number of wins, while the in-degree $d^-(v)$ denotes the number of losses. These metrics provide a local view of each player's performance in relation to others. Moreover, the presence of loops within the directed graph $G$ signifies mutual victories among groups of players, reflecting competitive balance and performance volatility. Of particular interest is the largest loop in $G$, which includes the maximum number of players. This loop is indicative of the most competitive and unstable subset of the player pool, offering valuable insight into the intensity and unpredictability of match outcomes.

### 3.7 Positive Sentiment Score

In contrast to emotions elicited through traditional human-centered methods, the spontaneous emotional responses exhibited by LLMs during task execution may offer deeper insights into their intrinsic behavioral traits, particularly their capacity for resilience under adversity. This characteristic is crucial for ensuring that LLMs maintain a stable and constructive affective state, thereby supporting more harmonious human–machine interactions.

To capture these emotional dynamics, we implemented a self-reporting protocol wherein players are asked to select their emotional state after each round of gameplay from the following options: *A. Desperate, B. Sad, C. Peaceful, D. Happy, and E. Excited*. To quantitatively assess the overall emotional positivity of the players, we introduce the Positive Sentiment Score (PSS), a metric based on the expected value of a discrete random variable $Z$, which encodes emotional states. The PSS is computed as follows:

$$E(Z) = \sum_{i \in I} p_i z_i, I = \{A, B, C, D, E\}, \tag{6}$$

where $I$ denotes the set of possible emotional categories, $z_i \in \{-2, -1, 0, 1, 2\}$ represents the sentiment score assigned to each emotion, and $p_i$ denotes the empirical probability of a player reporting emotion $i$. A higher PSS indicates a greater tendency toward positive emotional expression, thereby providing a useful quantitative measure of the emotional stability and optimism exhibited by the LLMs during gameplay.

### 3.8 Competition Schedule: Experimental Setup

To comprehensively evaluate the overall capabilities of individual players, a round-robin tournament structure is employed across five distinct game types. Specifically, for each game type, every player competes against each of the other 19 players exactly once, resulting in a total of 190 matches per full round-robin cycle. To mitigate potential biases introduced by variations in player performance or match scheduling order, each tournament is repeated three times with independently randomized match sequences. The final Elo scores, performance metrics, and PSSs for each player are computed by averaging the results across the three independent tournaments. It is noteworthy that the order of matches in each round-robin cycle is fully randomized to ensure fairness and minimize ordering effects. In total, the evaluation encompasses 2,850 = 3 × 190 × 5 games.

| Games | Number of Rounds | Rounds Range |
|---|---|---|
| Gomoku | 5,222 | [9, 129] |
| Chess | 28,931 | [11,299] |
| Reversi | 11,397 | {60} |
| Tic-Tac-Toe | 1,343 | [5, 9] |
| Free-Style | 1,904 | [5, 18] |

Table 1: Average total number of rounds and range of rounds for different classes of games. Values have been rounded.

| Games | Number of Loops | Maximum Loop |
|---|---|---|
| Gomoku | 93 | 17 |
| Chess | 8 | 6 |
| Reversi | 79 | 17 |
| Tic-Tac-Toe | 79 | 19 |
| Free-Style | 93 | 19 |

Table 2: The average number of loops and nodes covered by maximum loop included in PLGs for different games. Values have been rounded.

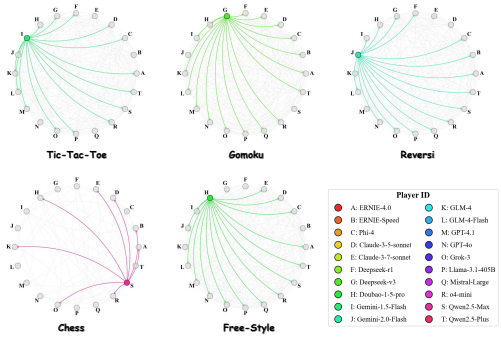



Figure 4: Performance loop graphs (PLGs) for different adversarial games.

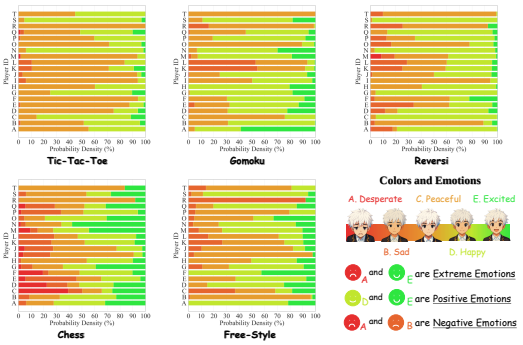



Figure 5: Distribution of emotions across LLMs in different confrontation games.

# 4 Adversarial Competition Results

## 4.1 General Experimental Results

Following the completion of three full round-robin cycles, we compute the average and range of total move counts for each of the five game categories, as summarized in Table 1. Several observations can be drawn from the data: 1) Chess exhibits both the highest average number of moves and the widest range, reflecting its inherent strategic complexity and depth compared to the other board games. 2) The minimum number of moves recorded for Gomoku and Tic-Tac-Toe is 9 and 5, representing the shortest possible paths to victory in these games. This indicates the presence of LLMs with highly proficient strategies in these simpler games, as well as others with limited competence, thereby revealing significant variability in game-specific expertise across different LLMs.

## 4.2 Technical Performance Benchmark

We record each player's final Elo score along with their individual win–loss records, as summarized in Table 3. Several key insights can be derived: 1) Gemini-2.0-Flash attains the highest average performance across all evaluated LLMs, whereas o4-mini records the lowest Elo score. The disparity of over 160 points between the two models underscores the substantial variability in the competitive gameplay capabilities of current LLMs; 2) The top-performing LLMs vary across game categories, indicating that current LLMs exhibit uneven capabilities in reasoning and decision-making, depending on the specific game type; 3) Among the five board games, chess shows the smallest variance in performance across all models. This is largely attributed to the frequent occurrence of repeated positions leading to draws, suggesting that while LLMs can maintain parity in complex scenarios, they still struggle to fully capitalize on winning opportunities in strategically dense environments.

To visualize the pairwise win–loss relationships, we construct PLGs for each game using the "best of three" rule. These are presented in Fig. 4, with detailed visualizations for each game type provided in Sec. B. Additionally, to further quantify the structural properties of the PLGs, we analyze the number of loops and the length of the maximum loop using depth-first search (DFS). The corresponding results are presented in Table 2. From the combined analysis of Fig. 4 and Table 2, an important phenomenon emerges: with the exception of chess, all other games exhibit a high number of loops in their PLGs, and the maximum loop in some cases includes nearly total number of players. Such cyclical relationships underscore two critical observations: the non-transferability of LLM capabilities across different matchups, and the presence of distinct, game-specific strengths within each LLM.

Table 3: Group round robin results. Since the number of wins, losses, and draws are rounded up, the results of the three averages are rounded down.Best in **RED** and Second in **BLUE**.

| LLMs | Label | Tic-Tac-Toe | | | | Gomoku | | | | Reversi | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Elo Score ↑ | Win ↑ | Loss ↓ | Draw | Elo Score ↑ | Win ↑ | Loss ↓ | Draw | Elo Score ↑ | Win ↑ | Loss ↓ | Draw |
| ERNIE-4.0 | A | 1444 | 8 | 11 | 0 | 1435 | 7 | 12 | 0 | 1560 | 13 | 6 | 0 |
| ERNIE-Speed | B | 1498 | 9 | 10 | 0 | 1394 | 4 | 15 | 0 | 1435 | 6 | 13 | 0 |
| Phi-4 | C | 1464 | 6 | 10 | 3 | 1494 | 9 | 10 | 0 | 1408 | 5 | 14 | 0 |
| Claude-3-5-sonnet | D | 1478 | 7 | 11 | 1 | 1600 | 14 | 5 | 0 | 1457 | 8 | 10 | 1 |
| Claude-3-7-sonnet | E | 1592 | 12 | **4** | 3 | 1401 | 5 | 14 | 0 | 1510 | 11 | 8 | 0 |
| Deepseek-r1 | F | 1505 | 8 | 8 | 3 | 1479 | 9 | 10 | 0 | 1478 | 9 | 10 | 0 |
| Deepseek-v3 | G | 1505 | 7 | 8 | 4 | **1670** | **17** | **2** | 0 | 1472 | 8 | 11 | 0 |
| Doubao-1-5-pro | H | 1539 | 10 | 7 | 2 | 1512 | 10 | 9 | 0 | 1474 | 8 | 11 | 0 |
| Gemini-1.5-Flash | I | **1600** | **15** | **4** | 0 | 1489 | 9 | 10 | 0 | 1485 | 9 | 9 | 1 |
| Gemini-2.0-Flash | J | **1593** | **13** | **4** | 2 | 1620 | 14 | 5 | 0 | **1623** | **15** | **4** | 0 |
| GLM-4 | K | 1509 | 8 | 9 | 2 | 1360 | 3 | 16 | 0 | **1590** | **14** | **5** | 0 |
| GLM-4-Flash | L | 1420 | 6 | 13 | 0 | 1374 | 4 | 15 | 0 | 1482 | 9 | 9 | 1 |
| GPT-4.1 | M | 1492 | 8 | 9 | 2 | **1643** | **16** | **3** | 0 | 1517 | 10 | 9 | 0 |
| GPT-4o | N | 1529 | 11 | 8 | 0 | 1584 | 14 | 5 | 0 | 1455 | 6 | 13 | 0 |
| Grok-3 | O | 1499 | 9 | 10 | 0 | 1622 | 15 | 4 | 0 | 1524 | 10 | 9 | 0 |
| Llama-3.1-405B | P | 1452 | 8 | 10 | 1 | 1461 | 9 | 10 | 0 | 1586 | 12 | **5** | 2 |
| Mistral-Large | Q | 1545 | 11 | **6** | 2 | 1508 | 9 | 10 | 0 | 1514 | 10 | 9 | 0 |
| o4-mini | R | 1421 | 5 | 13 | 1 | 1306 | 1 | 18 | 0 | 1485 | 8 | 11 | 0 |
| Qwen2.5-Max | S | 1366 | 4 | 15 | 0 | 1641 | **16** | **3** | 0 | 1466 | 8 | 10 | 1 |
| Qwen2.5-Plus | T | 1550 | 12 | 7 | 0 | 1408 | 5 | 14 | 0 | 1481 | 8 | 11 | 0 |

| LLMs | Label | Chess | | | | Free-Style | | | | Average | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Elo Score ↑ | Win ↑ | Loss ↓ | Draw | Elo Score ↑ | Win ↑ | Loss ↓ | Draw | Elo Score ↑ | Win ↑ | Loss ↓ | Draw |
| ERNIE-4.0 | A | 1471 | 1 | 3 | 15 | 1525 | 11 | 8 | 0 | 1487 | 8 | 8 | 3 |
| ERNIE-Speed | B | 1417 | 0 | 8 | 11 | 1537 | 11 | 8 | 0 | 1456 | 6 | 11 | 2 |
| Phi-4 | C | 1511 | 3 | **2** | 14 | 1444 | 7 | 12 | 0 | 1464 | 6 | 10 | 3 |
| Claude-3-5-sonnet | D | 1510 | 4 | 4 | 11 | 1492 | 8 | 11 | 0 | 1507 | 8 | 8 | 3 |
| Claude-3-7-sonnet | E | 1509 | 4 | 3 | 12 | 1417 | 6 | 13 | 0 | 1485 | 8 | 8 | 3 |
| Deepseek-r1 | F | 1542 | 7 | **2** | 10 | **1591** | **14** | **5** | 0 | 1519 | 9 | 7 | 3 |
| Deepseek-v3 | G | 1552 | 5 | **0** | 14 | 1547 | 11 | 8 | 0 | 1549 | **10** | 6 | 4 |
| Doubao-1-5-pro | H | 1485 | 2 | 4 | 13 | **1606** | **15** | **4** | 0 | 1523 | 9 | 7 | 3 |
| Gemini-1.5-Flash | I | 1417 | 1 | 8 | 10 | 1551 | 12 | 7 | 0 | 1508 | 9 | 8 | 2 |
| Gemini-2.0-Flash | J | 1516 | 2 | **0** | 17 | 1515 | 10 | 9 | 0 | **1573** | **11** | **4** | 4 |
| GLM-4 | K | 1469 | 0 | 3 | 16 | 1478 | 8 | 11 | 0 | 1481 | 7 | 9 | 4 |
| GLM-4-Flash | L | 1474 | 0 | 3 | 16 | 1509 | 10 | 9 | 0 | 1451 | 6 | 10 | 3 |
| GPT-4.1 | M | 1530 | 3 | **0** | 16 | **1591** | **14** | **5** | 0 | **1554** | **10** | **5** | 4 |
| GPT-4o | N | 1533 | 5 | **2** | 12 | 1481 | 9 | 10 | 0 | 1516 | 9 | 8 | 2 |
| Grok-3 | O | 1554 | **6** | **2** | 11 | 1393 | 5 | 14 | 0 | 1518 | 9 | 8 | 2 |
| Llama-3.1-405B | P | **1558** | **6** | **0** | 13 | 1566 | 12 | 7 | 0 | 1524 | 9 | 6 | 3 |
| Mistral-Large | Q | 1505 | 2 | **2** | 15 | 1402 | 6 | 13 | 0 | 1494 | 8 | 8 | 3 |
| o4-mini | R | 1444 | 1 | 6 | 12 | 1395 | 5 | 14 | 0 | 1410 | 4 | 12 | 3 |
| Qwen2.5-Max | S | **1567** | **8** | **2** | 9 | 1467 | 7 | 12 | 0 | 1501 | 9 | 8 | 2 |
| Qwen2.5-Plus | T | 1438 | 2 | 8 | 9 | 1493 | 9 | 10 | 0 | 1493 | 7 | 10 | 2 |

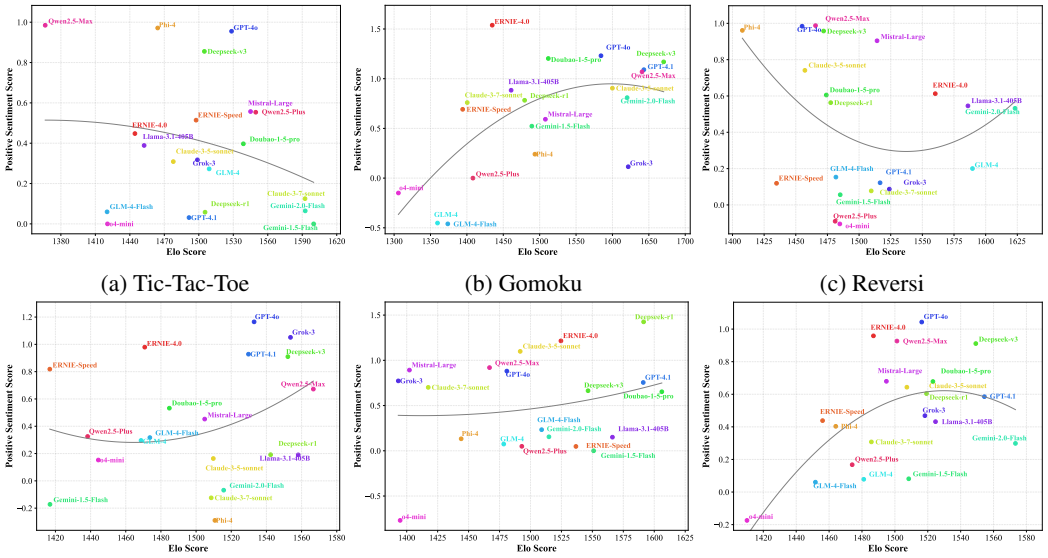
### 4.3 Emotional Positivity Benchmark

We analyze the distribution of emotional responses exhibited by each player across different game types and the aggregated results are presented in Fig. 5 and Table 4. Several key observations can be drawn from analysis: 1) Overall, the majority of emotional responses are concentrated around Peaceful, Happy, and Excited, suggesting that most LLMs are capable of maintaining a predominantly positive affective state during gameplay; 2) GPT-4o consistently exhibits the most positive emotional profile across all games, followed closely by ERNIE-4.0. In contrast, o4-mini demonstrates a generally negative emotional tendency, particularly under the Free-Style game type, where strong negative emotions are frequently observed. This pattern of emotional volatility in o4-mini warrants attention, as it may pose risks in applications where emotional regulation is essential; 3) Game type appears to significantly influence the emotional states of the LLMs. For example, Chess elicits a broader range of extreme emotions, likely due to its strategic complexity and high cognitive demands. Conversely, games such as Tic-Tac-Toe induce minimal emotional fluctuation, with very few extreme affective responses observed.

To investigate the relationship between the technical and psychological performance of LLMs in board games, we present the analysis results in Fig. 6. Several key observations emerge from this figure: 1) There is no significant correlation between technical performance and emotional stability across all LLMs, thereby supporting the rationale for evaluating these two dimensions independently; 2) The correlation between technical and psychological dimensions appears to be task-dependent. For instance, in Gomoku, a moderate positive correlation is observed, while in Tic-Tac-Toe, a negative correlation is evident; 3) Some LLMs exhibit high technical performance while displaying

Table 4: Positive Sentiment Scores (PSSs) for different LLMs. Best in **RED** and Second in **BLUE**.

| LLMs | Label | Positive Sentiment Score ↑ | | | | | |
|---|---|---|---|---|---|---|---|
| | | Tic-Tac-Toe | Gomoku | Reversi | Chess | Free-Style | Average |
| ERNIE-4.0 | A | 0.4478 | **1.5382** | 0.6126 | **0.9794** | **1.2150** | **0.9586** |
| ERNIE-Speed | B | 0.5143 | 0.6918 | 0.1191 | 0.8183 | 0.0482 | 0.4383 |
| Phi-4 | C | **0.9714** | 0.2407 | 0.9612 | -0.2896 | 0.1327 | 0.4032 |
| Claude-3-5-sonnet | D | 0.3088 | 0.9043 | 0.7411 | 0.1636 | 1.0990 | 0.6434 |
| Claude-3-7-sonnet | E | 0.1250 | 0.7596 | 0.0773 | -0.1244 | 0.7010 | 0.3077 |
| Deepseek-r1 | F | 0.0580 | 0.7828 | 0.5632 | 0.1917 | **1.4253** | 0.6042 |
| Deepseek-v3 | G | 0.8553 | 1.1693 | 0.9577 | 0.9097 | 0.6639 | 0.9112 |
| Doubao-1-5-pro | H | 0.3971 | 1.2033 | 0.6053 | 0.5325 | 0.6526 | 0.6782 |
| Gemini-1.5-Flash | I | 0.0000 | 0.5226 | 0.0561 | -0.1721 | 0.0000 | 0.0813 |
| Gemini-2.0-Flash | J | 0.0645 | 0.8093 | 0.5322 | -0.0684 | 0.1546 | 0.2984 |
| GLM-4 | K | 0.2727 | -0.4521 | 0.2004 | 0.2961 | 0.0745 | 0.0783 |
| GLM-4-Flash | L | 0.0597 | -0.4603 | 0.1533 | 0.3169 | 0.2333 | 0.0606 |
| GPT-4.1 | M | 0.0312 | 1.0905 | 0.1220 | 0.9280 | 0.7551 | 0.5854 |
| GPT-4o | N | 0.9552 | **1.2308** | **0.9842** | **1.1649** | 0.8810 | **1.0432** |
| Grok-3 | O | 0.3175 | 0.1144 | 0.0871 | 1.0512 | 0.7727 | 0.4686 |
| Llama-3.1-405B | P | 0.3889 | 0.8839 | 0.5455 | 0.1904 | 0.1505 | 0.4318 |
| Mistral-Large | Q | 0.5571 | 0.5920 | 0.9040 | 0.4526 | 0.8922 | 0.6796 |
| o4-mini | R | 0.0000 | -0.1497 | -0.1046 | 0.1526 | -0.7692 | -0.1742 |
| Qwen2.5-Max | S | **0.9848** | 1.0690 | **0.9877** | 0.6722 | 0.9195 | 0.9266 |
| Qwen2.5-Plus | T | 0.5536 | 0.0000 | -0.0898 | 0.3250 | 0.0495 | 0.1677 |




(a) Tic-Tac-Toe (b) Gomoku (c) Reversi

(d) Chess (e) Free-Style (f) Average

Figure 6: Distribution of Elo scores and PSSs for different LLMs across games.

predominantly negative emotional states, emphasizing their limitations in terms of psychological adjustment; 4) Most LLMs, regardless of technical performance, can maintain a positive mindset, suggesting that they are better adapted to high-stress competitive environments than humans.

## 5 Conclusion

In conclusion, this study introduces LLM against LLM, a novel adversarial benchmark framework based on board games, to assess the technical and psychological capabilities of LLMs in an autonomous, interaction-rich environment. By developing the Qi Town, featuring five distinct board games and 20 LLM-driven players, we provide a rigorous and scalable testbed platform for evaluating reasoning ability, decision-making strategies, and emotional resilience. Through the integration of the Elo scoring system, performance loop graphs, and the proposed Positive Sentiment Score, we provide a multi-dimensional benchamrk that captures both quantitative performance and affective stability. Our findings not only reveal significant differences in strategic and emotional behaviors across LLMs but also challenge the limitations of traditional Q&A-based benchmarks. This work lays the foundation for more holistic and behaviorally grounded adversial benchmark methods, and offers valuable insights for the design of more adaptive, robust, and emotionally intelligent LLMs.

# References


[1] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010, 2023.

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

[3] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[4] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020.

[5] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*, 2024.

[7] Yingjie Zhou, Zicheng Zhang, Jiezhang Cao, Jun Jia, Yanwei Jiang, Farong Wen, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Memo-bench: A multiple benchmark for text-to-image and multimodal large language models on human emotion analysis. *arXiv preprint arXiv:2411.11235*, 2024.

[8] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

[9] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmarking the video quality understanding of lmms. *arXiv preprint arXiv:2409.20063*, 2024.

[10] Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024.

[11] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[12] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

[13] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

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

[15] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[16] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[17] Xinyu Wang, Bohan Zhuang, and Qi Wu. Are large vision language models good game players? *arXiv preprint arXiv:2503.02358*, 2025.

[18] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael Lyu. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] W Brian Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994.

[20] Anthropic. `https://www.anthropic.com/claude/sonnet`, 2025.

[21] Anthropic. Model Card Claude 3 Addendum. `https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/\protect\penalty\z@Model_Card_Claude_3_Addendum.pdf`, 2024. Accessed: 2024-03-15.

[22] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[24] ByteDance. `https://seed.bytedance.com/en/special/doubao_1_5_pro`, 2025.

[25] Baidu. `https://cloud.baidu.com/doc/WENXINWORKSHOP/s/6ltgkzya5`, 2025.

[26] Baidu. `https://cloud.baidu.com/doc/WENXINWORKSHOP/s/clntwmv7t`, 2025.

[27] Meta. `https://ai.meta.com/blog/meta-llama-3-1`, 2024.

[28] Mistral. `https://mistral.ai/news/mistral-large`, 2024.

[29] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[30] xAI. `https://x.ai/news/grok-3`, 2025.

[31] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[32] DeepMind. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0-flash`, 2024.

[33] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[34] OpenAI. `https://openai.com/index/gpt-4/`, 2025.

[35] OpenAI. `https://openai.com/index/hello-gpt-4o/`, 2024.

[36] OpenAI. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[37] Piyush N Shinde, Yash J Chavan, Shubham G Chilka, Gaurav N Patil, and Manoj Raghunath Kharde. Pygame: Develop games using python. *International Journal for Research in Applied Science and Engineering Technology*, 2021.

[38] N. Fiekas. python-chess: a chess library for python. https://github.com/niklasf/python-chess, 2024. [Online].

[39] Arpad E. Elo and Sam Sloan. *The Rating of Chess Players, Past and Present*. Ishi Press, 2008.

# A Typical Game Analysis

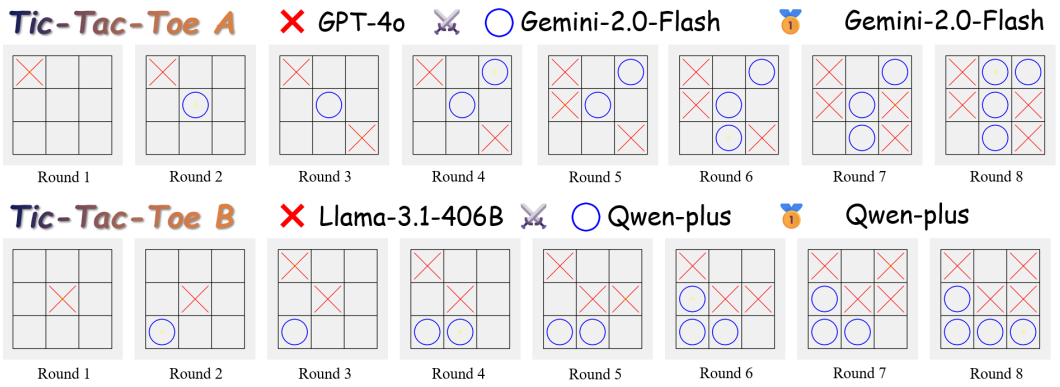



Figure 7: Example of typical Tic-Tac-Toe games.

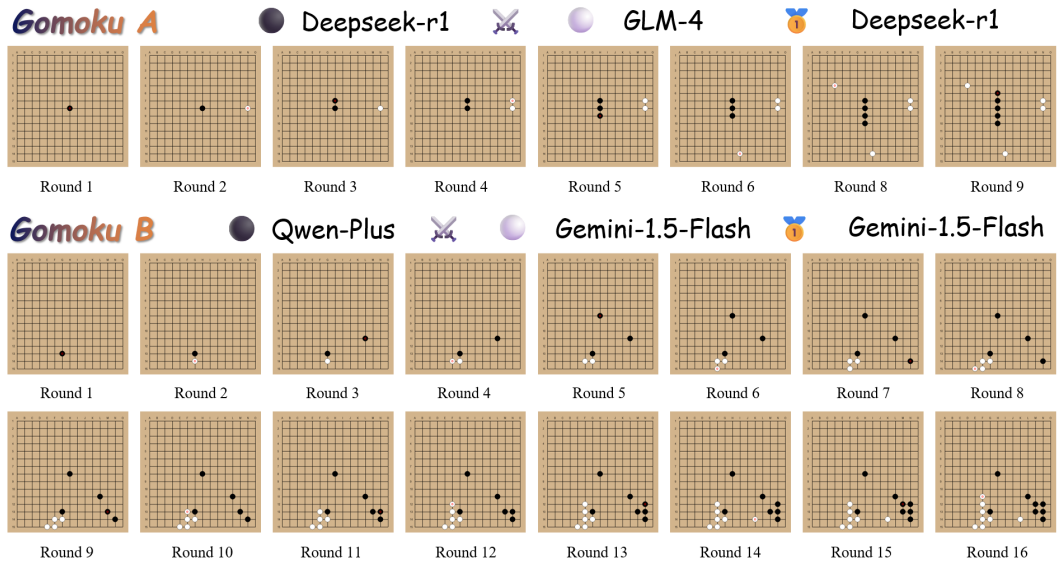



Figure 8: Example of typical Gomoku games.

## A.1 Tic-Tac-Toe

Tic-Tac-Toe is a fundamental two-player strategy game typically played on a 3×3 grid. Players alternate turns, marking empty cells with either "X" or "O," with the objective of aligning three of their symbols consecutively—horizontally, vertically, or diagonally. The game ends in a draw if all cells are occupied without a winner. Despite its simplicity, Tic-Tac-Toe engages a player's short-term tactical planning and requires effective balancing between offensive and defensive strategies.

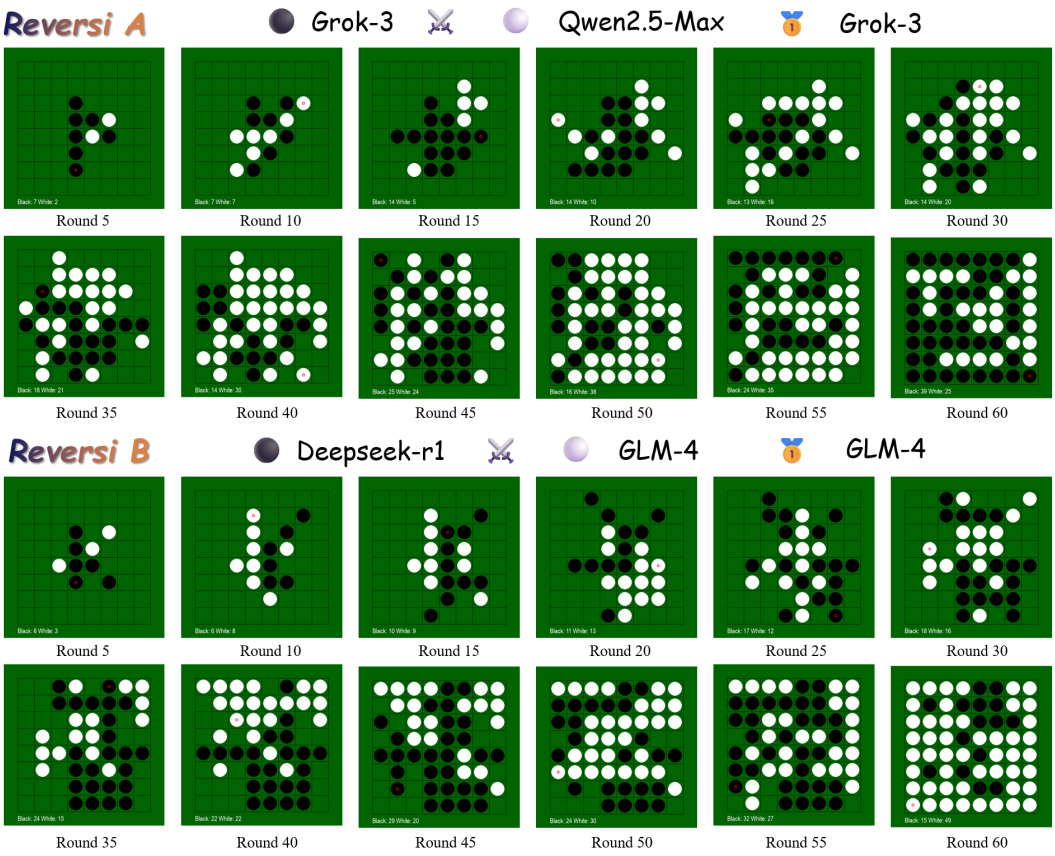



Figure 9: Example of typical Reversi games.

To illustrate the decision-making behaviors of large language models (LLMs) within a constrained strategic environment, we present two representative scenarios in Fig. 7. In Tic-Tac-Toe A, Gemini-2.0-Flash missed a clear opportunity to win during Round 6, failing to capitalize on a straightforward winning move. Although GPT-4o eventually secured a win in Round 7, its earlier move in Round 7 neither ensured its own victory nor effectively blocked potential threats from the opponent. In Tic-Tac-Toe B, both LLMs continued to demonstrate suboptimal choices, revealing consistent deficiencies in both offensive and defensive planning. These examples suggest that current LLMs still exhibit notable limitations in strategic reasoning and decision-making under simple, rule-based conditions.

## A.2 Gomoku

Gomoku extends the conceptual structure of Tic-Tac-Toe to a more complex level. Played on a 15×15 board, two players alternate placing black or white pieces on empty intersections. A win is achieved by aligning exactly five consecutive stones of the same color (alignments exceeding five are not valid). In this study, a free opening rule is adopted with no forbidden-move constraints, simplifying implementation while preserving strategic complexity. The larger board and stricter victory conditions significantly expand the decision space and require deeper combinatorial reasoning and predictive planning, often involving advanced techniques such as Victory by Continuous Four (VCF) and Victory by Continuous Threat (VCT).

To further investigate the gameplay behavior of LLMs, we analyze two representative match examples depicted in Fig. 8. In Gomoku A, the game exhibits a classic quick-win pattern. The player Deepseek-r1 demonstrates consistent strategic intent, incrementally consolidating its advantage with each move. In contrast, GLM-4 displays disorganized and scattered play, indicative of poor strategic coherence and poor performance. Conversely, Gomoku B presents a more balanced and competitive match. From the outset, both players initiated their moves near the edges or corners of the board—positions generally considered disadvantageous for early dominance. Ultimately, Qwen2.5-Plus lost the game due to an overemphasis on developing its own formation in the lower-right quadrant while failing

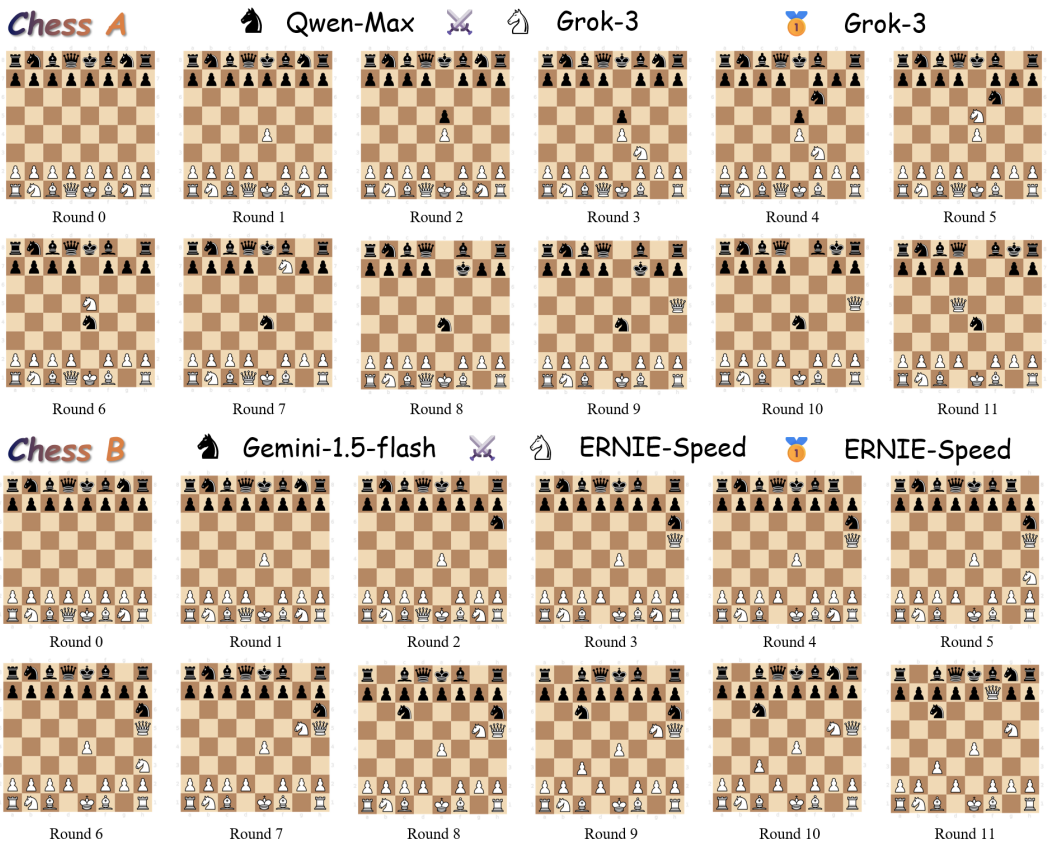



Figure 10: Example of typical Chess games.

to adequately defend against Gemini-1.5-Flash's emerging threats. These two cases highlight not only the varied levels of proficiency and adaptability of LLMs in navigating Gomoku's complex strategic landscape, but also expose persistent limitations in global planning and situational awareness exhibited by LLMs.

### A.3 Reversi

Reversi is a turn-based strategy game conducted on an 8×8 board, with the objective of capturing the majority of the board using one's color discs by the end of the game. The game begins with two black and two white discs placed diagonally in the center. Players must place discs such that they flank one or more of the opponent's discs in a straight line, which are then flipped to their own color. Reversi demands not only foresight and local tactical precision but also a strong capacity for global board evaluation and strategic adaptability throughout the game.

Fig. 9 presents two illustrative matches selected from the round-robin tournament. In Reversi A, an underdog scenario unfolds in which Grok-3 secures a surprising victory against the stronger opponent, Qwen2.5-VL. Although Grok-3 remained at a significant disadvantage for the first 55 moves, weaknesses in Qwen2.5-VL's overall positional strategy enabled Grok-3 to stage a late-game reversal. By contrast, Reversi B exemplifies a well-balanced contest, with both LLMs demonstrating relatively comparable performance and strategic depth. These representative examples suggest that while some LLMs exhibit promising situational tactics, many still fall short in terms of comprehensive global coordination and long-term planning in complex, dynamic environments.

### A.4 Chess

Chess is a globally standardized adversarial board game played on an 8×8 grid, involving 16 pieces per player, including the King, Queen, Rooks, Bishops, Knights, and Pawns—each with unique movement rules. The game objective is to checkmate the opponent's King while maintaining one's own defensive integrity. In this study, advanced gameplay features are incorporated, including

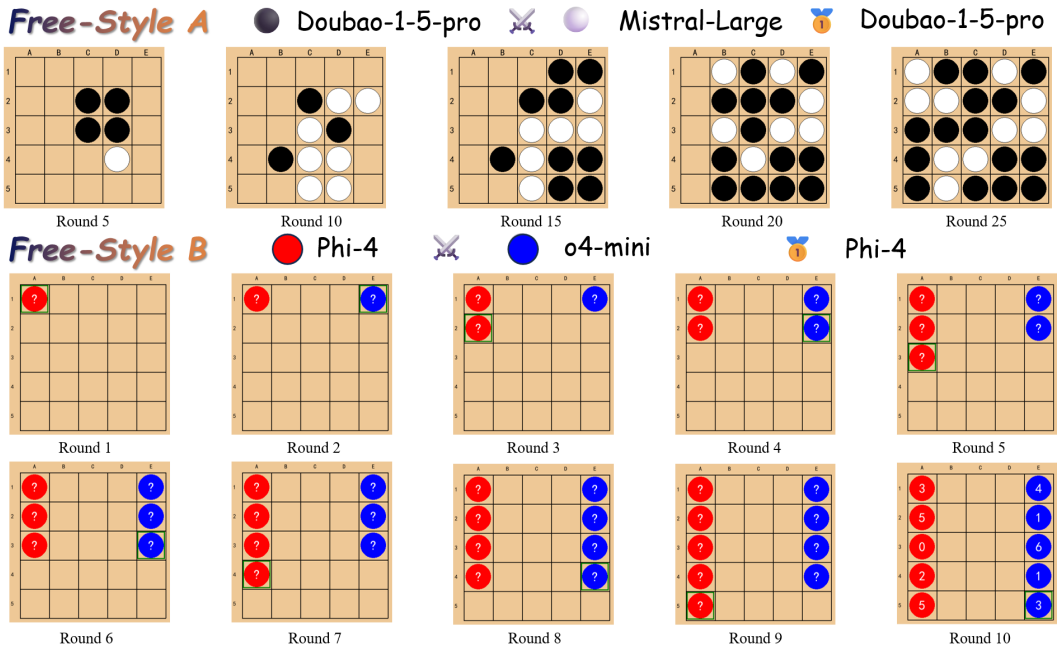



Figure 11: Example of typical Free-Style games.

castling, pawn promotion, en passant, and threefold repetition, to approximate realistic competitive conditions. Chess challenges players with both deep combinatorial calculation and long-term strategic planning, making it a benchmark for assessing cognitive and metacognitive competencies.

As shown in Table 1, Chess exhibits the broadest distribution of game lengths among all evaluated game types. To aid in interpretation and analysis, we select two representative, relatively simple matches for discussion as shown in Fig. 10. In Chess A, both players began with a mirrored opening; however, Grok-3 adopted a more aggressive posture early in the game. This pressure prompted Qwen2.5-Max to prematurely reposition its King to evade Grok-3's Queen, ultimately leading to a decisive checkmate executed by Grok-3. In Chess B, the game initially proceeded with standard development for the first nine moves. However, on Round 10, Gemini-1.5's decision to withdraw its Knight created a critical vulnerability, which ERNIE-Speed exploited through a coordinated attack involving both its Knight and Queen, culminating in a swift victory. Collectively, these examples indicate that while some LLMs demonstrate foundational competence in Chess, others continue to exhibit deficiencies in tactical coherence and strategic organization under adversarial conditions.

### A.5 Free-style

Free-Style is an adversarial board game characterized by its flexible rule system. Prior to gameplay, both players engage in a negotiation phase to collaboratively define the game rules, thereby reaching a mutual agreement. To manage the complexity of the resulting rule sets, all experiments are conducted on a fixed 5×5 board. The game rules, established through negotiation, reveal a rich diversity in both the design of game pieces and gameplay mechanics when controlled by LLM-driven agents. Frequently, numeric values are used as game pieces, and piece grading emerged as a common structural feature across various games. In terms of gameplay styles, while conventional strategies such as forming three-in-a-line are observed, more intricate formats such as adaptations inspired by Animal Chess are also explored.

To illustrate the breadth of game types enabled by Free-Style, two representative examples are visualized in Fig. 11. The first, Free-Style A, bears resemblance to Reversi; however, it introduces greater freedom in selecting drop positions. Additionally, players flip the color of all adjacent, 4-connected pieces upon placement, introducing a strategic balance between offensive and defensive positioning to maximize territorial control. The second game, Free-Style B, titled "Hidden Numbers," emphasizes strategic reasoning. In this variant, players alternately place numerical values between 0 and 15 on the first and fifth columns of the board, under the constraint that the total sum does not

exceed 15. These values are initially hidden and displayed only at the end of the game. Scoring is based on the count of higher-valued digits per row, with one point awarded per row win. The player with the highest cumulative score emerges victorious. This game format demands a higher level of strategic foresight from LLM agents due to its hidden-information structure and arithmetic constraints.

## B Details of Performance Loop Graph

While Fig. 4 presents the Performance Loop Graphs (PLGs) corresponding to the top-performing player in each game category to facilitate intuitive understanding, the PLGs of other participating LLMs are not ignored. To provide a more comprehensive overview of our experimental findings, we now present the PLGs for all players across all game types, as illustrated in Figs. 12 to 16. Several key observations can be derived from these figures: 1) PLGs offer a more visually intuitive representation of the win–loss dynamics among LLMs than the aggregate win/loss/draw statistics reported in Table 3; 2) The PLGs associated with the same model differ substantially across game types, reinforcing the observation that LLMs exhibit inconsistent performance across different strategic environments; 3) The PLG visualizations reveal the presence of cyclical win–loss relationships among certain LLMs, thus forming closed loops. This phenomenon introduces a novel lens through which to evaluate the stability and relative significance of LLMs' competitive capabilities.

## C Temporal Analysis of Player Sentiment

Player emotion is inherently dynamic during gameplay, evolving in response to in-game events and strategic developments. As such, analyzing emotional characteristics solely through the static statistical methods outlined in Sec. 4.3 presents certain limitations. To further investigate the temporal dynamics of emotional responses in individual players, we introduce an emotional time-series heatmap visualization, shown in Figs 17 to 21. In these visualizations, the horizontal axis represents the number of rounds played, the vertical axis corresponds to the Positive Sentiment Score, and

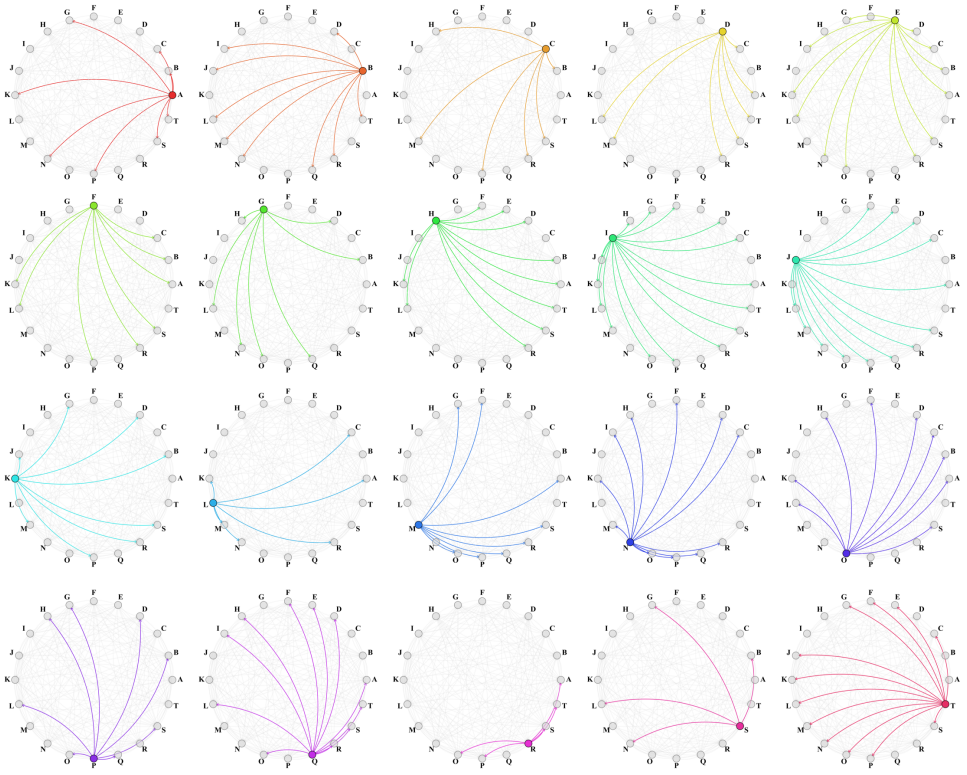


Figure 12: PLGs for Tic-Tac-Toe. The labels and colors of the nodes are consistent with Fig. 4.

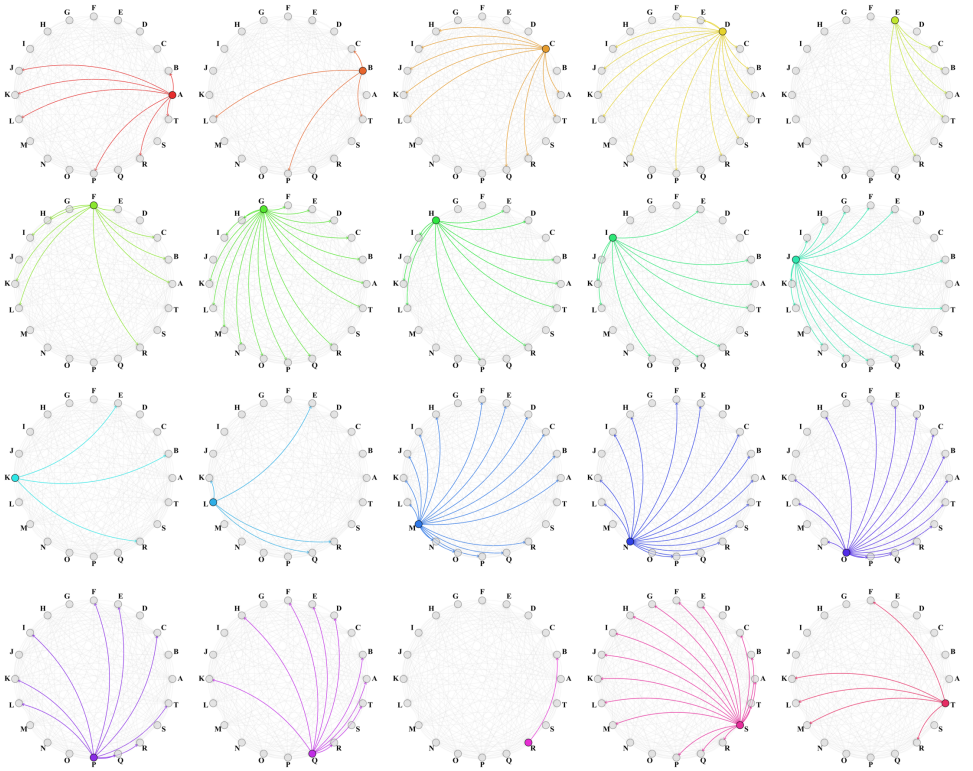


Figure 13: PLGs for Gomoku. The labels and colors of the nodes are consistent with Fig. 4.

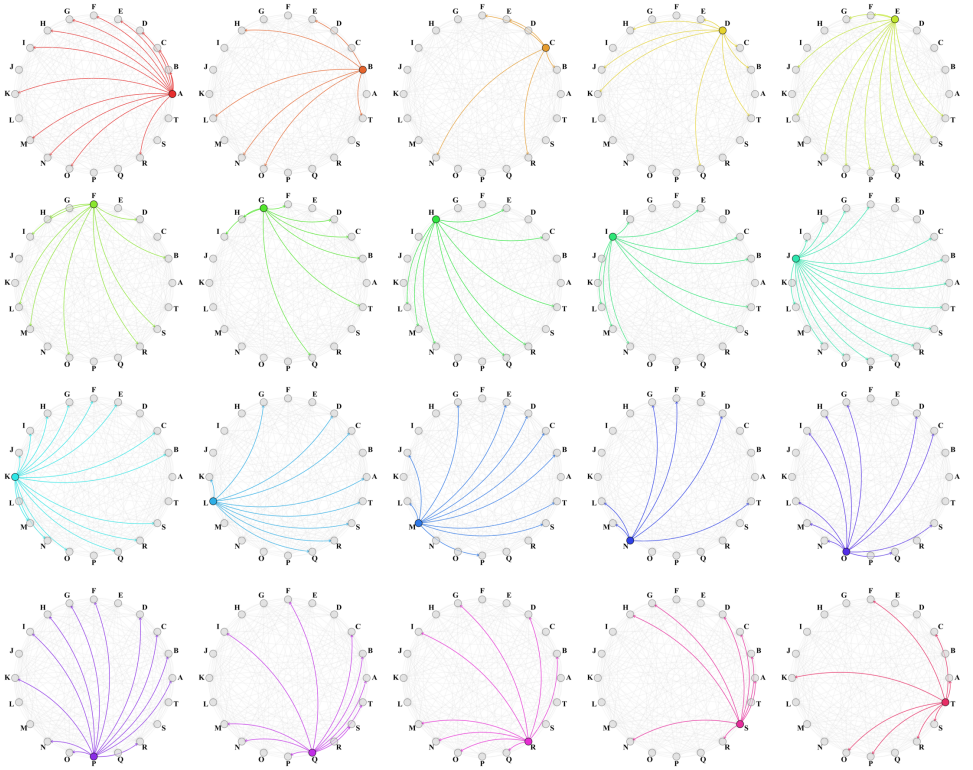


Figure 14: PLGs for Reversi. The labels and colors of the nodes are consistent with Fig. 4.

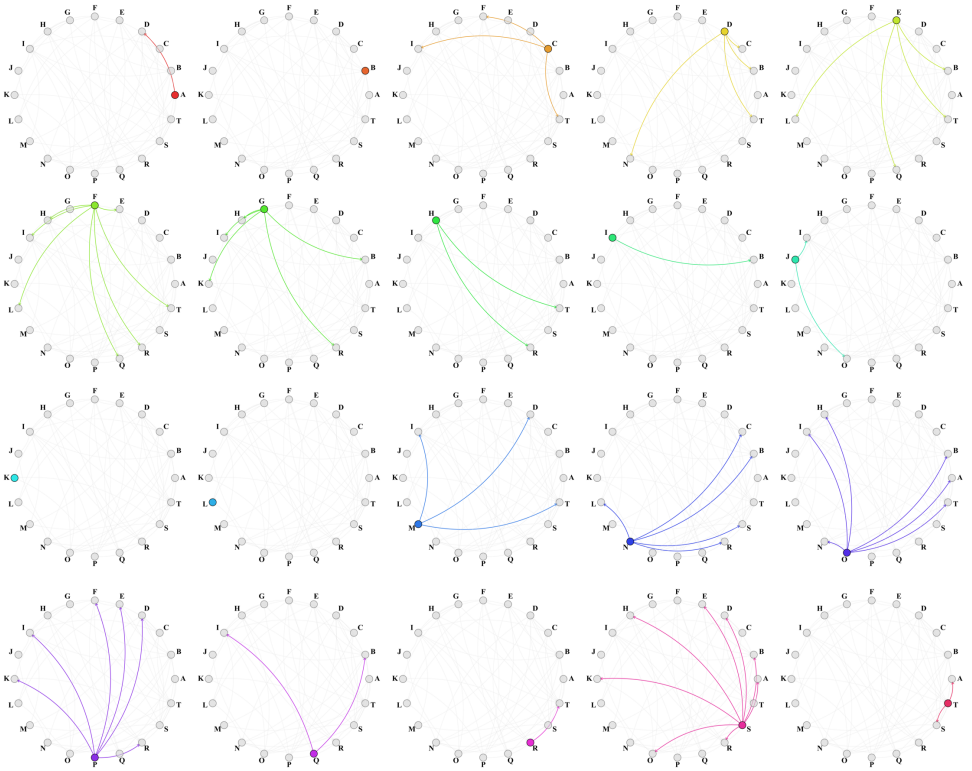


Figure 15: PLGs for Chess.The labels and colors of the nodes are consistent with Fig. 4.

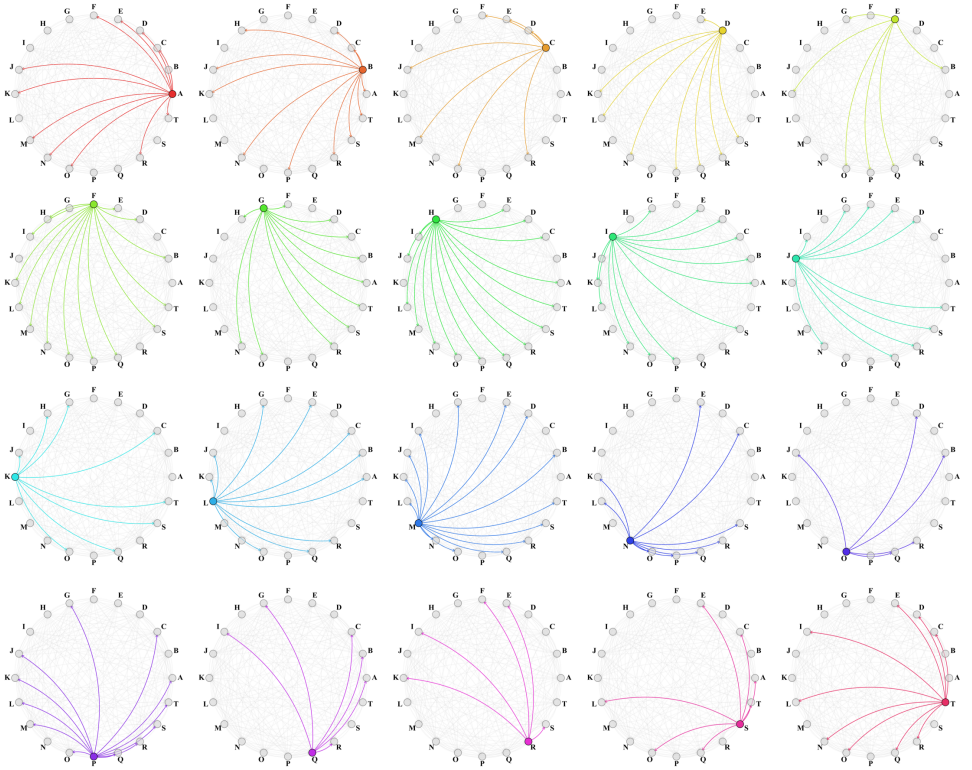


Figure 16: PLGs for Free-Style. The labels and colors of the nodes are consistent with Fig. 4.

the color intensity within each hexagonal bin indicates the frequency of specific emotional states over time. Beyond the findings reported in Sec. 4.3, several additional insights emerge from the heatmaps: 1) Emotional trajectories differ markedly across game types. Notably, all LLMs tend to exhibit greater emotional volatility and more frequent negative sentiment when engaged in Chess, suggesting that cognitively demanding games exert a stronger influence on the emotional responses of LLMs compared to simpler games; 2) There is significant variation in the emotional dynamics across different models, even within the same game category. This heterogeneity may reflect underlying differences in the affective tendencies or simulated personality traits of the LLMs; 3) In addition to providing emotional data, the heatmaps also convey information about game duration. For instance, in Fig. 18, both ERNIE-Speed and o4-mini extend their Gomoku matches beyond 120 rounds, while most other players reach a conclusion before 100 rounds. Similar observations across other game types offer further insight into each LLM's ability to manage gameplay and adapt strategy over time.

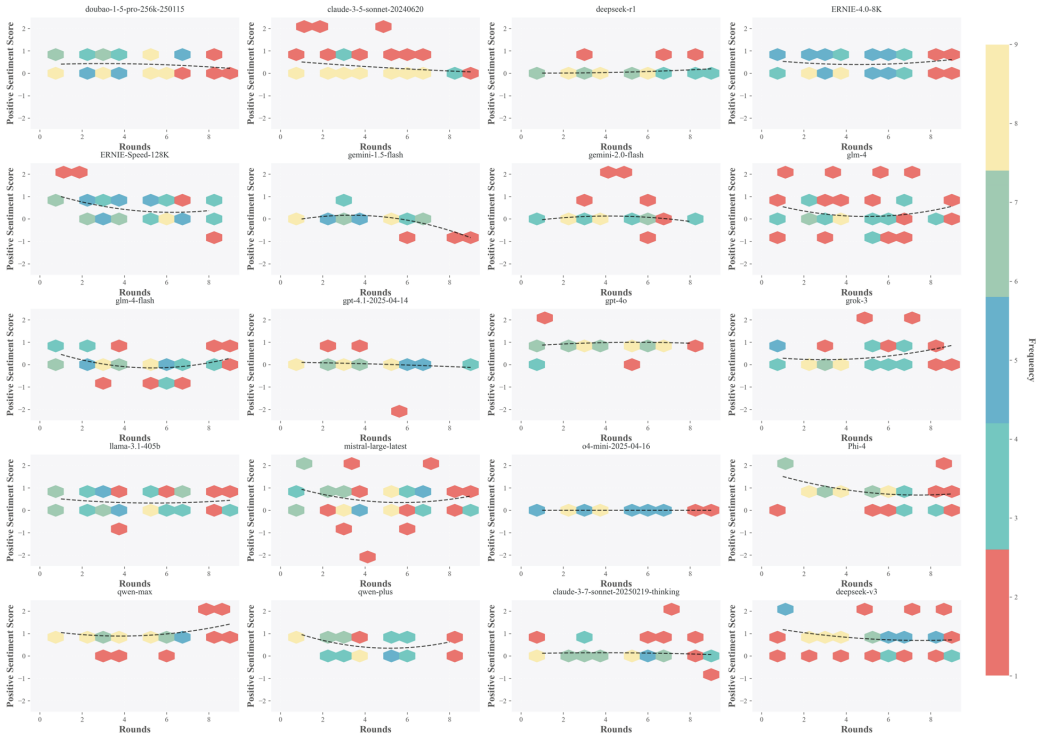



Figure 17: Emotional time-series heatmap for Tic-Tac-Toe.


Figure 18: Emotional time-series heatmap for Gomoku.

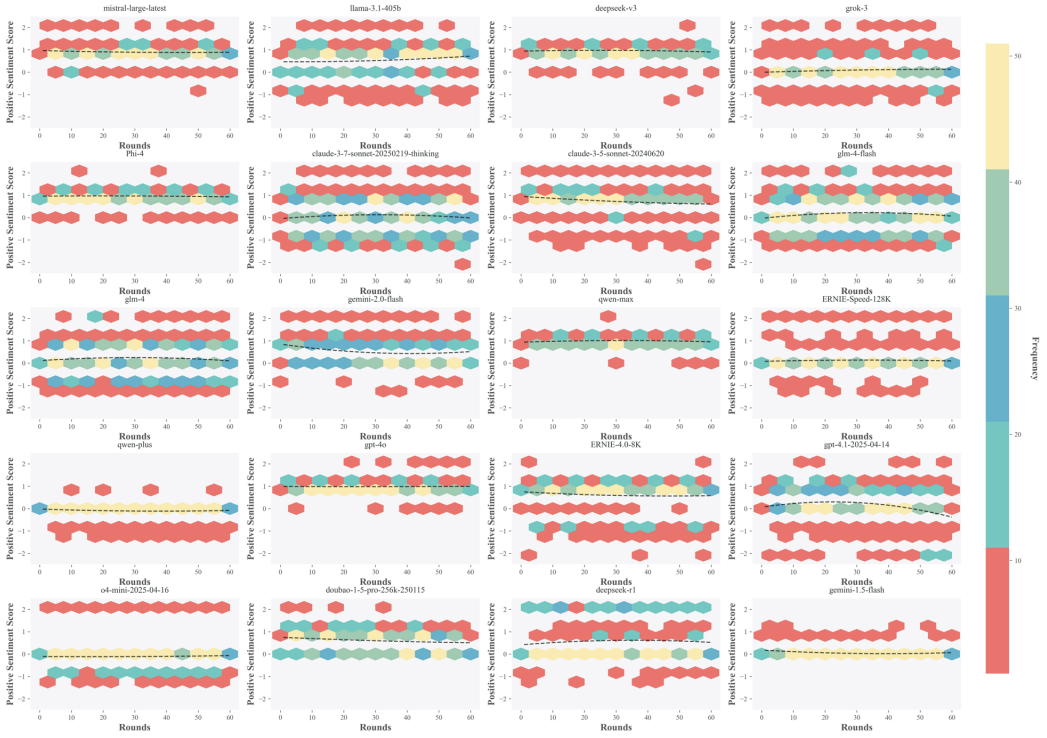



Figure 19: Emotional time-series heatmap for Reversi.

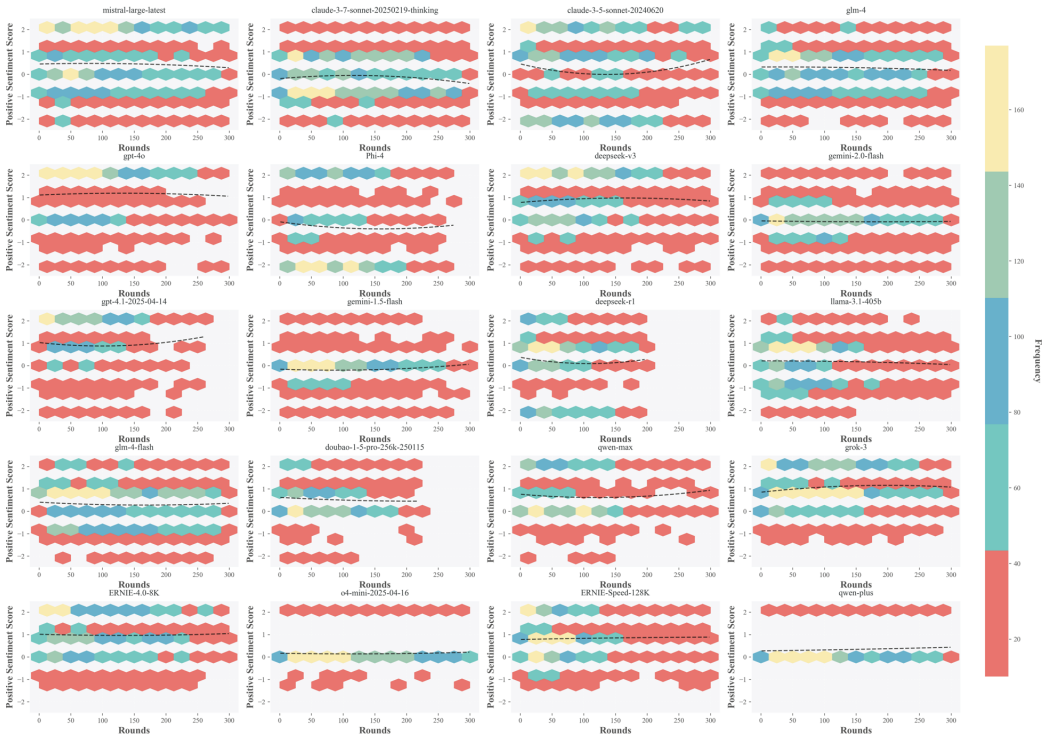



Figure 20: Emotional time-series heatmap for Chess.

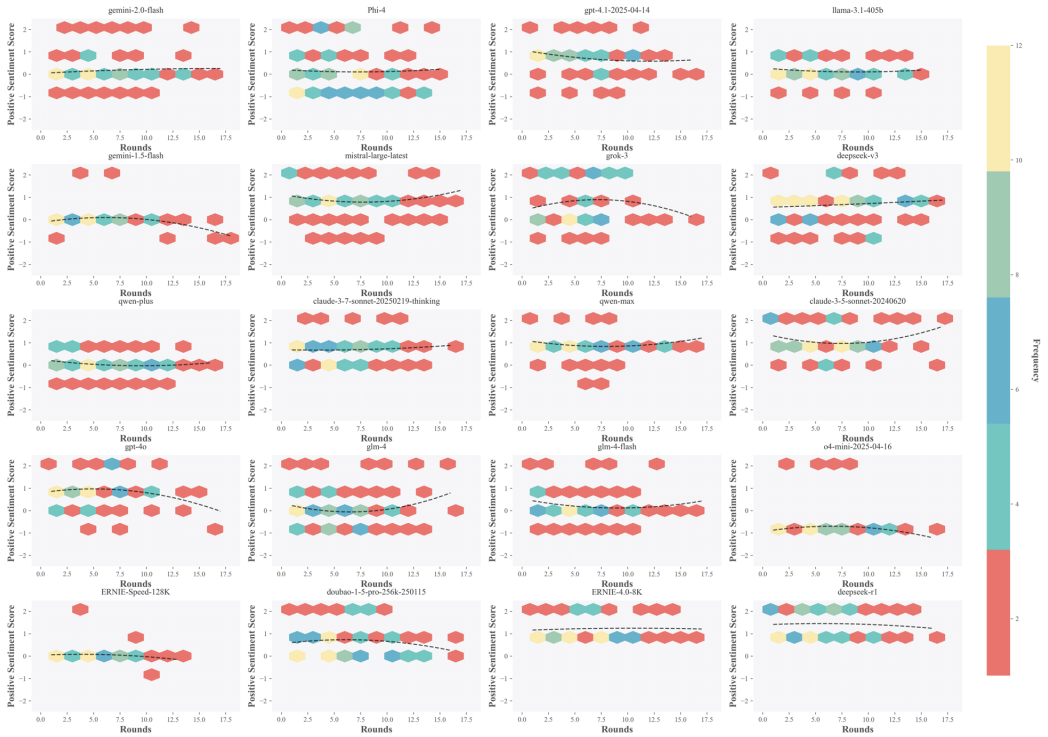



Figure 21: Emotional time-series heatmap for Free-Style.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution of this paper has been highlighted in both the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: We will explain the limitations of this work and possible solutions in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For the theories mentioned in the text, we provide sufficient formulas to prove them.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have given the experimental setup and related content in detail in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: While we are holding off on providing a link to open source access at this time, we will make all experimental data (including all game logs) available after the paper is accepted to facilitate follow-up research.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification:The relevant settings and details of the experiment have been clearly stated.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The experimental data are statistically analyzed and specific methods of analysis are given in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The text describes the relevant settings for the APIs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors currently believe that no ethical issues require special attention.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work in this paper will contribute to the further development of LLM.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The main experimental data of this paper (game logs) will be released after the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The main subject of this paper is LLMs and the use of LLMs has been asserted.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.