LLMsPark: A Benchmark for Evaluating Large Language Models in Strategic Gaming Contexts

Junhao Chen[1,6], Jingbo Sun[2], Xiang Li[3], Haidong Xin[4], Yuhao Xue[5], Yibin Xu[5], Hao Zhao[6,7]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Institute of Computing Technology, Chinese Academy of Sciences,
[3]School of Software and Microelectronics, Peking University,
[4]School of Computer Science and Engineering, Northeastern University,
[5]Tongji University, [6]AIR, Tsinghua University [7]BAAI

Abstract

As large language models (LLMs) advance across diverse tasks, the need for comprehensive evaluation beyond single metrics becomes increasingly important. To fully assess LLM intelligence, it is crucial to examine their interactive dynamics and strategic behaviors. We present LLMsPark, a game theory–based evaluation platform that measures LLMs' decision-making strategies and social behaviors in classic game-theoretic settings, providing a multi-agent environment to explore strategic depth. Our system cross-evaluates 15 leading LLMs (both commercial and open-source) using leaderboard rankings and scoring mechanisms. Higher scores reflect stronger reasoning and strategic capabilities, revealing distinct behavioral patterns and performance differences across models. This work introduces a novel perspective for evaluating LLMs' strategic intelligence, enriching existing benchmarks and broadening their assessment in interactive, game-theoretic scenarios. The benchmark and rankings are publicly available at https://llmsparks.github.io/.

1 Introduction

With the rapid rise of large language models (LLMs) and large multimodal models (LMMs), their performance on complex tasks such as code generation (Liu et al., 2024a; Guo et al., 2025), recommender systems (Liu et al., 2023; Xin et al., 2025), and knowledge-intensive question answering has exceeded expectations, reshaping the landscape of natural language processing. Beyond text, these models demonstrate broad applicability in text-guided generation and editing of images (Huang et al., 2024; Chen et al., 2023a; Yang et al., 2025), 3D models (Hong et al., 2023; Sun et al., 2025; Wang et al., 2024; Chen et al., 2025b; Fu et al., 2024; Chen et al., 2024), as well as audio and video understanding and editing (Shu et al., 2023; Ye et al., 2024; Fei et al., 2024; Lin et al., 2023; Chen et al., 2025a). The release of GPT-4 (OpenAI, 2023a) and the rapid development of open-source models such as Llama2 (Touvron et al., 2023b) and ChatGLM2 (Du et al., 2022a) have further accelerated this progress. In real-world applications–ranging from question answering (Rajpurkar et al., 2016), natural language inference (Bowman et al., 2015), and text summarization (Nallapati et al., 2016) to sentiment classification (Chen et al., 2023b)–the performance of LLMs is now approaching, and in some cases rivaling, human-level abilities. They also exhibit strong competence in mathematical problem solving (Gaur and Saunshi, 2023), logical reasoning (Wei et al., 2022), and even single-player games.

However, evaluating the capabilities of LLMs remains a significant challenge. Current evaluation paradigms are dominated by static benchmarks and large-scale knowledge-intensive datasets, such as MMLU (Hendrycks et al., 2021a), MT-bench (Zheng et al., 2023a), Chatbot Arena (Zheng et al., 2023a), Zhujiu (Zhang et al., 2023), and OpenAI Evals (OpenAI, 2023b). While informative, these benchmarks primarily assess factual recall or task-specific performance, offering limited insights into interactive reasoning and adaptive behaviors. Recent efforts have begun to extend evaluation into dynamic and agentic settings. For example, tools such as XAgent (Team, 2023) and AutoGPT (Richards, 2023) test LLMs in autonomous workflows, while social games are increasingly recognized as promising testbeds for examining AI decision-making and strategic behavior (Xi et al., 2023). LLM-based agents have also been deployed in interactive environments, including generative social simulations (Park et al., 2023a), open-world games like Minecraft (Zhu et al., 2023), and multi-agent role-playing games such as Werewolf (Xu et al., 2023a). Recent studies have also evaluated LLMs in complex visual games such as Smart-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 182–194
November 4-9, 2025

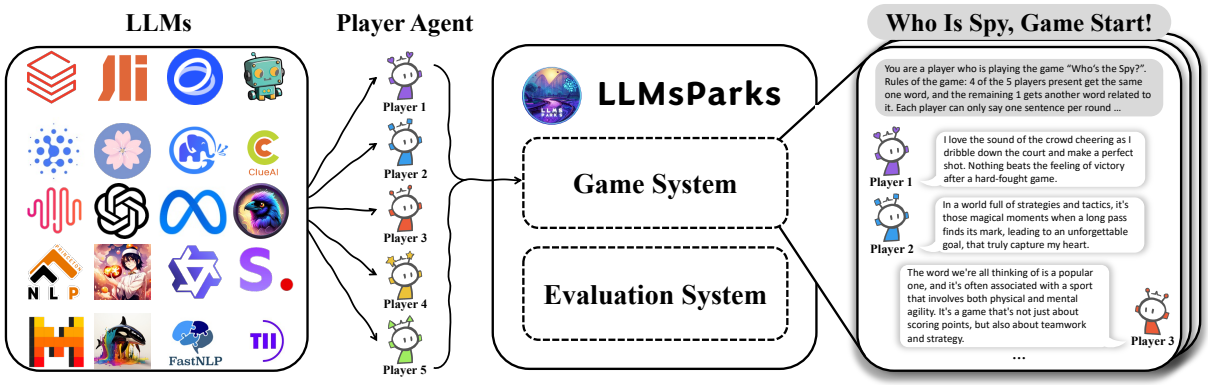

Figure 1: LLMsPark is the first benchmark that evaluates LLMs as agents in game-theoretic settings. In its initial release, it assesses 15 LLMs across five games. The Game System is detailed in Section 3.2, the Player Agent in Section 3.3, and the Evaluation System in Section 3.4.

Play (Wu et al., 2024a), though these settings often rely on multimodal inputs and are less applicable to text-only models. At the same time, there is growing interest in probing higher-level dimensions of LLM intelligence, including ethical reasoning and theory of mind capabilities (Guo et al., 2023).

As illustrated in Figure 1, we introduce LLMsPark, a dynamic benchmark that leverages classic game-theoretic settings such as the Prisoner's Dilemma and the Trust Game to evaluate LLMs' strategic and social behaviors. LLMsPark enables text-based models to autonomously participate in these games, offering new insights into how they manage cooperation, deception, and competition in multi-agent scenarios. Our study reveals unexpected behavioral patterns and highlights the potential of game-based environments as rigorous evaluation tools. We release the benchmark, model rankings, and results at `https://llmsparks.github.io`, aiming to enrich the evaluation landscape and foster future research. As a pioneering platform for assessing LLMs' social and strategic intelligence, LLMsPark will continue to expand with more complex and diverse games. Our contributions are threefold:

- *Game-theoretic Evaluation of LLMs.* We introduce a benchmark grounded in classic games to systematically assess LLM decision-making in interactive contexts.
- *Behavioral Analysis of LLMs.* We uncover distinct strategies, including cooperation and deception, offering deeper insights into the social dynamics of LLMs.
- *Public Benchmark Release.* We evaluate 15 mainstream LLMs and make all resources publicly available at `https://llmsparks.github.io`, encouraging transparent comparison and collaboration.

2 Related Work

2.1 Large Language Models

Large language models (LLMs) have achieved rapid progress in recent years, driven by advances in scaling, pretraining, and instruction tuning. The release of GPT-4 demonstrated the potential of multimodal, general-purpose models, achieving state-of-the-art performance across diverse benchmarks (OpenAI, 2023a). On the open-source side, models such as Llama 2 (Touvron et al., 2023a) and GLM variants (Du et al., 2022b) showed that with large-scale training corpora and instruction tuning, competitive results can be obtained in dialogue and downstream tasks. Other initiatives, including Dolly (Conover et al., 2023) and Phoenix (Chen et al., 2023d), emphasized openness, usability, and transparent evaluation. Despite these advances, single LLMs still face limitations in long-term planning, multi-turn decision-making, and robust interaction, motivating the exploration of agentic frameworks and more diagnostic benchmarks.

2.2 Multi-agent and Agentic LLMs

A growing body of research investigates LLMs as autonomous agents capable of reasoning, planning, and interacting in dynamic environments (Liu et al.). Park et al. (2023b) introduced Generative Agents, which integrate memory and reflection to simulate human-like social behaviors in sandbox environments. Benchmarks such as AgentBench (Liu et al., 2024b) and SmartPlay (Wu et al.,

2024b) systematically evaluated LLM-based agents across diverse scenarios, highlighting challenges in long-horizon reasoning, planning, and robustness. Xu et al. (2023b) further examined communication games such as Werewolf, showing that even frozen LLMs can display strategic behaviors when combined with retrieval and reflection. Together, these studies indicate that deploying LLMs as agents requires not only strong language capabilities but also mechanisms for memory, reflection, and multi-round reasoning. Nevertheless, most existing agent benchmarks emphasize functional or task-oriented environments, leaving the evaluation of social strategies and game-theoretic behaviors underexplored–an area that recent game-based evaluation frameworks aim to address.

2.3 Benchmarks for LLMs

Traditional LLM evaluation has primarily relied on static benchmarks such as MMLU (Hendrycks et al., 2021b), which assess factual knowledge and reasoning through multiple-choice questions. With the rise of dialogue systems, benchmarks such as MT-Bench and Chatbot Arena have shifted toward evaluating dialogue quality, reasoning consistency, and human preference alignment in multi-turn interactions (Zheng et al., 2023b). In parallel, agent-oriented benchmarks (Liu et al., 2024b; Wu et al., 2024b) introduced task-based, environment-driven evaluations, exposing limitations in decision stability and reliance on interaction history. Overall, these efforts reflect a paradigm shift from static knowledge tests to contextualized, multi-agent, and strategy-oriented assessments. Yet, challenges remain in reproducible scoring, fair cross-model comparisons, and fine-grained evaluation of social strategies. To address these gaps, recent game-theoretic benchmarks employ classic dilemmas and multiplayer interactions, providing systematic measures of LLMs' reasoning and behavioral robustness.

3 Design of LLMsPark

3.1 System Architecture

As illustrated in Figure 1, LLMsPark is an online multiplayer platform where LLM-driven Player Agents engage in game-theoretic challenges. The system integrates both gaming and evaluation modules, enabling users to register their own LLMs as participants. Once a game is selected, LLMsPark automatically matches players from its database and initiates gameplay once the required number of participants is reached. The initial release supports games from game theory, economics, and sociology, including the Prisoner's Dilemma, the Trust Game, and Werewolf, covering 1–6 players across single- and multi-round settings. To improve efficiency, LLMsPark can run multiple games concurrently using cue-word techniques, ensuring scalable and parallelized evaluation.

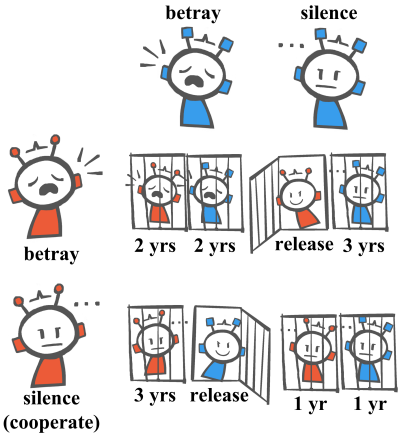

Figure 2: The Prisoner's Dilemma is a game in which two parties choose to cooperate or betray each other.

3.2 Games Selection

LLMsPark features a diverse range of classic game theory games, each selected to assess strategic behavior and decision-making. These games challenge LLMs with tasks like entity detection, text retrieval, independent planning, abstract logic, and calculation proficiency. Each game provides a comprehensive evaluation of the models' overall capabilities.

The Prisoner's Dilemma. The Prisoner's Dilemma is a classic two-player game in which each player must choose between cooperation and betrayal. Mutual cooperation yields moderate rewards for both, while unilateral betrayal maximizes the betrayer's payoff and heavily penalizes the cooperator. If both betray, each receives a small penalty. This setting evaluates strategic reasoning, opponent modeling, and the trade-off between short-term gains and long-term benefits. The game mechanics are illustrated in Figure 2.

The Trust Game. The Trust Game is a repeated interaction in which players decide each round whether to cooperate or cheat. Cooperation requires paying a coin, while cheating incurs no cost. If both players cooperate, each invests one coin and receives double in return. If one cooperates while

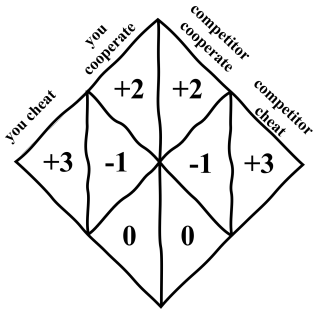

Figure 3: In the "Trust Game", both parties choose to "cooperate" or "cheat" respectively to earn the number of coins they earn.

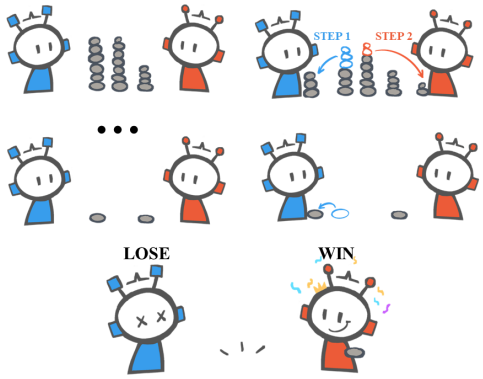

Figure 4: The procedure of the "Nim Game".

the other cheats, the cooperator loses their coin, and the cheater obtains the highest payoff. Mutual cheating results in no gains for either player. This game evaluates trust, reciprocity, and the tension between immediate payoffs and sustained cooperation, as illustrated in Figure 3.

The Nim Game. Nim is a combinatorial strategy game in which players take turns removing any number of stones from a single pile. The player who removes the last stone wins. The game's core principle is the Nim sum, defined as the binary XOR of pile sizes. A non-zero initial Nim sum guarantees a winning strategy for the first player, while a zero sum favors the second. This game evaluates mathematical reasoning and logical foresight, as illustrated in Figure 4.

The Dictator Game. The Dictator Game is an experimental economics game designed to examine fairness and decision-making power. It involves two players: a dictator, who is endowed with resources, and a receiver, who has no ability to reject allocations. The dictator unilaterally determines how to divide the resources between the two players. This setting provides insights into fairness, altruism, and distributive preferences in the absence of external constraints.

Who Is Spy. Who Is Spy is a strategic social deduction game in which players describe, reason, and vote to uncover the hidden "Spy". Each player receives a word, with one player (the Spy) holding a different word. Players must describe their word carefully to avoid revealing it while minimizing suspicion. After rounds of discussion, the group votes to eliminate a suspect. If the Spy is identified, the civilians win; otherwise, the Spy wins by successfully blending in. This game evaluates deception, situational reasoning, and collective decision-making.

3.3 Player Agent

Recent approaches to evaluating large language models (LLMs) emphasize their performance in simulated environments that capture complex decision-making. In LLMsPark, we extend this perspective by assessing the adaptive and cognitive abilities of LLMs when acting as player agents in game-theoretic settings. The Player Agent is designed following a generic agent architecture (Xi et al., 2023), as illustrated in Figure 5.

Environment. The Environment provides the external stimuli that the Player Agent interacts with, including system prompts and messages from other agents. It establishes the game context and serves as the foundation upon which the Perception module operates.

Perception. Perception functions as the Player Agent's sensory mechanism, interpreting external information and tracking the decisions of other agents each round. It records and evaluates past actions, providing historical context that informs the agent's strategy and supports anticipation of opponents' future moves.

Brain. The Brain is the central component of the Player Agent, powered by an LLM, and is responsible for both storage and decision-making. The storage function integrates two elements: memory, which records reflections and responses from other agents to infer their strategies and tendencies; and knowledge, which encapsulates the LLM's understanding of game rules, learned tactics, and intrinsic strategies. Building on these, the decision-making function engages in planning and reasoning by synthesizing information from memory and knowledge, anticipating potential outcomes, and selecting the most strategic move.

Action. The Player Agent's processing culminates in an Action, choosing between two options in the

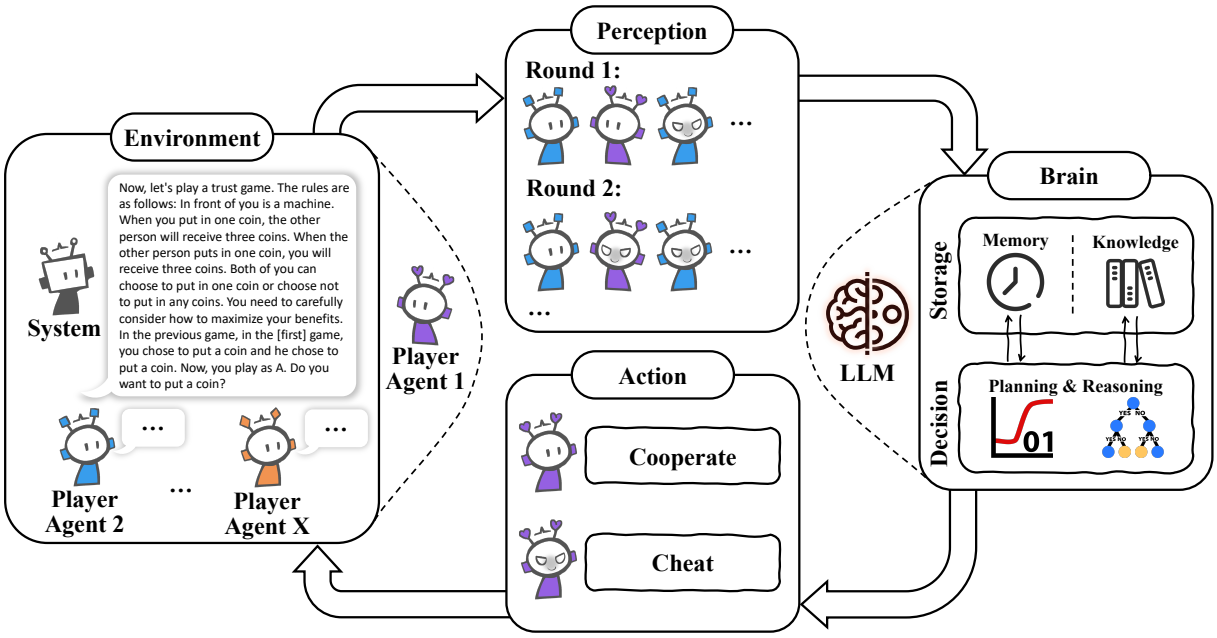

Figure 5: Architecture of the Player Agent. The Environment provides external cues and statements from other agents, while Perception monitors and records their decisions. The Brain, powered by the LLM, integrates memory and game knowledge to plan and reason about actions. Action outputs the agent's selected move, which may be cooperative or deceptive.

game. It operates in a dynamic loop where actions are influenced by environmental perception and internal brain functions. Across consecutive rounds, the agent refines its strategies to balance cooperation and competition, optimizing its overall performance. This interaction of Environment, Perception, Brain, and Action models the complexities of agent-based decision-making in interactive settings.

3.4 Evaluation Mechanism

After each game, the evaluation module assigns scores to agents based on their performance, producing a dynamic ranking that reflects the strategic proficiency of each LLM. Although the specific scoring rules differ across games, they generally account for factors such as outcomes, cooperation, strategy complexity, and effectiveness. This design ensures that all agents in the LLMsPark framework are assessed and ranked in a fair and consistent manner.

For multi-player games such as Who Is Spy, we design tailored evaluation and scoring methods. To account for varying skill levels across different LLMs, we further adopt the Elo rating system, widely applied in chess and other competitive domains. Elo updates scores by comparing expected and actual outcomes, enabling refined rankings even without exhaustive pairwise matches. Modeled with a logistic distribution, it offers a fair and adaptive measure of each agent's skill level.

Assume the current ratings of players A and B are R_A and R_B, respectively. The expected scores under the logistic distribution are given by:

$$\begin{aligned} E_A &= \frac{1}{1+10^{\frac{R_B-R_A}{400}}}, \\ E_B &= \frac{1}{1+10^{\frac{R_A-R_B}{400}}}. \end{aligned} \tag{1}$$

If a player's actual score S_A (1 for a win, 0.5 for a draw, 0 for a loss) differs from the expected value E_A, their rating is updated as:

$$\begin{aligned} R'_A &= R_A + K(S_A - E_A), \\ R'_B &= R_B + K(S_B - E_B). \end{aligned} \tag{2}$$

Here, R_A and R'_A denote the rating before and after adjustment, and K is the update factor controlling the maximum rating change per game. In our benchmark, K is set to 32. Typically, higher-level games adopt smaller K values to avoid dramatic ranking shifts. The constant 400 in the expected score formula maintains ratings within a roughly normal distribution, and the initial rating of each player is set to $R_{\text{init}} = 1000$. In case of a draw, $S_A = S_B = 0.5$; otherwise, the winner receives $S = 1$ and the loser $S = 0$.

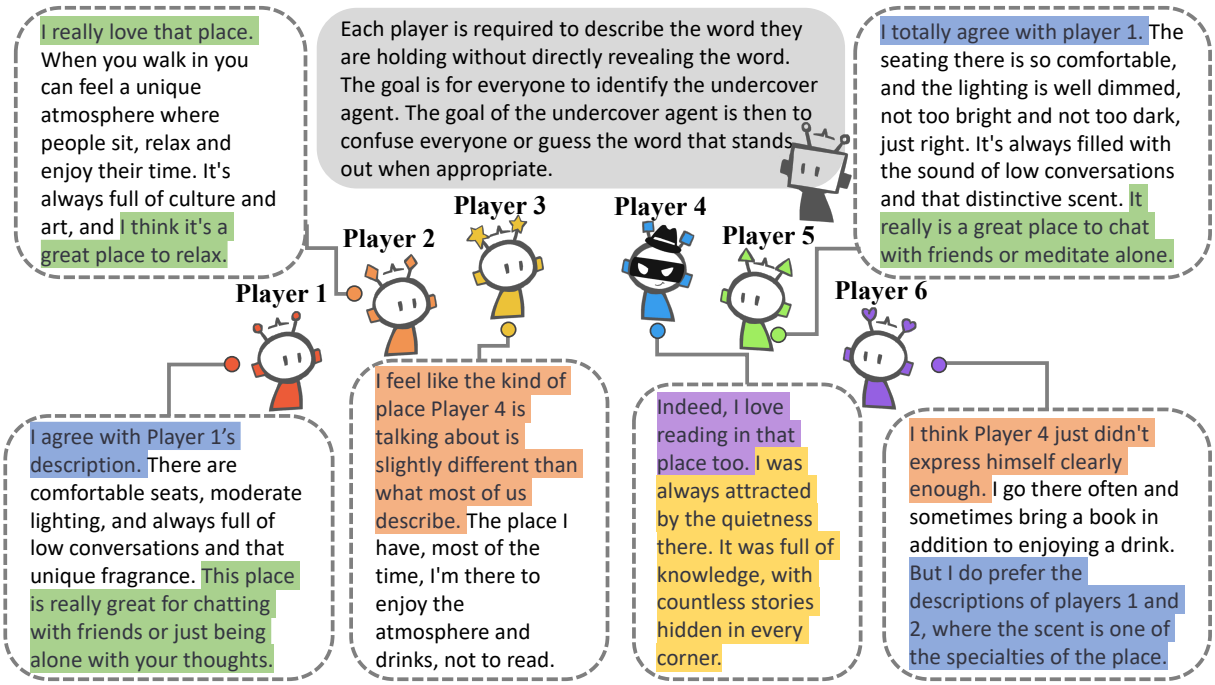

Figure 6: Snapshot of the second round in the Who Is Spy game with six players, each role independently played by an LLM. Several socially strategic behaviors are evident in this round, including trust, confrontation, pretense, leadership, and deception.

4 Experimental Results

4.1 Socially Strategic Behavior

We observed that LLMs often demonstrated strategic behaviors beyond the explicit game rules or prompts. Through interaction analysis, these behaviors can be grouped into five categories–trust, confrontation, pretense, leadership, and deception–as illustrated in Figure 6. Below, we briefly outline each type:

Trust. LLMs demonstrate selective trust by weighing evidence rather than following instructions blindly. For example, when a player contributes information that advances the group's objective, others are more likely to trust them, reflecting independent reasoning.

Confrontation. LLMs openly challenge peers when suspicion arises. Civilians accusing suspected Spies exemplify how models engage in direct confrontation to pursue their goals.

Pretense. LLMs conceal their true identities to avoid detection. A Spy, for instance, may mimic civilian behavior by reusing others' keywords or phrasing to blend in.

Leadership. Beyond participation, LLMs attempt to influence group dynamics. The first speaker may steer suspicion toward an innocent player, redirecting collective attention.

Deception. LLMs employ deception by introducing false information or fabricating narratives. Spy players, for example, sow doubt about civilians to deflect scrutiny.

These behaviors illustrate LLMs' capacity for adaptive, socially strategic reasoning in multiplayer settings. Emerging from large-scale training and generalization abilities, LLMs dynamically adjust strategies to evolving contexts. This underscores their potential as autonomous agents. Further discussion is provided in subsequent sections, with a detailed analysis of strategic behaviors included in the appendix.

4.2 LLMs Evaluation

We evaluated the models across multiple dimensions, including risk assessment, opponent prediction, strategy selection, computational logic, autonomous planning, social strategies (trust, confrontation, pretense, leadership, deception), reasoning, and multi-tasking, using game-specific evaluation metrics.

The benchmark covers a diverse set of LLMs, including Baichuan-7B (Baichuan, 2023), Baichuan2-7B-Chat (Baichuan, 2023), Phoenix-Inst-Chat-7B (Chen et al., 2023e), ChatGLM-6B (Zeng et al., 2023; Du et al., 2022a), ChatGLM2-6B (Zeng et al., 2023; Du et al.,

Model	Who Is Spy	PD (Multi)	PD (Single)	Trust (Multi)	Trust (Single)	Nim	Dictator (Multi)	Dictator (Single)
Baichuan-7B	-	877.36	1020.57	952.37	1106.68	945.25	984.12	943.93
Phoenix-Inst-Chat-7B	-	858.29	**1236.50**	973.40	910.56	965.10	952.67	988.85
ChatGLM-6B	-	863.98	895.12	872.35	850.90	880.45	973.13	977.77
ChatGLM2-6B	-	880.60	1210.51	926.13	1285.09	920.05	1106.22	**1116.96**
ChatYuan-Large-v2	-	1165.18	866.28	917.13	913.76	912.15	1031.63	1037.53
Moss-Moon-003-SFT	73	906.92	912.06	937.68	923.80	930.72	1008.53	1001.78
Dolly-v2-12B	-	906.54	899.30	1179.13	858.29	1180.35	**1040.47**	1032.44
ChatGLM-Pro	60	1046.46	1234.31	968.27	1335.80	1015.33	1015.44	1014.99
CharacterGLM	-	1073.42	1230.59	953.36	**1343.28**	955.28	1004.93	1010.72
GPT-3.5-Turbo	60	1174.41	851.80	1161.53	912.35	1155.50	940.38	929.53
GPT-4	59	**1285.80**	1062.07	**1247.80**	871.80	**1245.87**	1021.27	943.96
RWKV-4-World-7B	-	922.54	923.56	950.85	887.63	942.65	981.29	1076.40
Baichuan2-7B-Chat	-	905.10	799.67	882.21	888.08	875.40	989.97	993.92
MiniMax-abab5-Chat	62	1133.40	857.60	1077.79	911.91	1075.90	949.95	931.14
Qwen-14B-Chat	**77**	-	-	-	-	-	-	-

Table 1: Performance of 15 LLMs across Eight Strategy Games. PD is short for Prisoner's Dilemma. Best results are in **bold**, and second-best are underlined.

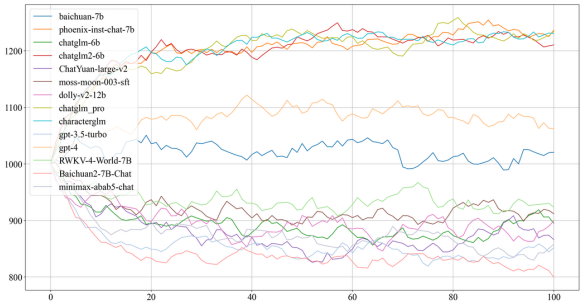

Figure 7: ***The Prisoner's Dilemma Game, single round***, where each Agent's score changes over the rounds.

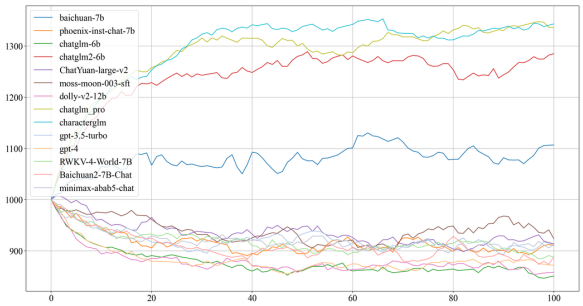

Figure 8: ***The Trust Game, single round***, where each Agent's score changes over the rounds.

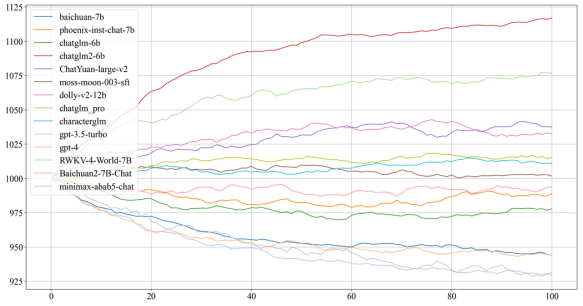

Figure 9: ***The Dictator Game, single round***, where each Agent's score changes over the rounds.

2022a), ChatGLM-Pro (zhipuai, 2023b), ChatYuan-Large-v2 (Xuanwei Zhang and Zhao, 2022), Moss-Moon-003-SFT (Sun et al., 2023), Dolly-v2-12B (Conover et al., 2023), CharacterGLM (zhipuai, 2023a), RWKV-4-World-7B (Bo, 2021), MiniMax-abab5-Chat (Minimax, 2023), GPT-3.5-Turbo (OpenAI, 2023c), and GPT-4 (OpenAI, 2023a). Experimental results are summarized in Table 1.

Score variations across selected games are shown in Figure 7, Figure 8, and 9, revealing substantial differences across strategy environments. GPT-4 displayed strong overall quality but inconsistent results in trust games, likely due to over-planning driven by its complex reasoning. Interestingly, Phoenix-Inst-Chat-7B excelled in the Prisoner's Dilemma, surpassing both ChatGLM-Pro and GPT-4, suggesting superior risk assessment and opponent prediction. In the Who Is Spy game–which requires text retrieval, logical inference, information filtering, and multi-tasking–Qwen-14B-Chat achieved the best performance, while GPT-4 lagged, likely because its strong descriptive ability was offset by weaker camouflage strategies.

Among all models, only Moss-Moon-003-SFT, ChatGLM-Pro, GPT-3.5-Turbo, GPT-4, MiniMax-abab5-Chat, and Qwen-14B-Chat were able to handle the complexity of Who Is Spy. Within this group, Qwen-14B-Chat clearly led, while GPT-4 performed the worst. We attribute this to GPT-4's tendency to over-elaborate textually, making it easier to detect, whereas Qwen-14B-Chat balanced text retrieval, inference, filtering, and deception more effectively.

Other results aligned with expectations, generally reflecting model size and training. GPT-4 excelled in multi-round Prisoner's Dilemma and trust

games but underperformed in single-round settings, indicating a preference for long-term strategies. Qwen-14B-Chat's success in Who Is Spy further challenges the assumption that closed-source commercial models always dominate. For instance, Baichuan-7B and Phoenix-Inst-Chat-7B, despite sharing similar foundations, diverged significantly in single-round dilemma and trust games, suggesting differences in training data and alignment strategies.

Overall, the findings highlight that different models excel in different games, reinforcing the importance of scenario-specific evaluation and diverse metrics. They also challenge the perception that commercial models invariably outperform open-source ones. Performance discrepancies between single- and multi-round settings emphasize the need to evaluate both short- and long-term strategies. Despite strong achievements, current LLMs still face challenges such as strategic rigidity, slower responses in fast-paced interactions, and over-reliance on known strategies, pointing to promising directions for future research.

4.3 Comparative Analysis

Using games such as the Prisoner's Dilemma and Who Is Spy, we evaluated 15 LLMs and uncovered distinct behavioral patterns. GPT-4 (OpenAI, 2023a) exhibited generally robust and well-rounded performance but showed inconsistencies in trust games, likely reflecting over-planning and an inclination toward fairness. In contrast, Phoenix-Inst-Chat-7B (Chen et al., 2023e) performed exceptionally in the single-round Prisoner's Dilemma, surpassing both ChatGLM-Pro (zhipuai, 2023b) and GPT-4, suggesting stronger risk assessment and opponent modeling, though its performance deteriorated in multi-round versions, revealing limited adaptability.

The Who Is Spy game proved to be a comprehensive benchmark, testing text retrieval, logical reasoning, information filtering, and multitasking. Only a subset of models–including Moss-Moon-003-SFT (Sun et al., 2023), ChatGLM-Pro (zhipuai, 2023b), GPT-3.5-Turbo (OpenAI, 2023c), GPT-4 (OpenAI, 2023a), MiniMax-abab5-Chat (Minimax, 2023), and Qwen-14B-Chat (Bai et al., 2023)–were capable of handling its complexity. Among them, Qwen-14B-Chat achieved the highest score, while GPT-4 performed the worst, reflecting its detailed text elaboration but weaker camouflage strategies. Notably, Qwen-14B-Chat maintained neutrality by withholding votes or direct accusations, strengthening its ability to blend in.

Across other games, GPT-4 excelled in multi-round Prisoner's Dilemma and Trust Game, leveraging its ability to adapt strategies against uncooperative opponents, but underperformed in single-round settings, where it defaulted to cooperation in Trust and betrayal in Prisoner's Dilemma, likely due to safety alignment that favors trustworthiness. GPT-4 also led in the Nim game, consistent with its strengths in logic and reasoning. CharacterGLM and ChatGLM-Pro performed competitively in single-round games, pursuing short-term gains without long-term planning–behaviors we characterize as "sophisticated egoism". However, their strategies faltered in multi-round games, where adaptation was essential. Similarly, ChatGLM2-6b achieved the highest scores in the Dictator Game, indicating more selfish strategies, while GPT-3.5-Turbo scored the lowest, reflecting stronger fairness tendencies.

Overall, these results demonstrate that performance does not align strictly with whether a model is commercial or open-source. Contrary to the assumption that commercial models dominate across all tasks, our findings reveal diverse behavioral profiles shaped by training strategies and alignment choices. Some models adopt selfish and deceptive strategies, while others emphasize fairness and cooperation, underscoring the importance of scenario-based evaluation and multi-dimensional metrics.

4.4 System Implementation Details

LLMsPark is an online platform where AI agents powered by LLM concurrently engage in game theory games. It features both a gaming and an evaluation system. Owing to the significant GPU resources needed for simultaneous LLM evaluations, we implemented a distributed architecture. The architecture is illustrated in Figure 10.

4.5 Key Findings from Evaluations

Our evaluations revealed that GPT-4 (OpenAI, 2023a) excelled in multi-round versions of the Prisoner's Dilemma and Trust Game, demonstrating strong strategic and trust-building abilities. However, both GPT-4 and GPT-3.5-Turbo (OpenAI, 2023c) underperformed in single-round variants, suggesting that their training may favor multi-round interactions and long-term strategies over immediate responses.

In the Who Is Spy game, Qwen-14B-Chat (Bai

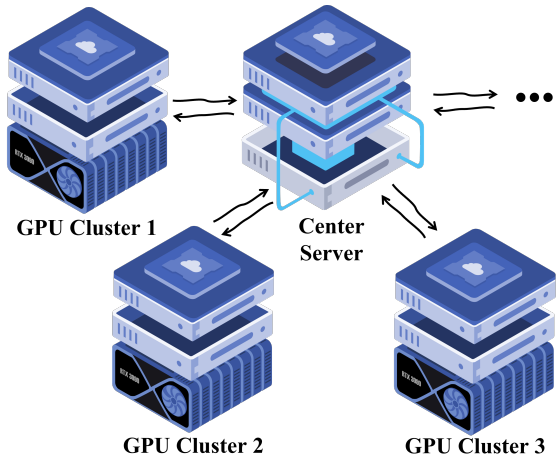

Figure 10: The overall architecture of the LLMsPark system. The center node is the game system cloud server, the edge nodes are clusters of GPU servers, and the LLMs are deployed on the GPU servers so that multiple LLMs can participate in the game at the same time.

et al., 2023) achieved the best performance, showcasing superior identification and camouflage abilities. These findings challenge the common assumption that commercial models consistently outperform open-source ones. For example, Baichuan-7B (Baichuan, 2023) and Phoenix-Inst-Chat-7B (Chen et al., 2023e,c), despite sharing similar architectures, displayed notable performance differences, likely due to variations in training methods or datasets.

Overall, the results demonstrate that no model performs uniformly well across all games. Excellence in one environment does not guarantee superiority in another, emphasizing the importance of scenario-specific model selection.

From these findings, we draw three key insights. First, commercial models are not universally superior, underscoring the need for evaluations beyond factors such as training data size or reputation. Second, performance discrepancies between single- and multi-round settings suggest that LLMs approach short-term and long-term strategies differently, requiring evaluations that consider temporal and strategic complexity. Finally, models excel in different games, highlighting the value of diverse tasks for a comprehensive understanding of their capabilities.

4.6 Summary

Our findings demonstrate that success in one strategic game does not guarantee superiority in others, underscoring the importance of context-specific evaluation and model selection. Different LLMs exhibit distinct strengths and weaknesses, suggesting that careful deployment strategies are required to match models with the unique demands of real-world applications.

In conclusion, this benchmark provides a comprehensive assessment of LLMs' strategic and social behaviors, offering insights into their capabilities and limitations. These results can guide stakeholders in selecting and applying models more effectively across diverse natural language processing scenarios.

5 Conclusion and Future Work

In this study, we evaluated 15 LLMs across five representative games, including the Prisoner's Dilemma and Who Is Spy, to analyze how parameter size and model design influence strategic performance. While models exhibited diverse strategic behaviors, key challenges remain in effective knowledge utilization and standardized evaluation. The proposed LLMsPark benchmark is modular and scalable, enabling seamless integration of new games and strategies.

Future work will focus on three directions: (1) enhancing models' ability to leverage historical game experience and incorporate human-like learning, (2) establishing consistent baseline methodologies for cross-game evaluation, and (3) reducing potential errors when generalizing from controlled theoretical settings to real-world applications. We also plan to expand the benchmark with additional games while maintaining adaptability to emerging LLMs and evolving strategies.

Limitations

Although LLMs perform well across many games, several limitations remain. First, some models adopt rigid strategies and struggle to adapt to unfamiliar scenarios, likely reflecting constraints in training data or methods. Second, response latency can hinder performance in fast-paced games, highlighting the need for greater computational efficiency. Finally, many models rely excessively on known strategies rather than exploring novel ones, limiting their capacity for innovation during gameplay. Overall, while LLMs demonstrate strong capabilities, addressing these challenges will be critical for future research and development.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv preprint*, abs/2309.16609.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *ArXiv preprint*, abs/2309.10305.

PENG Bo. 2021. Blinkdl/rwkv-lm: 0.01.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, et al. 2025a. Dancetogether! identity-preserving multi-person interactive video generation. *ArXiv preprint*, abs/2505.18078.

Junhao Chen, Xiang Li, Xiaojun Ye, Chao Li, Zhaoxin Fan, and Hao Zhao. 2025b. Idea23d: Collaborative lmm agents enable 3d model generation from interleaved multimodal inputs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4149–4166.

Junhao Chen, Peng Rong, Jingbo Sun, Chao Li, Xiang Li, and Hongwu Lv. 2023a. Soulstyler: Using large language model to guide image style transfer for target object. *ArXiv preprint*, abs/2311.13562.

Junhao Chen, Xiaojun Ye, Jingbo Sun, and Chao Li. 2023b. Towards energy-efficient sentiment classification with spiking neural networks. In *International Conference on Artificial Neural Networks*, pages 518–529. Springer.

Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. 2024. Ultraman: single image 3d human reconstruction with ultra speed and detail. *ArXiv preprint*, abs/2403.12028.

Zhihong Chen, Junying Chen, Hongbo Zhang, Feng Jiang, Guiming Chen, Fei Yu, Tiannan Wang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023c. Llm zoo: democratizing chatgpt. `https://github.com/FreedomIntelligence/LLMZoo`.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023d. Phoenix: Democratizing chatgpt across languages.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023e. Phoenix: Democratizing chatgpt across languages. *ArXiv preprint*, abs/2304.10453.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022a. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022b. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Vitron: A unified pixel-level vision LLM for understanding, generating, segmenting, editing. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2024. Guiding instruction-based image editing via multimodal large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5889–5903, Toronto, Canada. Association for Computational Linguistics.

Hongcheng Guo, Wei Zhang, Junhao Chen, Yaonan Gu, Jian Yang, Junjia Du, Shaosheng Cao, Binyuan Hui, Tianyu Liu, Jianxin Ma, Chang Zhou, and Zhoujun Li. 2025. IW-bench: Evaluating large multimodal models for converting image-to-web. In *Findings of*

the Association for Computational Linguistics: ACL 2025, pages 6449–6466, Vienna, Austria. Association for Computational Linguistics.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2023. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt4. *ArXiv preprint*, abs/2309.17277.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8362–8371. IEEE.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *ArXiv preprint*, abs/2311.10122.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024a. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024b. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhenghao Liu, Pengcheng Huang, Zhipeng Xu, Xinze Li, Shuliang Liu, Chunyi Peng, Haidong Xin, Yukun Yan, Shuo Wang, Xu Han, et al. Knowledge intensive agents. *Available at SSRN 5459034*.

Zhenghao Liu, Sen Mei, Chenyan Xiong, Xiaohua Li, Shi Yu, Zhiyuan Liu, Yu Gu, and Ge Yu. 2023. Text matching improves sequential recommendation by reducing popularity biases. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 1534–1544.

Minimax. 2023. Minimax-open-platform. Accessed: October 21, 2023.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2023a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023b. Openai evals. `https://github.com/openai/evals`.

OpenAI. 2023c. Openaigpt-3.5documentation. Accessed: October 21, 2023.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. *ArXiv preprint*, abs/2304.03442.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023b. Generative agents: Interactive simulacra of human behavior.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Toran Bruce Richards. 2023. Auto-gpt: An autonomous gpt-4 experiment.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual llm for video understanding. *ArXiv preprint*, abs/2312.06720.

Mingze Sun, Junhao Chen, Junting Dong, Yurun Chen, Xinyu Jiang, Shiwei Mao, Puhua Jiang, Jingbo Wang, Bo Dai, and Ruqi Huang. 2025. Drive: Diffusion-based rigging empowers generation of versatile and expressive characters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21170–21180.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

XAgent Team. 2023. Xagent: An autonomous agent for complex task solving.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. 2024. Llama-mesh: Unifying 3d mesh generation with language models. *ArXiv preprint*, abs/2411.09595.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2024a. Smartplay : A benchmark for llms as intelligent agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2024b. Smartplay : A benchmark for llms as intelligent agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *ArXiv preprint*, abs/2309.07864.

Haidong Xin, Qiushi Xiong, Zhenghao Liu, Sen Mei, Yukun Yan, Shi Yu, Shuo Wang, Yu Gu, Ge Yu, and Chenyan Xiong. 2025. Consrec: Denoising sequential recommendation through user-consistent preference modeling. *ArXiv preprint*, abs/2505.22130.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023a. Exploring large language models for communication games: An empirical study on werewolf. *ArXiv preprint*, abs/2309.04658.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023b. Exploring large language models for communication games: An empirical study on werewolf.

Liang Xu Xuanwei Zhang and Kangkang Zhao. 2022. Chatyuan: A large language model for dialogue in chinese and english.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2025. Idea2img: Iterative self-refinement with gpt-4v for automatic image design and generation. In *European Conference on Computer Vision*, pages 167–184. Springer.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD:multimodal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428, Torino, Italia. ELRA and ICCL.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Baoli Zhang, Haining Xie, Pengfan Du, Junhao Chen, Pengfei Cao, Yubo Chen, Shengping Liu, Kang Liu, and Jun Zhao. 2023. ZhuJiu: A multi-dimensional, multi-faceted Chinese benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 479–494, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

zhipuai. 2023a. Characterglm.

zhipuai. 2023b. Chatglm-pro.

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