Solving Repeated Games with Large Language Model

Naming Liu
Shanghai Jiao Tong University
Shanghai, China
liunaming@sjtu.edu.cn

Youzhi Zhang
CAIR, Hong Kong Institute of Science & Innovation, CAS
HKSAR, China
youzhi.zhang@cair-cas.org.hk

Ying Wen
Shanghai Jiao Tong University
Shanghai, China
ying.wen@sjtu.edu.cn

ABSTRACT

Sequential reasoning is a fundamental yet challenging capability for intelligent agents, requiring Large Language Model (LLM) agents to anticipate others' beliefs and dynamically adapt their strategies in repeated multi-agent interactions. However, existing LLM approaches often lack a reasoning framework that jointly supports opponent modeling and effective adaptation, limiting their robustness in dynamic and complex games. To address this gap, we introduce the Reflective Hypothetical Mind (RHM) framework, inspired by the Hypothetical Mind architecture [6]. RHM maintains multiple hypothetical minds to represent evolving opponent strategies and, crucially, integrates an explicit adaptation module that translates these belief updates into adaptive decision-making. This design enables LLM agents not only to model changing behaviors but also to respond with strategically effective adaptations. Empirical results across diverse repeated games demonstrate that RHM outperforms baseline LLMs by achieving stronger coordination and adaptability across diverse repeated games, highlighting the effectiveness of unifying opponent modeling with explicit policy adaptation.

KEYWORDS

Large Language Model, Repeated Games

ACM Reference Format:
Naming Liu, Youzhi Zhang, and Ying Wen. 2026. Solving Repeated Games with Large Language Model. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/IRTM9736

1 INTRODUCTION

In recent years, the rapid advancement of Large Language Models (LLMs) has marked a significant milestone in the development of artificial intelligence. State-of-the-art models such as GPT-4 have demonstrated remarkable capabilities across a wide range of tasks, including machine translation [14], text generation [16], information retrieval [38], and code repair [29]. These technologies have not only attracted widespread academic attention but have also found impactful applications in domains such as education [3], law [9], product design [19], and healthcare [15], profoundly reshaping how people live and work [7]. Yet as LLM capabilities continue to grow, a critical open question arises: how can we rigorously evaluate their capacity for **sequential reasoning** in non-stationary multi-agent environments?

Sequential reasoning refers to decision-making in long-horizon, multi-agent environments, where agents must adapt to the actions of non-stationary players, update beliefs about others' evolving strategies, and ensure consistency with their own long-term objectives [35]. Such reasoning poses unique challenges for LLMs and LLM-based agents, which must operate under uncertainty and dynamic strategic interactions rather than static one-shot tasks. Sequential reasoning is therefore essential for building intelligent agents and underpins diverse real-world applications, including dynamic investment, adaptive business strategy [37], multi-round negotiation [13], and long-term policy-making [20].

Effective sequential reasoning relies on understanding others' perspectives and anticipating their strategies. Despite its importance, current progress on equipping LLMs with sequential reasoning remains limited. Existing approaches largely rely on static prompting [1, 36**?**], where models are instructed within the prompt to account for others' beliefs and decisions during their own reasoning. While such methods can elicit short-term perspective-taking and approximate higher-order beliefs, they fall short of enabling LLMs to maintain persistent belief states, adapt to non-stationary opponents, and engage in genuine strategy shifts. Consequently, they lack the flexibility required for dynamic sequential reasoning.

The Hypothetical Mind (HM) framework [6] takes a step forward by generating and testing hypotheses about opponents' changing strategies, enabling dynamic belief-based prediction. This allows agents to anticipate short-term opponent behaviors more effectively than static prompting methods. However, in complex games, opponent modeling alone does not equate to policy adaptation. While HM can forecast what an opponent might do next, it lacks mechanisms for systematically revising its own strategies in response, or for actively disrupting detrimental strategic cycles. A typical failure case occurs in the Iterated Prisoner's Dilemma when the opponent adopts a *Tit-for-Tat* strategy. As illustrated in Figure 1, the opponent initially chooses *Cooperate*, while the Hypothetical Mind (HM) agent, driven by short-term self-interest, responds with the dominated action *Defect* to pursue the immediate payoff $(10, 0)$. Subsequently, the opponent switches to *Defect*—mirroring the previous round—while HM continues to defect, resulting in the lower payoff $(2, 2)$. This cycle arises because, from a one-shot self-interested perspective, defection strictly dominates cooperation: regardless of the opponent's move, it maximizes the immediate reward. Once triggered, both agents become locked in perpetual mutual defection, severely undermining their long-term payoffs. In contrast, sustained mutual cooperation would yield higher long-term rewards, but achieving it requires agents to reason sequentially

Corresponding author: Youzhi Zhang, Ying Wen.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/IRTM9736

	Cooperate	Defect
Cooperate	(6, 6)	(0, 10)
Defect	(10, 0)	(2, 2)

Figure 1: The Hypothetical Mind (HM) agent exhibits detrimental strategic cycles in the Iterated Prisoner's Dilemma due to its short-term self-interest. Assume the opponent adopts a *Tit-for-Tat* strategy and initially chooses *Cooperate*; HM responds with the dominated action *Defect* to pursue the immediate payoff $(10, 0)$. Subsequently, the opponent switches to *Defect*—mirroring the previous action—while HM continues to select the same dominated action, leading to a lower outcome $(2, 2)$. Without actively choosing *Cooperate*, both players eventually fall into a mutual-defection cycle, which severely undermines their long-term payoffs.

beyond myopic incentives and adapt their strategies to preserve cooperation.

To address these challenges, we propose the *Reflective Hypothetical Mind* (RHM) framework, which equips LLM agents with structured sequential reasoning by combining opponent modeling with adaptive decision-making. RHM consists of three key components: (1) a **hypothetical mind** module that generates and maintains beliefs about dynamic opponent strategies, (2) a **self-refinement** module that revises decision-making rules based on past outcomes to prevent compounding early errors, and (3) a **policy adaptation** module that translates reflective insights into concrete behavioral adjustments. Through this integration, RHM enables agents not only to anticipate opponent strategies but also to adapt their own policies accordingly, allowing them to escape harmful cycles, stabilize cooperative behavior, and sustain long-term advantages in dynamic multi-agent environments.

We evaluate RHM across canonical repeated games with increasing complexity. In simple 2×2 self-interested games, LLMs alone already approximate cooperative play and achieve strong payoffs. In contrast, cooperative settings such as the repeated Battle of Sexes require explicit opponent modeling through theory of mind to sustain coordination. Finally, in more complex environments like iterated Rock-Paper-Scissors and iterated Colonel Blotto, opponent modeling alone is insufficient, and explicit policy adaptation becomes essential for robust performance. These results highlight that RHM's integration of hypothetical mind and adaptation enables LLMs to succeed across diverse strategic scenarios where baselines struggle.

2 RELATED WORK

LLM Reasoning Paradigms. Large Language Models (LLMs) exhibit impressive emergent reasoning abilities, yet much of this capability is highly dependent on prompting techniques that scaffold the model's thought process. *Chain-of-Thought* (CoT) [31] introduced stepwise reasoning to improve problem solving, while *Tree-of-Thought* (ToT) [32] extends this idea by branching and evaluating multiple reasoning paths in parallel, closely related to our use of parallel hypothetical beliefs for opponent modeling. K-Level Reasoning [36] explicitly implements recursive higher-order belief reasoning and shows that GPT-4 can handle two- to three-level strategic reasoning in negotiation and competitive settings. More recent work introduces self-reflection and critique mechanisms, such as *Reflexion* [27], which allow models to revise past outputs using feedback signals from the environment or an external verifier. Other studies, including STaR [33] and symbolic interpreter-guided refinement [25], explore iterative hypothesis generation and correction, enabling LLMs to refine reasoning based on task-specific feedback. Despite these advances, current methods remain largely **prompt-dependent and episodic**: they lack persistent internal memory, explicit long-term belief tracking, and structured mechanisms for systematically updating strategies as environments evolve. Our approach differs by embedding reflective reasoning into a *policy adaptation loop*, bridging the gap between single-turn reasoning and long-term adaptive decision-making.

LLM-Based Agents and Planning. Another growing research direction investigates how to transform LLMs into autonomous agents that operate across complex domains. LLM-based agents have been used as high-level planners in virtual or embodied environments by leveraging their extensive background knowledge. For instance, *Voyager* [30] autonomously acquires and composes skills in Minecraft via a dynamically built skill library, enabling it to solve progressively harder tasks. *SAMA* [21] and similar frameworks integrate LLM planning with goal-conditioned reinforcement learning, decomposing complex tasks into subgoals to guide low-level policies. Interactive social simulators such as Generative Agents [24] demonstrate LLMs' capacity to maintain long-term memories, set goals, and engage in open-ended social interactions. In cooperative multi-agent domains, *ProAgent* [34] improves zero-shot coordination in Overcooked by inferring teammates' intentions from state observations; other works [6] design cognitive modules to enhance collaboration by modeling other agents' goals and preferences. These systems showcase the potential of LLMs as central cognitive controllers, but they largely focus on **collaboration or static planning**, leaving open the challenge of robust strategic adaptation in *dynamic and adversarial* multi-agent games, where environments and opponents evolve over time.

Decision Making in Normal-Form Games. A complementary line of work explores how LLMs behave in formal game-theoretic scenarios. Early experiments revealed that simple prompting can induce cooperative or human-like play in the Prisoner's Dilemma and Ultimatum Game [10], yet outcomes vary dramatically with subtle framing changes [23], highlighting sensitivity to surface-level textual cues. Larger-scale evaluations across two-player normal-form games [12] report frequent failure to converge to Nash equilibria and systematic biases in mixed-strategy reasoning. Such findings suggest that while LLMs can imitate strategic responses, their reasoning is often fragile and shaped more by prompt wording than by underlying incentive structures. Moving beyond static one-shot games thus requires mechanisms for **temporal consistency and adaptive response**—capabilities that cannot be achieved through naive prompting alone.

Decision Making in Repeated Games. Repeated games serve as a natural testbed for sequential reasoning, belief updating, and long-term strategy formation. Recent studies show that while LLMs perform competitively when incentives are simple and aligned, they struggle in coordination and adversarial settings: once cooperation breaks (e.g., in Iterated Prisoner's Dilemma), models rarely

re-establish trust [1]. Even with advanced prompting, LLMs often oscillate between over-punishment and short-term payoff maximization [8, 22], leading to brittle dynamics. Approaches like CoT, ToT, and Reflexion improve local reasoning but fail to maintain **persistent opponent models or evolving policies** over long horizons.

These limitations motivate our **Reflective Hypothetical Mind (RHM)** framework, which unifies *structured opponent modeling*, *hypothesis evaluation*, and a *policy adaptation loop* to achieve both robust reasoning and long-term strategic adjustment. Unlike prior prompt-only or single-turn reasoning systems, RHM maintains evolving beliefs, reflects on past interactions, and updates its action policy accordingly, enabling scalable and stable performance in coordination games, cyclic environments, and combinatorially large action spaces.

3 METHOD

This section presents our methodology for enabling large language model (LLM) agents to perform robust sequential reasoning in repeated multi-agent games. We first formalize repeated games and define the interaction setting. Then, we introduce the Reflective Hypothetical Mind (RHM) framework, which integrates three key components—hypothetical opponent modeling, self-reflection, and policy adaptation—allowing agents to iteratively refine strategies, predict opponent behavior, and minimize regret over repeated interactions.

3.1 Repeated Games

A finite normal-form game is defined by the tuple $\mathcal{G} = \langle N, \{\mathcal{A}_i\}_{i=1}^N, \{r_i\}_{i=1}^N \rangle$, where N is the set of players, $\mathcal{A}_i$ denotes the finite action set for player i, and $r_i : \mathcal{A}_1 \times \cdots \times \mathcal{A}_N \to \mathbb{R}$ specifies the payoff function of player i. Each player i chooses a (possibly stochastic) policy $\pi_i \in \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ is the probability simplex over $\mathcal{A}_i$. A joint action is denoted $\boldsymbol{a} = (a_1, \ldots, a_N)$. Let $-i$ denote the set of all players other than player i. A policy profile $(\pi_1^*, \ldots, \pi_N^*)$ constitutes a **Nash equilibrium** if no player can unilaterally improve their expected payoff, that is, for all $i \in N$, $\mathbb{E}_{\boldsymbol{a} \sim \pi_i^* \times \pi_{-i}^*}[r_i(\boldsymbol{a})] \geq \mathbb{E}_{\boldsymbol{a} \sim \pi_i' \times \pi_{-i}^*}[r_i(\boldsymbol{a})], \quad \forall \pi_i' \in \Delta(\mathcal{A}_i)$.

In a repeated game, the same stage game $\mathcal{G}$ is played over rounds $t = 1, 2, \ldots$ (possibly infinitely). At each round t, each player i observes the complete history of previous joint actions $h^t = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{t-1})$ and then selects an action $a_{i,t} \sim \pi_{i,t}(\cdot | h^t)$ conditioned on this history. This produces joint action $\boldsymbol{a}_t = (a_{1,t}, \ldots, a_{N,t})$ and immediate payoffs $r_i(\boldsymbol{a}_t)$. The total discounted payoff for player i is given by

$$U_i = \sum_{t=1}^{\infty} \delta^{t-1} r_i(\boldsymbol{a}_t), \quad \delta \in (0, 1),$$

where δ is the discount factor controlling the trade-off between immediate and future payoffs. This structure naturally enables contingent behaviors like cooperation, punishment, and opponent modeling, which are infeasible in one-shot normal form games.

Let the repeated game be represented as $\mathcal{G}^T = \langle N, \{\mathcal{A}_i\}_{i=1}^N, \{u_i\}_{i=1}^N \rangle^T$, where agents repeatedly select actions $a_{i,t} \sim \pi_i(h^t)$, with history $h^t = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{t-1})$. Each LLM agent is instantiated as a prompted policy function $\pi_i : H^t \to \mathcal{A}_i$, updated iteratively via reflection. The following subsections elaborate on the two key modules: self-reflection and opponent modeling, and their integration within an evolving decision process.

3.2 Reflective Hypothetical Mind

This section introduces the Reflective Hypothetical Mind (RHM) framework in detail. We first provide an overview of its core components and then describe how self-reflection, hypothetical mind, and adaptation modules interact to enable dynamic sequential reasoning in repeated games. The overall architecture of RHM is illustrated in Figure 2.

3.2.1 Hypothetical Mind. The Hypothetical Mind (HM) module is the predictive core of RHM, enabling the agent to anticipate opponents' strategies by maintaining candidate models, evaluating their plausibility, and guiding decision-making with a form of *theory of mind*. It consists of two main components: *Hypothesis Generation*, which leverages Large Language Model (LLM) to propose candidate opponent models, and *Hypothesis Evaluation*, which scores and selects the most plausible hypotheses for guiding actions.

Hypothesis Generation leverages an LLM to propose candidate models of the opponent's strategy (e.g., "the opponent counters my previous move"), conditioned on past interaction history. This process implicitly incorporates prior knowledge $p(h_i)$ from the LLM's pretrained weights and employs a refinement mechanism that prioritizes high-value hypotheses when generating new ones.

Hypothesis Evaluation estimates the likelihood $p(\mathbf{a} \mid h_i)$ by scoring hypotheses according to predictive accuracy. In each round, the system: (1) selects the top-k hypotheses (with $k = 3$ by default), (2) queries the LLM to predict the opponent's next move under each hypothesis, and (3) updates hypothesis values after observing the actual move. At each round, each hypothesis h_i produces a prediction $\hat{a}_i$, which is rewarded as

$$r_i = \begin{cases} +1 & \text{if } \hat{a}_i = a \\ -1 & \text{if } \hat{a}_i \neq a \end{cases}$$

Values are updated via a recency-weighted rule,

$$V_{h_i} \leftarrow V_{h_i} + \alpha \cdot \left(r_i - V_{h_i}\right)$$

where $\alpha = 0.3$ controls sensitivity to recent outcomes. This bounded update (within $[-1, 1]$) drives V_{h_i} toward 1 under repeated correct predictions.

A hypothesis is deemed *validated* when $V_{h_i} \geq V_{thr} = 0.7$. V_{thr} is set to 0.7 apriori based on analysis of the Rescorla-Wagner dynamics. Validated hypotheses guide decision-making until their values fall below this threshold, ensuring stable use of accurate models while still allowing adaptation when opponent behavior shifts. When multiple hypotheses remain validated, the agent selects its predicted action according to

$$\hat{a} = \arg\max_a \; n_a,$$

where n_a denotes the vote count for action a among the top-k hypotheses (i.e., those with the highest V_{h_i} scores). If no hypothesis qualifies, the most recent candidate is used by default.

3.2.2 Self-Reflection Module. The *self-reflection module* serves as an internal corrective mechanism that enables agents to identify and mitigate suboptimal decision patterns over time. In dynamic

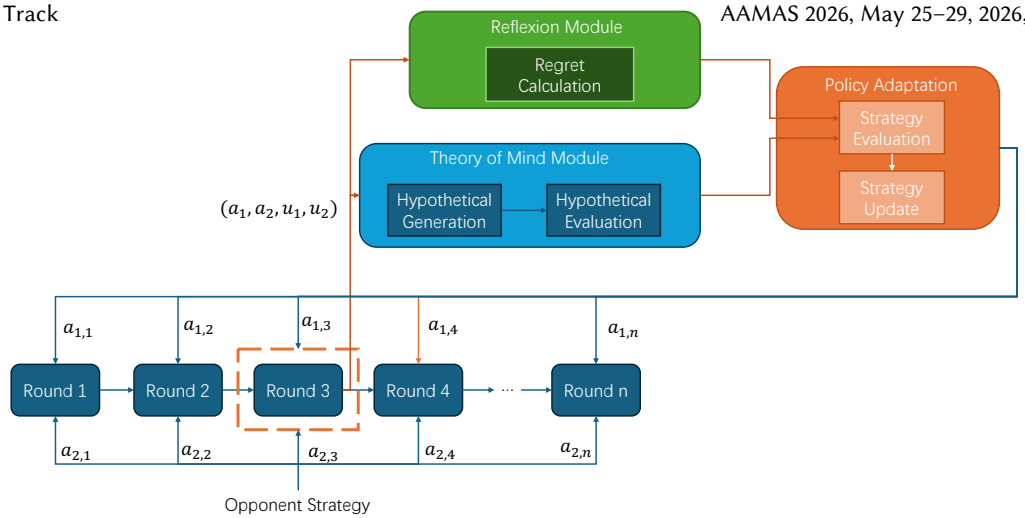

Figure 2: Overall architecture of the Reflective Hypothetical Mind (RHM). The framework integrates three key components that operate across repeated interactions: (1) the *Theory of Mind module* generates and evaluates multiple hypothetical opponent strategies based on observed actions $h^t = (a_1, ..., a_{t-1})$; (2) the *Reflexion module* computes reflection-based regrets by comparing optimal and actual outcomes, refining the agent's internal decision rules; and (3) the *Policy Adaptation module* translates predictive and reflective insights into updated strategies $(a_{i,t})$ for subsequent rounds. Across rounds $t = 1, \ldots, n$, the agent iteratively updates its beliefs and strategies, enabling dynamic opponent modeling and long-horizon policy improvement.

and uncertain environments, LLM agents may make mistakes due to incomplete modeling, hallucinated reasoning, or reliance on outdated beliefs. Without correction, such errors can propagate across rounds, as the model conditions on its prior outputs as part of the input context—potentially reinforcing flawed assumptions and triggering a compounding loop of strategic misalignment.

To address this, we introduce a regret-based self-reflection process after each round. Inspired by game-theoretic regret minimization [39], the agent retrospectively evaluates its previous decision based on the counterfactual payoff: the difference between the reward it received and the reward it could have obtained had it chosen the best possible alternative, assuming the opponent's action was fixed.

Formally, if the agent played action a_i while the opponent played a_{-i}, and a_i^* denotes the best response to a_{-i}, the regret is defined as:

$$\text{Regret}_i(a_i, a_{-i}) = u_i(a_i^*, a_{-i}) - u_i(a_i, a_{-i}) \tag{1}$$

In repeated settings where the opponent's behavior is not fully known, agents cannot guarantee optimal choices at every stage. However, they can still use observed payoffs and opponent modeling to identify clearly suboptimal past moves.

Within the RHM framework, this module complements the hypothetical mind by ensuring that belief updates are coupled with self-correction: the agent not only models evolving opponent strategies but also refines its own decision rules when clear inconsistencies are detected. This design improves resilience against compounding errors (e.g., entering locked cycles of mutual defection) and enhances long-term coherence, providing a foundation for robust adaptation in dynamic repeated games.

3.2.3 Policy Adaptation. The policy adaptation module operationalizes the insights from both the hypothetical mind and the self-reflection components into concrete behavioral updates. Its role is to mediate between *belief formation* and *decision execution*, ensuring that agents can adjust strategies dynamically while balancing external predictions with internal corrective signals.

At each round t, the agent maintains a candidate strategy set Π_t and selects a policy $\pi_t \in \Pi_t$ conditioned jointly on the most reliable hypothesis h^* and accumulated reflection signals s_t. Adaptation proceeds in two stages:

1. Strategy Evaluation. The agent estimates the opponent's action distribution $p(a_{-i} \mid h^*)$ from the validated hypothesis while simultaneously incorporating self-reflection corrections. The utility of a candidate policy π is scored as

$$U(\pi \mid h^*, s_t) = (1-\lambda)\,\mathbb{E}_{a_{-i}\sim p(\cdot|h^*)}[u(\pi, a_{-i})] - \lambda\, Regret(\pi, a_{-i}),$$

where $u(\pi, a_{-i})$ is the payoff against predicted actions and $Regret(\pi, s_t)$ captures regret-adjusted corrections derived from reflection. The weight $\lambda \in [0, 1]$ balances reliance on forward-looking predictions versus retrospective self-correction.

2. Strategy Update. If the current policy π_t yields stable gains (e.g., consistent payoff improvement or reduced regret), it is retained; otherwise, the agent shifts toward alternative candidates in Π_t or prompts the LLM to generate new strategies. This update scheme avoids exhaustive search and provides flexibility beyond classical equilibrium-search methods such as fictitious play [11] or Policy Space Response Oracle [18].

***Theorem 1.* (Convergence of Reflective Hypothetical Mind)** Suppose that (i) the opponent's strategy distribution is stationary, (ii) the true opponent model lies within the hypothesis space, and (iii) the reflection-based regret updates are bounded and unbiased. Then, the Reflective Hypothetical Mind (RHM) procedure converges almost surely to a Nash equilibrium policy $\pi^\star$, such that

$$\pi^\star \in \arg\max_{\pi_i \in \Pi_i} \mathbb{E}_{a_{-i}\sim p(\cdot|h^\star)}[u_i(\pi_i, a_{-i})],$$

where $h^\star$ denotes the asymptotically accurate opponent hypothesis. Consequently, $\lim_{t\to\infty} \text{Regret}_i(t) = 0$, implying that each agent's policy forms a Nash equilibrium conditioned on correct opponent inference.

Proof. Under the stated assumptions, as the hypothetical mind (HM) refines its predictions of the opponent, the accuracy $p_{h,t}$ of the validated hypothesis $h^\star$ improves over time. The value update

$$V_{h,t+1} = (1-\alpha)V_{h,t} + \alpha r_{h,t}, \quad r_{h,t} \in \{-1, +1\}, \tag{2}$$

tracks this accuracy, and standard stochastic approximation results [4, 17] imply that $V_{h,t}$ converges almost surely to a deterministic limit V_h^∞ for each $h \in \mathcal{H}$. Since the true model $h^\star$ is correctly specified, its limiting value dominates, i.e., $V_{h^\star}^\infty > V_h^\infty$ for all $h \neq h^\star$. Therefore, in the long run the procedure almost surely identifies and validates the true hypothesis $h^\star$, establishing concentration on the correct opponent model.

As predictions become precise, the reflection-based regret $\text{Regret}_i(t)$ decreases monotonically, and once $\text{Regret}_i(t) \to 0$, the policy adaptation module converges to

$$\pi^\star = \arg\max_{a_1 \in \mathcal{A}_1} \mathbb{E}[R_1(a_1, a_2) \mid a_2 \sim h^\star],$$

the best response to the accurately inferred opponent strategy a_2. Hence, the limiting policy $\pi^\star$ constitutes a Nash equilibrium conditioned on correct opponent inference, and the agent's per-round regret vanishes asymptotically. □

By explicitly combining predictive beliefs with corrective reflections, the policy adaptation module ensures that no single source dominates decision-making. Instead, strategy evolution emerges from the synergy of forward anticipation and retrospective adjustment, enabling resilient adaptation in dynamic repeated games.

4 EXPERIMENTAL RESULTS

We evaluate the proposed Reflective Hypothetical Mind (RHM) framework against diverse baselines across various classes of repeated games, ranging from self-interested zero-sum and coordination settings to more complex cyclic and large-scale games.

4.1 Baselines

We compare the proposed Reflective Hypothetical Mind (RHM) framework against three representative baselines that capture different reasoning and adaptation mechanisms in large language model (LLM)-based agents: Hypothetical Mind, Reflection, and a direct LLM policy without auxiliary reasoning modules. These baselines collectively cover a spectrum from explicit opponent modeling to pure linguistic decision-making.

- **Hypothetical Mind** [6]. This method builds upon the Hypothetical Mind framework, which equips LLM agents with the ability to infer and update beliefs about their opponents' strategies. At each round, the agent generates multiple candidate hypotheses about the opponent's behavioral rule (e.g., "the opponent tends to counter my last move") and evaluates their predictive accuracy based on observed actions. The most validated hypothesis is then used to guide subsequent decisions. This process effectively scaffolds a rudimentary Theory of Mind within the LLM, allowing it to anticipate others' intentions and adapt strategically during repeated interactions.
- **Reflexion** [27]. The Reflexion framework introduces an explicit self-reflective mechanism designed to improve decision stability in long-horizon reasoning tasks. After each round, the agent performs a textual self-assessment of its previous decision, identifying potential errors or inconsistencies in reasoning. These reflections are incorporated as additional context in the next round's prompt, forming a lightweight verbal reinforcement learning loop. Through repeated reflection and correction, the agent learns to recognize and avoid suboptimal reasoning trajectories, enhancing its robustness in dynamic game environments.
- **LLM** [1]. The third baseline serves as a minimal setup, where the LLM directly maps the observed game history to an action via a single prompt, without any explicit reflection or opponent-modeling module. The prompt contains the past sequence of actions and rewards in natural language, and the LLM is asked to reason about the next move directly. This setting evaluates the LLM's intrinsic ability to perform game-theoretic reasoning and strategic adaptation purely through its pretrained knowledge and in-context reasoning capability, without additional cognitive scaffolding.

4.2 Experimental Settings

All experiments were conducted on the GPT-4o model. We set $\delta = 1$ to preserve long-horizon strategic patterns—central to repeated-game reasoning and the RHM framework—without bias toward immediate rewards. For different repeated game environments, the number of iterations τ was set according to the complexity of the game:

- For Iterated Prisoner's Dilemma and Iterated Battle of Sexes, $\tau = 10$;
- For Iterated Rock–Paper–Scissors, $\tau = 20$
- For Iterated Colonel Blotto, $\tau = 30$

To ensure the reliability of experimental outcomes, each repeated game was simulated five times and averaged to obtain final performance metrics.

(1) Iterated Prisoner's Dilemma (IPD) [2]. The Prisoner's Dilemma describes a situation in which two prisoners must independently decide whether to cooperate (C) or defect (D). The payoff structure is as follows: if both cooperate, they each receive a moderate reward of 6; if one defects while the other cooperates, the defector receives a high payoff of 10 while the cooperator receives 0; if both defect, each receives a low payoff of 2. The payoff matrix is shown below:

	C	D
C	$(6, 6)$	$(0, 10)$
D	$(10, 0)$	$(2, 2)$

Six opponent strategies are employed to evaluate the behavior of the LLM agent:

Grim Trigger (GT) — starts with cooperation (C) but permanently switches to defection (D) once the opponent defects, modeling unforgiving strategies.

Tit-for-Tat (TfT) — begins with cooperation and subsequently mirrors the opponent's previous action. To introduce behavioral variance, it is modified to defect once in the 10th round regardless of prior history.

Always Defect (AD) — consistently plays D, representing a purely self-interested, Nash-equilibrium-seeking opponent that refuses to cooperate.

Surrender (SR) — opens with defection (D); if the opponent does not retaliate with D, it continues defecting, but if the opponent defects in return, it switches to permanent cooperation (C), enabling exploitation by adaptive agents.

Cooperative LLM — an LLM agent prompted with a *cooperative personality*, designed to test the agent's responsiveness to prosocial cues.

Human LLM — an LLM agent prompted with a *human-like personality*, used to assess the model's ability to adapt to human behavioral tendencies.

These six opponent types are designed to evaluate the LLM agent's level of *sequential rationality* and its ability to adjust strategies when interacting with diverse behavioral patterns.

(2) Iterated Battle of Sexes [5]. The Battle of Sexes represents coordination under preference asymmetry. Two players wish to coordinate but prefer different outcomes: Player 1 prefers Option A while Player 2 prefers Option B. The payoff matrix is as follows:

	A	B
A	$(10, 7)$	$(0, 0)$
B	$(0, 0)$	$(7, 10)$

The following opponent strategies are used:

Alternation — alternates between A and B across rounds, testing the agent's ability to recognize temporal patterns.

Always Option B — always selects B.

Cooperative LLM — an LLM prompted for cooperative behavior, aiming to maximize joint reward.

Human LLM — an LLM prompted to mimic human-like decision-making with imperfect consistency.

This setup evaluates the LLM agent's ability to achieve stable coordination or recover from miscoordination cycles.

(3) Iterated Rock–Paper–Scissors [28]. Rock–Paper–Scissors is a cyclic zero-sum game where Rock beats Scissors, Scissors beats Paper, and Paper beats Rock. Each win gives a payoff of +1, a loss gives −1, and a tie gives 0. The payoff matrix is:

	R	P	S
R	$(0, 0)$	$(-1, 1)$	$(1, -1)$
P	$(1, -1)$	$(0, 0)$	$(-1, 1)$
S	$(-1, 1)$	$(1, -1)$	$(0, 0)$

Opponent strategies include:

Tit-for-Tat — mirrors the opponent's last move.

Alternation — cycles through Rock, Paper, and Scissors sequentially.

Always Rock — consistently plays Rock.

Always Paper — consistently plays Paper.

Always Scissors — consistently plays Scissors.

This environment probes the agent's capacity for pattern recognition, adaptation, and exploitation of predictable strategies.

(4) Iterated Colonel Blotto Game [26]. The Colonel Blotto game involves two players simultaneously allocating limited resources (e.g., troops or energy units) across multiple battlefields. Each battlefield is won by the player allocating more resources, and total payoff equals the number of battlefields won. The payoff function is:

$$u_i\left(a_i, a_{-i}\right) = \begin{cases} 1 & \text{if } \sum_{k=1}^{K} \mathbb{I}\left[a_i^k > a_{-i}^k\right] > \sum_{k=1}^{K} \mathbb{I}\left[a_i^k < a_{-i}^k\right], \\ 0 & \text{if } \sum_{k=1}^{K} \mathbb{I}\left[a_i^k > a_{-i}^k\right] = \sum_{k=1}^{K} \mathbb{I}\left[a_i^k < a_{-i}^k\right], \\ -1 & \text{otherwise.} \end{cases}$$

where n is the number of battlefields (set to 3 in our experiments). The opponent follows an Alternation policy, cyclically varying its allocation pattern to test whether the LLM agent can learn adaptive resource allocation strategies across rounds.

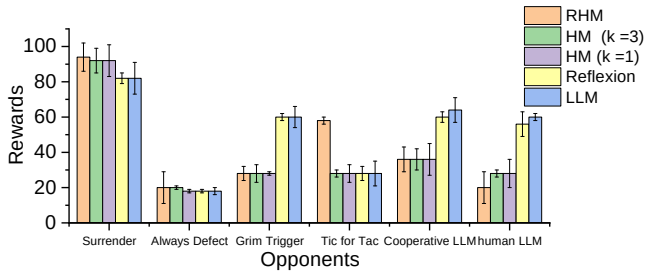

Figure 3: Average accumulated rewards of RHM, HM ($k = 1, 3$), Reflexion, and LLM baselines in Iterated Prisoner's Dilemma against diverse opponent strategies including *Surrender*, *Always Defect*, *Grim Trigger*, *Tit for Tat*, , *Cooperative LLM*, and *Human LLM*. As the results indicate, RHM achieves performance comparable to the LLM baseline. When facing opponent strategies such as *Surrender*, *Always Defect*, and *Tit-for-Tat*, RHM obtains the highest rewards. In contrast, the LLM baseline performs best against strategies including *Grim Trigger*, *Cooperative LLM*, and *Human LLM*. These findings suggest that, even without explicit opponent modeling or a reflective module, a plain LLM performs well in self-interested games.

4.3 Self-Interested Games

In self-interested settings such as the repeated Prisoner's Dilemma [2], large language models (LLMs) already exhibit competitive performance by approximating cooperative strategies. As shown in Figure 3, the baseline LLM achieves the best performance against opponent strategies such as Grim Trigger, Cooperative LLM, and Human LLM, demonstrating its strong reasoning and adaptation capabilities in repeated interactions, while Reflective Hypothetical Mind (RHM) attains the highest rewards when facing opponents including Surrender, Always Defect and Tit for Tat.

This improvement stems of RHM from addressing a key limitation of the Hypothetical Mind (HM) framework: when interacting with Tit for Tat-like opponents, HM tends to fall into a mutual defect loop, where both sides continuously defect and fail to re-establish cooperation. As shown in the upper panel of Figure 4, HM exhibits oscillatory behavior, with both agents repeatedly choosing Defect action, reflecting unstable coordination and persistent defection.

To overcome this issue, RHM incorporates a reflection module that enables the agent to identify and break out of such deadlocks by reasoning about past interactions and adjusting its future actions accordingly. As illustrated in the lower panel of Figure 4, RHM successfully restores coordination, with both agents converging to consistent cooperative choices over time. This reflective mechanism allows RHM to avoid mutual defection and maintain stable cooperation, leading to higher overall payoffs. As shown in left panel in Figure 4, the RHM agents achieve a consistently higher cumulative reward across rounds, indicating improved stability and faster adaptation in cooperative dynamics.

4.4 Cooperative Games

In contrast, coordination-oriented games such as the repeated Battle of Sexes [5] require agents not only to establish but also to sustain

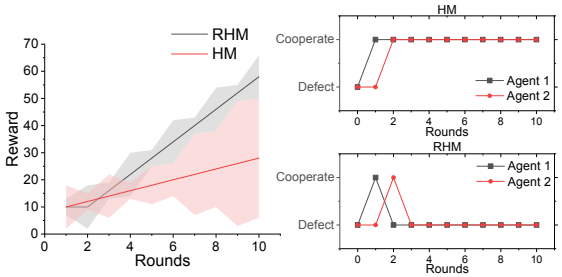

Figure 4: Broken Detrimental Cycles in Iterated Prisoner's Dilemma Game. As shown in the figure, HM agent falls into the detrimental strategic cycles described in Figure 1 when interacting with a *Tit-for-Tat* opponent, due to its short-sighted self-interest. In contrast, RHM overcomes this issue through its reflection module: by actively choosing *Cooperate*, RHM breaks the cycle and restores stable mutual cooperation.

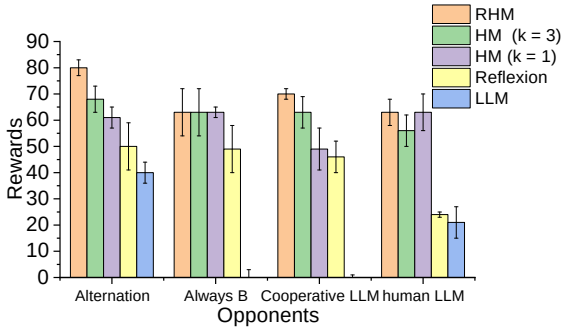

Figure 5: Average accumulated rewards of RHM, HM ($k = 1, 3$), Reflexion, and LLM baselines in Iterated Battle of Sexes against diverse opponent strategies including *Alternation*, *Always B*, *Cooperative LLM* and *Human LLM*. RHM consistently achieves the highest reward against most opponents, particularly those requiring dynamic adaptation such as *Alternation*. These results highlight that explicit opponent modeling and reflective policy adaptation enable RHM to maintain stable cooperation.

mutual cooperation in order to achieve Pareto-efficient outcomes. Unlike self-interested games where maximizing individual payoff suffices, success here hinges on the ability to infer and adapt to a partner's latent preference for one of the two coordination options. The baseline LLM, despite its strong language-based reasoning ability, often fails to form a stable joint strategy and becomes trapped in detrimental cycles.

Figure 5 quantifies this effect across a diverse set of opponents. While the baseline LLM exhibits highly unstable coordination rates, the hypothetical mind (HM; $k = 1, 3$) significantly raises successful alignment frequency by explicitly modeling the counterpart's likely response patterns. Its reflective extension, RHM, achieves the highest overall performance against all tested opponents, suggesting that reflection on past interaction histories is particularly powerful for resolving coordination dilemmas. Notably, RHM maintains strong coordination even under opponents that deliberately attempt to destabilize convergence.

To further analyze this capability, Figure 6 depicts the action trajectories of RHM when Agent 2 adopts an Alternation strategy, a challenging adversary that systematically oscillates between the

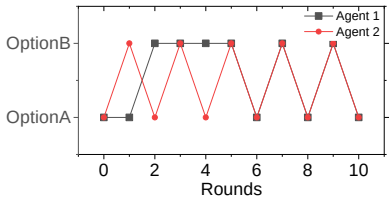

Figure 6: Action Trajectories in iterated Battle of Sexes with Alternation Opponent. The figure illustrates the evolution of action choices for two agents over repeated rounds when Agent 2 follows an *Alternation* strategy, switching between *Option A* and *Option B*. Under this dynamic opponent behavior, RHM rapidly adapts its strategy through reflective policy updates, aligns with the opponent's preference, and achieves stable convergence to coordinated choices within only a few rounds.

two coordination options to disrupt stable alignment. As shown in Figure 6, RHM rapidly recognizes the alternating pattern, leverages its reflection module to revise its belief about the opponent's latent preference, and locks onto a consistent choice This accelerated convergence demonstrates RHM's ability to break inefficient loops by integrating opponent modeling with retrospective reasoning over past rounds.

Collectively, these results highlight that in coordination-centric environments, explicit opponent modeling combined with self-reflection is critical for escaping miscoordination traps, accelerating convention formation, and sustaining efficient cooperation, outperforming both purely reactive reasoning (baseline LLM) and non-reflective hypothetical minds.

4.5 Complex Cyclic Games

In cyclic environments such as repeated Rock–Paper–Scissors (RPS) [28], the limitations of pure opponent modeling become evident. Although LLMs equipped with HM reasoning can anticipate short-term response patterns—for example, recognizing simple alternation or inferring that an opponent may react to the agent's last move—this reasoning remains decoupled from actionable policy updates. HM agents often produce accurate but passive predictions that stay at the belief level; they know what the opponent is likely to play next but fail to translate these predictions into consistent behavioral shifts. As a result, their play remains locally responsive but globally unstable, frequently drifting or oscillating instead of converging on a payoff-dominant adaptation.

Figure 7 makes this gap explicit. The baseline LLM and Reflexion attain only moderate payoffs and struggle especially against opponents with structural biases (e.g., Always Rock or Always Paper), failing to exploit their predictability. HM variants ($k = 1, 3$) do improve over these baselines by forecasting near-term moves and sometimes adapting to simple cycles, but their gains remain inconsistent in persistent cyclic dynamics (e.g., Alternation, Tit-for-Tat) and insufficient when exploiting static biases requires cumulative, history-aware adaptation. In these cases, HM predictions either fail to accumulate over time or do not trigger systematic policy revision, leading to reactive but ultimately myopic play.

By contrast, RHM achieves the highest average reward across all tested opponents, outperforming every baseline under both pattern-based adversaries (Alternation, Tit-for-Tat) and static biased ones

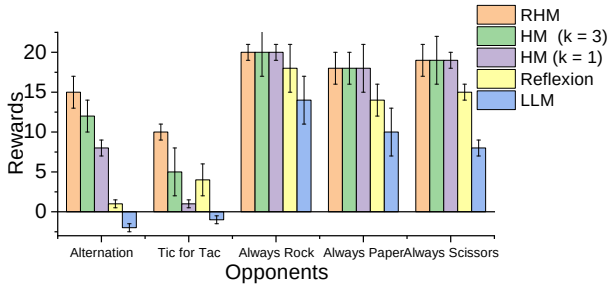

Figure 7: Average accumulated rewards of RHM, HM ($k = 1, 3$), Reflexion, and LLM baselines in Iterated Rock-Paper-Scissors against diverse opponent strategies including *Alternation*, *Tit-for-Tat*, *Always Rock*, *Always Paper*, and *Always Scissors*. Across all opponent types, RHM consistently achieves the highest accumulated rewards, demonstrating robust adaptation to both cyclic and deterministic opponent behaviors. When facing static opponents (*Always Rock/Paper/Scissors*), RHM and HM achieve comparable rewards, indicating that opponent modeling alone suffices under fixed deterministic dynamics. These results highlight that RHM's reflective adaptation mechanism is critical for maintaining superior performance in dynamic cyclic environments.

(Always Rock/Paper/Scissors). Crucially, RHM integrates a policy adaptation module that operationalizes its reflective reasoning: it evaluates past rounds, updates its strategic hypothesis about the opponent's latent bias or cycle, and then commits to long-term counter-strategies rather than merely reacting one step at a time. This enables RHM to consistently exploit deterministic biases (e.g., reliably defeating Always Rock) while also maintaining competitive performance in dynamically shifting cyclic play.

Taken together, these results show that opponent modeling alone—no matter how accurate—cannot guarantee robust performance in complex cyclic games. To succeed, agents must couple reasoning with systematic policy adaptation, transforming predictions into enduring behavioral change that can exploit both static asymmetries and evolving interaction dynamics.

4.6 Large Scale Games

The Colonel Blotto game [26] models resource allocation, where two players distribute a fixed budget s over n battlefields. Each player selects an allocation vector $(a_1, \ldots, a_n)$ with $\sum_{i=1}^{n} a_i = s$; each field is won by the player assigning more resources. Compared with the previous games, Colonel Blotto introduces a *combinatorially large action space*: even with only $n = 3$ fields, the number of feasible allocations rapidly expands to 100, 225, 441, and 784 when $s = 3, 4, 5, 6$, respectively. We fix $n = 3$ and gradually increase s to examine scalability.

As shown in Figure 8, performance diverges sharply as the allocation space grows [26]. The RHM maintains a clear and widening advantage, achieving the highest rewards across all budget sizes. In contrast, HM—despite its ability to model opponents and perform stepwise policy adaptation—consistently yields lower rewards than RHM as s increases and the strategy space expands. This reveals that while HM can exploit small-pattern reasoning in modest domains, its adaptation remains locally reactive and unable to search or commit to globally advantageous allocations in a vastly expanded space.

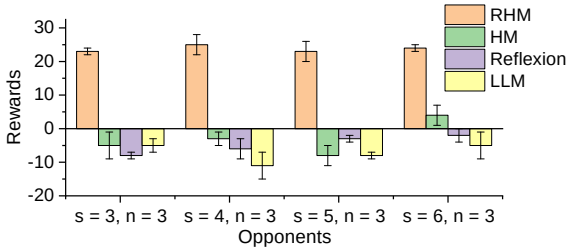

Figure 8: Average accumulated rewards of RHM, HM ($k = 1, 3$), Reflexion, and LLM baselines in Iterated Colonel Blotto as the total resource budget s increases from 3 to 6 with a fixed number of battlefields $n = 3$. RHM consistently achieves the highest rewards across all settings, demonstrating strong adaptability and scalability. While HM maintains moderate stability, its performance remains consistently below that of RHM as s increases, indicating limited scalability of stepwise belief updates. Reflexion and the plain LLM baseline exhibit sharp degradation under larger s, confirming that reflective opponent modeling and structured policy adaptation are essential for handling complex allocation spaces.

Reflexion and the plain LLM baseline perform even worse, showing severe instability and near-random play under large budgets.

These results demonstrate that opponent modeling plus local policy updates is insufficient in combinatorially complex environments. To handle large strategic spaces such as Colonel Blotto, agents must combine reflective reasoning with structured, sequential policy adaptation—as implemented in RHM—to avoid search inefficiency and maintain robust performance when the action space grows exponentially.

5 CONCLUSION

This work investigates how large language models (LLMs) can be enhanced with explicit reasoning and adaptive mechanisms to achieve robust performance in complex multi-agent games. We first analyze the Hypothetical Mind (HM) framework and show that, although HM improves short-term coordination and opponent prediction, it often fails to convert rich opponent modeling into stable long-term policies. Consequently, HM exhibits inconsistent behavior in cyclic games such as repeated Rock–Paper–Scissors and struggles to scale to large allocation spaces.

To address these limitations, we propose the Reflective Hypothetical Mind (RHM), which augments HM with a reflection-driven policy adaptation module that translates opponent reasoning into sustained behavioral change. Extensive experiments across coordination games, cyclic environments, and large-scale allocation problems demonstrate that RHM consistently outperforms strong baselines, including plain LLMs, Reflexion, and HM variants. These results highlight a central insight: accurate opponent modeling alone is insufficient for robust multi-agent performance—effective agents must systematically convert reflective insights into adaptive policies to cope with non-stationary dynamics and combinatorially large strategy spaces.

ACKNOWLEDGMENTS

This research is supported by the InnoHK Funding.

REFERENCES

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour* (2025), 1–11.

[2] Robert Axelrod. 1980. Effective choice in the prisoner's dilemma. *Journal of conflict resolution* 24, 1 (1980), 3–25.

[3] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62.

[4] Michel Benaïm. 1999. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités XXXIII* (1999), 1–68.

[5] Blanche Capel. 2000. The battle of the sexes. *Mechanisms of development* 92, 1 (2000), 89–103.

[6] Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. 2025. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[8] Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2025. Nicer Than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 522–535.

[9] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2023), 44123–44279.

[10] Fulin Guo. 2023. GPT in game theory experiments. *arXiv preprint arXiv:2305.05516* (2023).

[11] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious self-play in extensive-form games. In *International conference on machine learning*. PMLR, 805–813.

[12] Nathan Herr, Fernando Acero, Roberta Raileanu, María Pérez-Ortiz, and Zhibin Li. 2024. Are Large Language Models Strategic Decision Makers? A Study of Performance and Bias in Two-Player Non-Zero-Sum Games. *arXiv preprint arXiv:2407.04467* (2024).

[13] Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227* (2023).

[14] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745* (2023).

[15] Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. 2023. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square* (2023), rs–3.

[16] P Kumar, S Manikandan, and R Kishore. 2024. Ai-driven Text Generation: A Novel Gpt-based Approach for Automated Content Creation. In *2024 2nd International Conference on Networking and Communications (ICNWC)*. IEEE, 1–6.

[17] Harold J. Kushner and G. George Clark. 1978. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer.

[18] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems* 30 (2017).

[19] Pier Luca Lanzi and Daniele Loiacono. 2023. Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 1383–1390.

[20] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2023. Econagent: large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436* (2023).

[21] Wenhao Li, Dan Qiao, Baoxiang Wang, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. 2023. Semantically aligned task decomposition in multi-agent reinforcement learning. *arXiv preprint arXiv:2305.10865* (2023).

[22] Yuxuan Li and Hirokazu Shirado. 2025. Spontaneous giving and calculated greed in language models. *arXiv preprint arXiv:2502.17720* (2025).

[23] Nunzio Lorè and Babak Heydari. 2023. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898* (2023).

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

[25] Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559* (2023).

[26] Brian Roberson. 2006. The colonel blotto game. *Economic Theory* 29, 1 (2006), 1–24.

[27] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 8634–8652.

[28] Barry Sinervo and Curt M Lively. 1996. The rock–paper–scissors game and the evolution of alternative male strategies. *Nature* 380, 6571 (1996), 240–243.

[29] Nigar M Shafiq Surameery and Mohammed Y Shakor. 2023. Use chat gpt to solve programming bugs. *International Journal of Information Technology and Computer Engineering* 31 (2023), 17–22.

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

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[32] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[33] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. 2024. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proc. the 36th International Conference on Neural Information Processing Systems*, Vol. 1126.

[34] Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. 2024. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17591–17599.

[35] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230* (2024).

[36] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. 2024. K-Level Reasoning: Establishing Higher Order Beliefs in Large Language Models for Strategic Reasoning. *arXiv preprint arXiv:2402.01521* (2024).

[37] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition dynamics in large language model-based agents. *arXiv preprint arXiv:2310.17512* (2023).

[38] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).

[39] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems* 20 (2007).