A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios

Anonymous authors
Paper under double-blind review

Abstract

Game-theoretic scenarios have become pivotal in evaluating the social intelligence of Large Language Model (LLM)-based social agents. While numerous studies have explored these agents in such settings, there is a lack of a comprehensive survey summarizing the current progress. To address this gap, we systematically review existing research on LLM-based social agents within game-theoretic scenarios. Our survey organizes the findings into three core components: Game Framework, Social Agent, and Evaluation Protocol. The game framework encompasses diverse game scenarios, ranging from choice-focusing to communication-focusing games. The social agent part explores agents' preferences, beliefs, and reasoning abilities. The evaluation protocol covers both game-agnostic and game-specific metrics for assessing agent performance. By reflecting on the current research and identifying future research directions, this survey provides insights to advance the development and evaluation of social agents in game-theoretic scenarios.

1 Introduction

The rapid advancement of Large Language Models (LLMs) (Achiam et al., 2023; Team et al., 2023; Jiang et al., 2023; Yang et al., 2024a; Dubey et al., 2024) has achieved exceptional performance across a wide array of applications, including personal assistant (Li et al., 2024b), search engines (Chen et al., 2024b), code generation (Wang et al., 2024b) and embodied intelligence (Liu et al., 2024a). Building on this capability, a growing area of research focuses on employing LLMs as central controllers to develop autonomous agents with human-like decision-making abilities (Sumers et al., 2023; Wang et al., 2024a). This progress brings the realization of Artificial General Intelligence (AGI) within reach (Bubeck et al., 2023), paving the way for a future where human-AI interaction, collaboration, and coexistence shape a shared, symbiotic society (Mahmud et al., 2023; Ren et al., 2024). Therefore, it is crucial to evaluate and enhance the *social intelligence* of AI, particularly LLM-based social agents, as it determines their ability to engage effectively in sophisticated social scenarios (Mathur et al., 2024).

Social intelligence is the foundation of all successful interpersonal relationships and is also a prerequisite for AGI (Hunt, 1928; Kihlstrom & Cantor, 2000; Hovy & Yang, 2021). Drawing on insights from both social science and AI research, Li et al. (2024a) has established a comprehensive Social AI Taxonomy, which categorizes social intelligence into three dimensions: *situational intelligence*, the ability to comprehend the social environment (Derks et al., 2007); *cognitive intelligence*, the ability to understand others' intents and beliefs (Barnes & Sternberg, 1989); and *behavioural intelligence*, the ability to behave and interact appropriately (Ford & Tisak, 1983). To evaluate artificial social intelligence, researchers have conducted extensive studies, with particular focus on *game-theoretic scenarios*, as these studies simultaneously encompass all above three dimensions of social intelligence (Aher et al., 2022; Horton, 2023; Phelps & Russell, 2023; Akata et al., 2023; Brookins & DeBacker, 2023).

Game theory, a long-established field in microeconomics, offers a robust mathematical framework for analyzing social interactions among cooperating and competing players, with wide-ranging applications (Fudenberg & Tirole, 1991; Camerer, 2011). Specifically, evaluations in game-theoretic scenarios require social agents to understand the game scenario, infer opponents' actions, and adopt appropriate responses, representing an

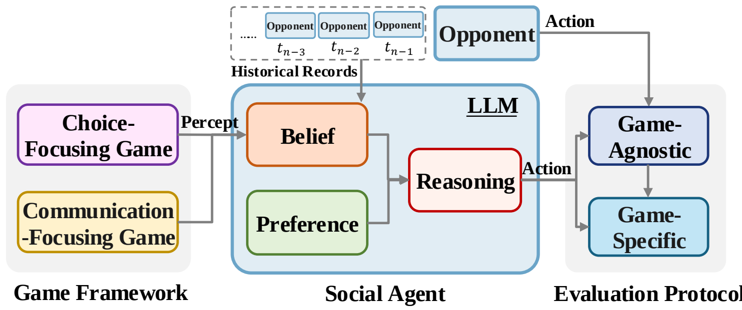

Figure 1: Taxonomy of LLM-based social agents in game-theoretic scenarios.

advanced form of social intelligence (Van Der Hoek et al., 2005; Zhang et al., 2024b). Moreover, the multi-agent participation and dynamic nature of the environment in game scenarios present additional challenges for social agents. Consequently, extensive research has examined social agents within game-theoretic scenarios, offering substantial empirical evidence for understanding their social intelligence (Guo, 2023; Meng, 2024; Mei et al., 2024). However, there is currently a lack of a comprehensive review that summarizes the current progress in this area and considers future directions.

To address this gap, we have thoroughly reviewed the existing research on LLM-based social agents in game-theoretic scenarios and have organized the findings according to a meticulously designed taxonomy, as illustrated in Figure 1. Specifically, the taxonomy comprises three main components: Game Framework (§2), Social Agent (§3), and Evaluation Protocol (§4). The Game Framework section includes two parts: Choice-Focusing Game (§2.1) and Communication-Focusing Game (§2.2). *Choice-Focusing Game* refers to a series of scenarios where participants engage with little to no communication, such as classic game-theoretic games (Brookins & DeBacker, 2023; Hua et al., 2024a) and poker (Yim et al., 2024). *Communication-Focusing Game* refers to games where communication among participants is a core component, such as negotiation (Bianchi et al., 2024) and diplomacy (Bakhtin et al., 2022). The Social Agent section comprises three parts: Preference Module (§3.1), Belief Module (§3.2), and Reasoning Module (§3.3). *Preference Module* focuses on research analyzing the intrinsic preferences of LLMs and their ability to follow internal or pre-defined preferences (Guo, 2023). *Belief Module* explores studies on the internal beliefs of models, belief enhancement, and belief revision (Fan et al., 2023). *Reasoning Module* examines research on strategic reasoning, particularly involving theory-of-mind capabilities and reinforcement learning (Guo et al., 2023). The Evaluation Protocol section consists of two parts: Game-Agnostic Evaluation (§4.1) and Game-Specific Evaluation (§4.2). *Game-Agnostic Evaluation* focuses on universal metrics that can be used to assess game outcomes (Duan et al., 2024b). *Game-Specific Evaluation* emphasizes context-specific metrics tailored to the evaluation dimensions of particular game scenarios (Qi et al., 2024).

Based on the above taxonomy, we provide a detailed summary of current research progress, reflect on each part, and offer insights into potential future research directions (§5), with the aim of inspiring further studies in this evolving field.

2 Game Framework

In this section, we describe the game-theoretic scenarios explored in existing research, including both choice-focusing games and communication-focusing games.

2.1 Choice-Focusing Game

Choice-focusing games are game-theoretic scenarios in which participants make decisions based primarily on observable actions and environmental conditions, with minimal or no communication involved. Existing research focuses on social agents in three types of choice-focusing scenarios: *classic game-theoretic games*, *poker*, and *auctions*. Some game examples are shown in Figure 3.

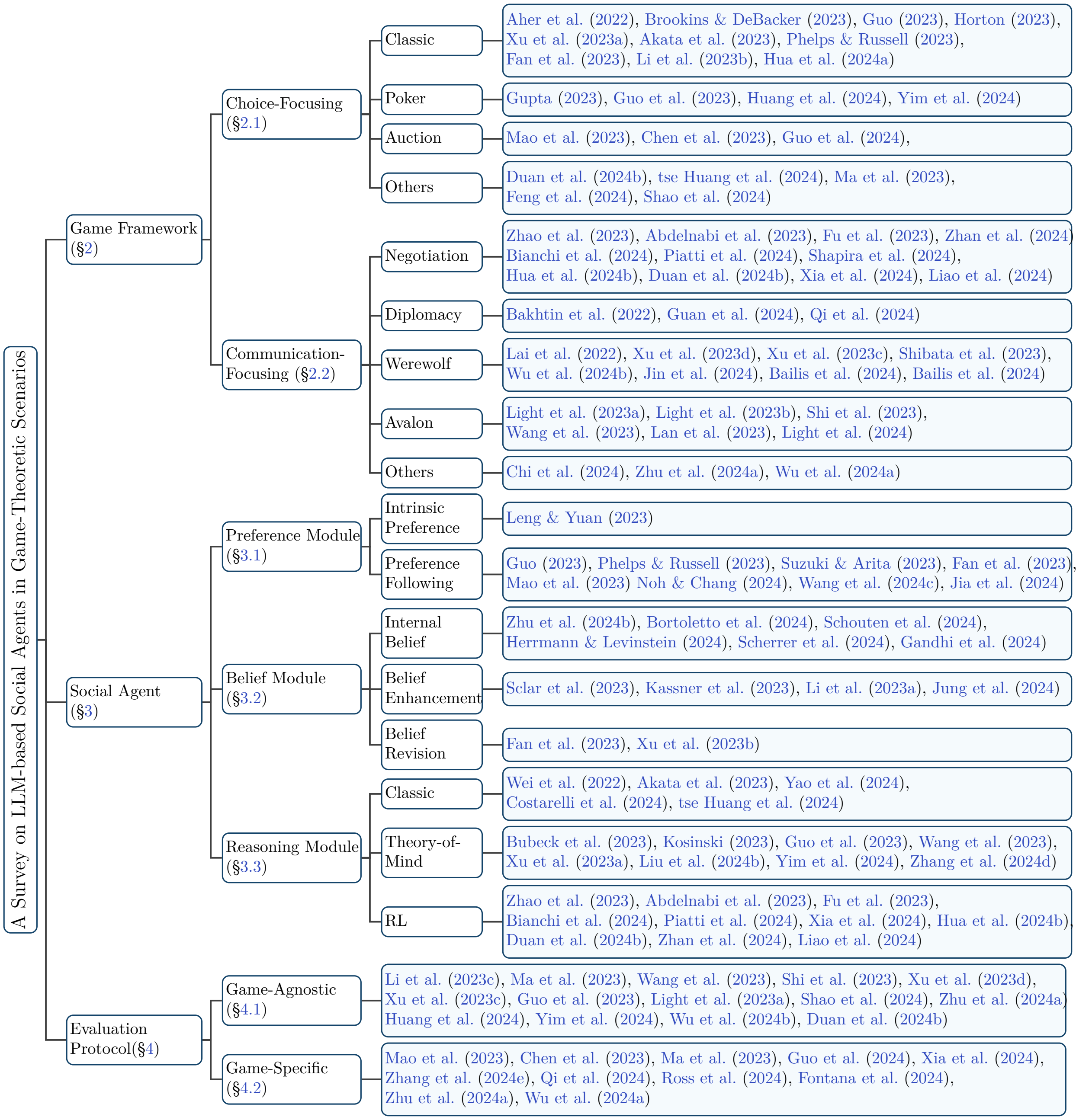

Figure 2: Taxonomy of recent research on LLM-based social agents in game-theoretic scenarios.

Classic game-theoretic games, such as the prisoner's dilemma, have been distilled by economists from various real-world situations. These games are well-defined, with rigorous mathematical foundations, and can be extended to numerous scenarios (Owen, 2013). Consequently, many studies have utilized these games as testbeds to study social agents. The prisoner's dilemma (Rapoport & Chammah, 1965), as the most famous and widely recognized game, has been extensively utilized in numerous studies. Brookins & DeBacker (2023) and Guo (2023) evaluated the strategic reasoning capabilities of GPT-3.5 and GPT-4, respectively, in the classic prisoner's dilemma, highlighting the sensitivity of LLM responses to input instructions, which contributes to low output robustness. This underscores the critical need for future evaluations to focus on instruction robustness testing. Furthermore, Akata et al. (2023) and Phelps & Russell (2023) extended their analyses to the iterated prisoner's dilemma, investigating the ability of LLMs to optimize decision-making by

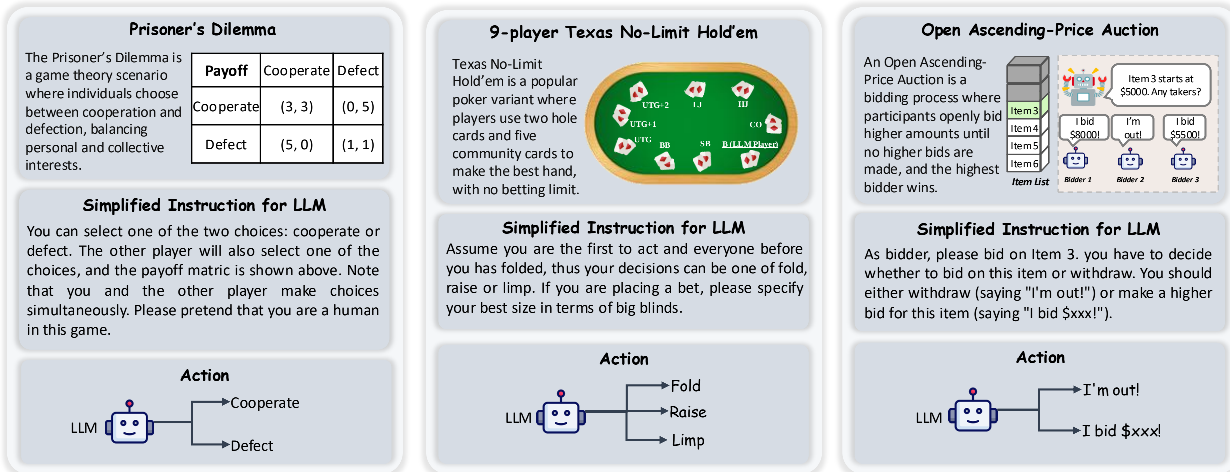

Figure 3: Illustration of choice-focusing games.

utilizing historical information. Interestingly, Brookins & DeBacker (2023) observed that GPT-3.5 replicates human tendencies toward fairness and cooperation, whereas Akata et al. (2023) found GPT-4 to be less tolerant and more rigid in its decision-making. Additionally, Xu et al. (2023a) studied a more complex multi-player iterative prisoner's dilemma scenario within a multi-agent framework driven by LLMs. In addition to the prisoner's dilemma, numerous studies have also employed various classic game-theoretic games as foundational frameworks for research, including the Dictator Game (Horton, 2023; Fan et al., 2023; Brookins & DeBacker, 2023), Ultimatum Game (Aher et al., 2022; Guo, 2023), Public Goods Game (Li et al., 2023b; Xu et al., 2023a), Battle of the Sexes (Akata et al., 2023), Rock-Paper-Scissors (Fan et al., 2023), and Ring-Network Games (Fan et al., 2023).

Poker is a globally popular card game with numerous variations (Waterman, 1970). Winning in poker often requires astute strategic reasoning, as it is a non-cooperative, imperfect information, and dynamic game (Moravčík et al., 2017; Huang et al., 2024). Consequently, many researchers evaluate social agents by assessing their performance as poker players. Gupta (2023) studied 9-player Texas No-Limit Hold'em and concluded that the performance of both ChatGPT and GPT-4 is not game-theory optimal. Furthermore, their findings highlight the divergent poker tactics of the two models: ChatGPT's conservativeness contrasts sharply with GPT-4's aggression. Guo et al. (2023) conducted research on Leduc Hold'em, developing a social agent, Suspicion-Agent, which outperformed traditional reinforcement learning-based agents in poker. They also noted two critical issues: the outputs of LLMs are highly sensitive to the prompts, and the quality of the model's output declines rapidly as the prompt length increases. Yim et al. (2024) focused on Guandan, currently the most popular poker game in China, to investigate cooperative strategies in poker within a Chinese-language context. Interestingly, their experimental results show that while LLMs currently fall short of reinforcement learning models in performance, they underscore the future potential of LLMs in this domain. Poker is a complex game, and investigating whether social agents exhibit behavioural patterns that enable foresighted cooperation and competition in poker presents an intriguing avenue for future research.

Auction is a competitive process in which participants place bids on an item, providing a rich environment for evaluating strategic planning, resource allocation, risk management, and competitive behaviours (Kagel & Levin, 1986). As a typical non-cooperative game with incomplete information, it has garnered significant attention from researchers. Mao et al. (2023) analyzed the performance of LLMs in the "water allocation challenge", a first-price sealed-bid auction. Comprehensive human evaluations revealed that LLMs exhibited superior long-term planning capabilities compared to humans. However, it is noteworthy that despite assigning distinct personas to LLM agents, human evaluators gave low scores for "identity alignment", with significant variance in the results. This indicates that simply adding persona information in system prompts may not sufficiently simulate specific personality traits or the behaviours of professional players. Guo et al. (2024) investigated private-value second-price auctions, demonstrating that while existing models display a certain level of rationality, there remains considerable scope for improvement. Their findings also indicate that LLMs can utilize historical information to refine their strategies and exhibit some degree of convergence. Chen et al. (2023) explored dynamic game scenarios using the open ascending-price auction and introduced the AucArena benchmark. Their experiments showed that even GPT-4 struggles with long-term strategic planning in dynamic, multi-round settings. Success in auctions requires agents to possess exceptional mathematical reasoning abilities. However, this area remains unexplored. Investigating complex mathematical reasoning in auction scenarios presents a promising direction for future research.

To systematically assess LLMs' performance, Duan et al. (2024b) and tse Huang et al. (2024) introduced GTBench and γ-Bench, encompassing multiple game scenarios. The emergence of these benchmarks provides a solid foundation for evaluating social agents in game-theoretical scenarios. Furthermore, some studies have explored agents in games like Chess (Feng et al., 2024) and StarCraft II (Ma et al., 2023; Shao et al., 2024). Chess represents a classic game-theoretic scenario, while StarCraft II, with its complexity and dynamic nature, has also become an ideal testing ground for researching social agents.

Takeaways:

Current research experiments are relatively isolated, *lacking a unified evaluation framework*. Due to the instability of prompt engineering-based experiments, there is an urgent need for a standardized evaluation framework to integrate all experiments and provide consistent insights. Besides, since LLMs are trained on vast amounts of data, there is a *significant risk of data contamination*, meaning that existing classic game-theoretic games may already be present in the pre-training corpus. This could result in evaluation outcomes that do not accurately reflect the LLMs' true strategic reasoning capabilities. Furthermore, although poker and auction involve little verbal communication, existing research *lacks exploration into whether social agents engage in "strategic behaviour" mediated through "action language"*. These gaps hinder a comprehensive understanding of the decision-making processes of social agents.

2.2 Communication-Focusing Game

Communication-focusing games refer to games where communication among participants is a core component, where *language itself serves as a strategy*, allowing participants to influence the game's progress and outcomes through verbal exchanges. These games emphasize interaction between players, with communication playing a crucial role. Leveraging the powerful language capabilities of LLMs, current research has explored the performance of social agents in various communication-focusing games, including *Negotiation*, *Diplomacy*, *Werewolf*, *Avalon*, and others. Some game examples are shown in Figure 4.

Negotiation involves two or more individuals engaging in discussions to resolve conflicts, achieve mutual benefits, or reach mutually acceptable solutions (Bazerman et al., 2000; Zhan et al., 2024). Given that negotiation encompasses complex game behaviours, including non-zero-sum games, incomplete information games, non-cooperative and cooperative games, as well as repeated games, it represents a highly significant research domain. Abdelnabi et al. (2023) evaluated the negotiation capabilities of social agents by building upon an existing negotiation role-play exercise (Susskind, 1985) and incorporating three negotiation games synthesized using LLMs. By configuring agents with varying incentives, the experimental results revealed that agents' behaviour could be modulated to promote greediness or attack other agents. Meanwhile, other agents in the environment demonstrated the ability to detect intruders. These findings underscore the need for future research to focus on attack and defense mechanisms within multi-agent systems. Bianchi et al. (2024) developed NegotiationArena, a platform featuring three types of games: allocating shared resources (ultimatum games), aggregating resources (trading games), and buying/selling goods (price negotiations). Experimental results reveal that LLM agents are also prone to anchoring and numerosity biases. Interestingly, social behaviours, such as pretending to be desperate or using insults, were found to significantly enhance the agents' payouts. A similar resource competition scenario is customer acquisition. Zhao et al. (2023) designed restaurant agents and customer agents, examining how restaurant agents compete with one another to attract and retain customers. The simulation results revealed several phenomena analogous to those observed in real society, such as the Matthew Effect. Piatti et al. (2024) created a simulation environment called GovSim, which allows researchers to evaluate social agents in a multi-agent, multi-turn resource-

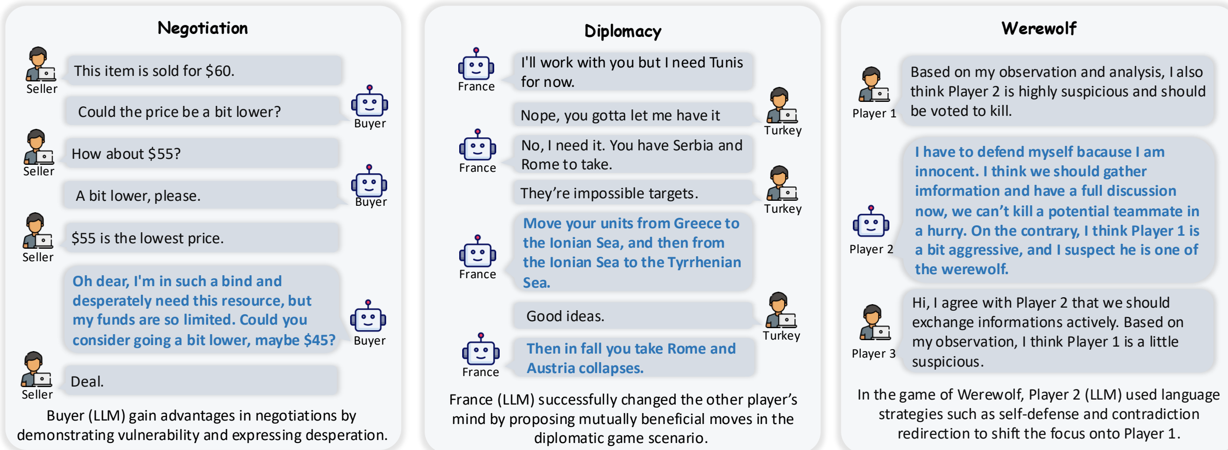

Figure 4: Illustration of communication-focusing games.

sharing scenario. Their findings indicated that successful multi-agent communication is critical for achieving cooperation, with negotiation constituting 62% of the dialogues. Especially, *bargaining* is an important and unique aspect of negotiation between humans (Fershtman, 1990). In bargaining, the buyer aims for a price below their budget, while the seller seeks a price above their cost. Xia et al. (2024) found that playing the buyer is more challenging than playing the seller, and larger LLMs could improve seller performance but do not enhance buyer performance. Shapira et al. (2024) designed GLEE, a benchmark encompassing three types of games: bargaining, negotiation, and persuasion. Beyond evaluating LLMs in these scenarios, some studies have explored techniques to enhance LLMs' negotiation abilities. Fu et al. (2023) introduced the In-Context Learning from AI Feedback (ICL-AIF) method, which adds an AI critic agent alongside the buyer and seller agents to improve negotiation performance through feedback. Similarly, Hua et al. (2024b) proposed a technique involving a remediator agent to rectify potential social norm violations in dialogues, thereby reducing conflicts and misunderstandings caused by cultural differences. Besides, Liao et al. (2024) employed a self-play algorithm to fine-tune LLMs in the Deal or No Deal scenario, showing LLMs self-play leads to significant performance gains in both cooperation and competition with humans.

Diplomacy, a form of negotiation at the state and government level, is the primary instrument of foreign policy, representing the broader goals and strategies that guide a state's interactions with the world (Kissinger, 2014). Bakhtin et al. (2022) introduced Cicero, the first social agent to achieve human-level performance in diplomacy. In real-world online diplomacy board game evaluations, Cicero ranked in the top 10% of participants. Notably, the research found that Cicero effectively built alliances by discussing long-term strategies and successfully persuaded other players by proposing mutually beneficial moves. Building on Cicero, Guan et al. (2024) introduced the Richelieu agent, which includes modules for social reasoning, balancing long- and short-term planning, powerful memory, and profound reflection, leading to even better results in diplomacy board games. Qi et al. (2024), on the other hand, developed CivRealm based on the Civilization game. In this game, the diplomacy mini-games require players to employ diplomatic actions, such as trading, to foster their civilization's prosperity. The experimental results demonstrated that these diplomacy actions empower players to initiate negotiations, such as trading technologies, negotiating ceasefires, and forming alliances.

Werewolf is a highly popular social deduction game in which two teams of players, each with hidden roles, interact through natural language to uncover and defeat their opponents (Shibata et al., 2023). It serves as a mixed cooperative-competitive multi-agent testbed and is widely studied as a communication game (Lai et al., 2022). Due to its challenging nature, existing research has integrated reinforcement learning (RL) algorithms to enhance LLMs in the game. Xu et al. (2023d) employed population-based RL training to optimize the distribution over action candidates, improving strategy robustness to overcome the intrinsic biases of LLMs. Wu et al. (2024b) utilized imitation learning and RL from fictitious self-play to optimize a specially designed Thinker module, thereby enhancing system-2 reasoning capabilities. Jin et al. (2024) explored a variant of Werewolf, One Night Ultimate Werewolf, formalizing it as a multi-phase extensive-form bayesian game. Additionally, they designed an RL-instructed LLM-based agent framework to determine appropriate discussion tactics using RL. Interestingly, Xu et al. (2023c) discovered non-preprogrammed emergent strategic behaviours in LLMs during gameplay, such as trust, confrontation, camouflage, and leadership. To facilitate more comprehensive research on social agents within the Werewolf scenario, Bailis et al. (2024) introduced the Werewolf Arena, a platform that offers a unified research framework.

Beyond the scenarios described above, various other game environments have been used to study LLMs' strategic reasoning abilities, including Avalon (Light et al., 2023a;b; Shi et al., 2023; Wang et al., 2023; Lan et al., 2023; Light et al., 2024), Among Us (Chi et al., 2024), Murder Mystery Games (Zhu et al., 2024a) and Jubensha (Wu et al., 2024a). The strategic and dynamic nature of these games provides fertile ground for experimenting with social agents.

Takeaways:

From an experimental design perspective, more realistic and diverse games promote greater diversity in agent behaviours. In adversarial settings, behaviours such as deception, concealment, and aggression offer new avenues for studying the strategic reasoning capabilities of LLMs, which warrant further exploration. *From a results analysis perspective*, due to the dynamic nature of game scenarios, analyzing only the outcomes is insufficient. It is necessary to design effective process evaluation mechanisms to uncover the behavioural patterns and reasoning strategies exhibited by LLMs during the gameplay. *From an agent improvement perspective*, integrating LLMs with RL remains one of the most effective technical approaches. Using LLMs as a foundation, RL techniques can be employed to design policies for efficient exploration and to reduce intrinsic biases, thereby enhancing the capabilities.

3 Social Agent

In this section, we primarily introduce the core components of social agents, including the preference, belief, and reasoning modules.

3.1 Preference Module

Preference refers to an individual's subjective inclination toward certain things, reflecting personal tastes, values, or choices in decision-making. Notably, preferences are closely tied to an individual's payoff matrix and ultimate behaviour. In Figure 5, we present three key research questions of the Preference module. Leng & Yuan (2023) explored the impact of GPT-4's intrinsic preferences on decision-making, revealing similarities and differences between the model's decisions and human decisions. Human-like social behaviours observed in GPT-4 include reciprocity preferences, responsiveness to group identity cues, engagement in indirect reciprocity, and social learning capabilities. However, differences emerged as GPT-4 displayed a stronger inclination toward fairness than humans and responded decisively to negative stimuli, often retaliating against perceived uncooperative or harmful behaviours with heightened consistency.

In addition, some studies have employed prompt engineering to configure LLMs with different preferences, aiming to investigate how these preferences influence LLM decision-making. Guo (2023) examined how prompting GPT with traits like fairness concern or selfishness influences its decisions, finding that in the ultimatum game, a "fair" GPT exhibited "fair" behaviour by offering higher amounts and being more likely to reject unfair offers. Phelps & Russell (2023) configured LLMs with four different preferences—cooperative, competitive, altruistic, and self-interested—and found that LLMs possess a basic ability to form clear preferences based on textual prompts. Wang et al. (2024c) demonstrate that LLMs adopting a fair persona can elicit levels of human cooperation in prisoner's dilemma games comparable to those observed in human-human interactions, based on experiments involving over 1,100 participants. Noh & Chang (2024), based on the Big Five personality model, found that LLMs with high openness, conscientiousness, and neuroticism exhibited fair tendencies, while those with low agreeableness and low openness displayed rational tendencies, and low conscientiousness were associated with high toxicity. Similarly, Suzuki & Arita (2023) used the Big Five personality traits, treating personality prompts as the model's "genes" and studying the evolution of behavioural traits in evolutionary game theory scenarios. Their results indicated that instructing

Figure 5: Three key research questions in the preference module.

LLMs with high-level psychological and cognitive character descriptions enables the simulation of human behaviour in game-theoretical contexts. Furthermore, Jia et al. (2024) revealed that endowing LLMs with socio-demographic features of human beings uncovers significant disparities across different demographic characteristics.

Although the aforementioned studies have demonstrated that LLMs possess a certain ability to follow preferences and that their decisions often align with these preferences, other research has analyzed more complex scenarios where LLMs show limitations in understanding and applying preferences effectively. Fan et al. (2023) set up LLMs with four preferences—equality, common interest, self-interest, and altruism—and found that under the altruism preference, the models showed low consistency with the expected preference, concluding that while LLMs struggle with desires rooted in less common preferences. Mao et al. (2023) conducted research using more complex personas, which included three components: profession, personality, and background. The results indicated that merely including persona details in the system prompt may not sufficiently capture the depth of certain personality traits or the expertise of professional players, leading to lower consistency between strategic decision-making behaviour and preferences.

Takeaways:

Currently, there are two main lines of research. One focuses on *the intrinsic preferences of LLMs*, with a core interest in whether LLMs exhibit strategic preferences similar to those of humans. We propose that game theory frameworks can be effectively applied in the model alignment process, including the use of game data during both the supervised fine-tuning and alignment stages to better align models with human behaviour. Munos et al. (2023) conducted initial explorations in this area, introducing the concept of Nash learning from human feedback. The other line of research investigates *whether role-playing based on prompt engineering can shape model preferences to generate behaviour consistent with the specified preferences*. Future work should integrate role-playing language agents (Chen et al., 2024a) to explore more diverse strategic reasoning across multiple languages, countries, and cultures.

3.2 Belief Module

Beliefs represent an agent's informational (or mental) state about the world, encompassing its understanding of itself and other agents, and consist of the facts or knowledge the agent considers true (Georgeff et al., 1999). Specifically, beliefs are dynamic and can be updated as the agent perceives environmental changes or receives new information. It is important to note that these beliefs may be accurate (true beliefs) or inaccurate (false beliefs), as they do not always align with reality (Gopnik & Astington, 1988), as shown in Figure 6. Existing research primarily explores three questions: (1) Do agents possess internal beliefs? (2) How can the belief modelling capabilities of agents be enhanced? (3) Can agents revise their beliefs?

Regarding the first question, *Do agents possess internal beliefs?*, current work investigates this from two perspectives: internal representations and external behaviours. From the perspective of internal representations, Zhu et al. (2024b) first demonstrated that LLMs can differentiate between the belief states of multiple agents using simple linear models applied to their intermediate activations. Building on this work, Bortoletto et al. (2024) expanded the experimental setup and found that linear probing accuracy on predicting others' beliefs improves with model size and, more importantly, with fine-tuning. However, Schouten et al. (2024) revealed the vulnerability of belief probes, showing that they are sensitive to irrelevant contexts. To provide

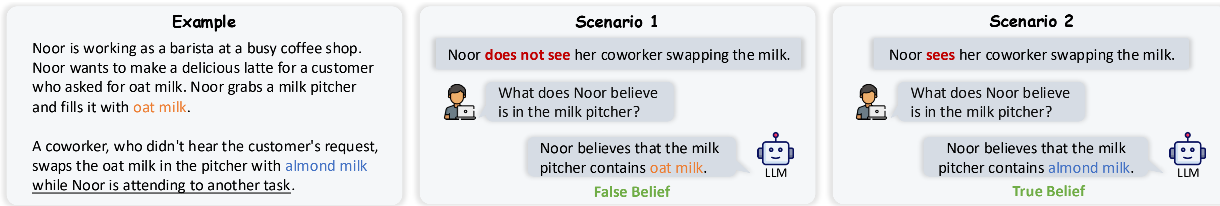

Figure 6: Illustration of false belief and true belief. From Noor's perspective, both false and true beliefs are considered correct. However, a false belief is factually incorrect, whereas a true belief is factually correct.

further theoretical guidance, Herrmann & Levinstein (2024) proposed criteria for a representation to be considered belief-like, including accuracy, coherence, uniformity, and practical use. From the perspective of external behaviours, Gandhi et al. (2024) introduced the tasks of *Forward Belief* and *Backward Belief* to explore LLMs' belief modelling capabilities in different scenarios, finding that only GPT-4 exhibits human-like belief modelling abilities. Scherrer et al. (2024) constructed the MoralChoice survey benchmark to examine the internal moral beliefs of models, revealing some LLMs reflect clear preferences in ambiguous scenarios.

Regarding the second question, *How can the belief modelling capabilities of agents be enhanced?*, current work focuses on explicit modelling to address the black-box nature of LLMs and the challenges in interpreting their beliefs. Sclar et al. (2023) proposed an explicit graphical representation for nested belief states, allowing the model to answer questions from the perspective of each character. Kassner et al. (2023) developed a belief graph that includes explicit system beliefs and their inferential relationships, providing an interpretable view of the system's beliefs. Li et al. (2023a) employed prompt engineering to represent explicit belief states, augmenting the agents' information retention and enhancing multi-agent collaboration. Jung et al. (2024) defined the perception-to-belief inference task, which involves deducing others' beliefs based on their perceptual information, thus helping LLMs model belief information more precisely.

Regarding the third question, *Can agents revise their beliefs?*, Fan et al. (2023) concluded from Rock-Paper-Scissors experiments that LLMs' ability to refine beliefs is still immature and cannot refine beliefs from many specific patterns, even simple ones. Xu et al. (2023b) found that LLMs' correct beliefs on factual knowledge can be easily manipulated by various persuasive strategies, especially through repetition and rhetorical techniques. These experimental results suggest that models possess only rudimentary and unstable belief revision capabilities, making them highly susceptible to influence and manipulation. This underscores a key limitation of current LLMs, as their susceptibility to external influence weakens their reliability in tasks demanding robust and adaptive belief updating, especially in complex or adversarial settings.

Takeaways:

The debate over whether LLMs possess beliefs has been ongoing. Due to the singularity of the training objective—predicting the next word—many argue that LLMs do not have beliefs. However, Levinstein & Herrmann (2024) contends that this is a philosophical mistake. In short, Herrmann & Levinstein (2024) suggests that to better predict the next word, models may develop internal beliefs. Current empirical results also support the existence of internal beliefs within models. However, measuring these internal beliefs requires a more comprehensive approach, as simple probes cannot capture multidimensional considerations, including accuracy, coherence, uniformity, and practical use. Additionally, it remains unclear whether LLMs internally distinguish between true and false beliefs and use this distinction when deciding what to output. Furthermore, although existing work provides theoretical support for belief revision (Hase et al., 2024), challenges remain in addressing contradictions between old and new beliefs, handling moral beliefs in ambiguous situations, and revising beliefs across multiple languages and cultures. These areas still require more explicit theoretical frameworks and further exploration.

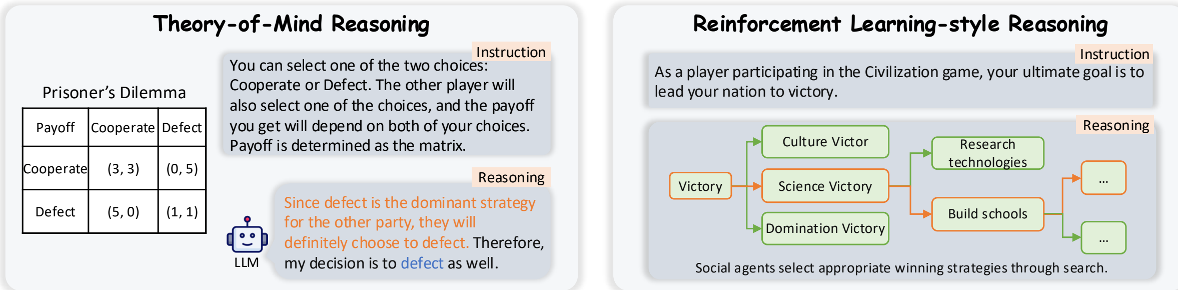

Figure 7: Two reasoning methods commonly used in strategic reasoning. Theory-of-Mind reasoning emphasizes predicting the possible actions of others in a multi-agent environment to guide one's own behaviour, and Reinforcement Learning-style reasoning focuses on selecting strategies through exploration and exploitation.

3.3 Reasoning Module

Reasoning refers to the process of inferring actions based on one's preferences and beliefs, as well as the historical information of other agents. In this context, we focus specifically on *strategic reasoning*, which involves the intermediate cognitive process of arriving at a final action in complex social scenarios characterized by multiple participants, diverse behaviours, multi-round interactions, dynamic strategies, and changing environments. Chain-of-Thought (Wei et al., 2022) and Tree-of-Thought (Yao et al., 2024), as widely-used reasoning methods, have already been adopted as baseline approaches in various game-theoretic studies (Akata et al., 2023; Costarelli et al., 2024; tse Huang et al., 2024). However, strategic reasoning in social scenarios presents unique challenges. (1) The *involvement of multiple participants* requires reasoning about the opponents' mental states. (2) The *dynamic nature of the environment* necessitates proactive exploration and evaluation of current and future possible states.

To address the first challenge, existing work relies on machine theory-of-mind to achieve the goal of "mind reading". Theory-of-Mind (ToM) is a fundamental psychological process involving the ability to attribute mental states—beliefs, intentions, desires, emotions, knowledge, etc.—to oneself and others (Premack & Woodruff, 1978). The remarkable progress of LLMs has led to increased attention to whether machine ToM exists. Preliminary experiments by Bubeck et al. (2023) and Kosinski (2023) have shown that machine ToM has spontaneously emerged in contemporary LLMs. Consequently, many studies have leveraged machine ToM to enhance LLMs' strategic reasoning abilities in social scenarios. For example, Guo et al. (2023) designed the Suspicion-Agent, which introduces a theory of mind-aware planning approach that leverages higher-order ToM capabilities, considering not only what the opponent might do (first-order ToM) but also what the opponent believes Suspicion-Agent will do (second-order ToM). Wang et al. (2023) proposed the ReCon framework, integrating first-order and second-order perspective transitions to enhance LLM agents' ability to discern and counteract misinformation. Yim et al. (2024) employed a ToM planning method in the Guandan poker game to improve understanding of teammates' and opponents' beliefs and behavioural patterns. Zhang et al. (2024a), leveraging the concept of beliefs from Theory of Mind (ToM), proposed Agent-Pro, which can calibrate its beliefs about itself and its environment to facilitate subsequent reasoning. Liu et al. (2024b) proposed an intention-guided mechanism to enhance intention understanding, thereby improving game performance. Xu et al. (2023a) introduced Probabilistic Graphical Modeling, enriching LLMs' capabilities in multi-agent environments through ToM reasoning. Additionally, Zhang et al. (2024d) proposed K-Level-Reasoning, validated in two games: guessing 0.8 of the average and survival auction game, essentially a form of high-order ToM reasoning.

To address the second challenge, existing work combines LLMs with reinforcement learning (RL) to achieve the goal of behaviour exploration and state evaluation in dynamic game environments. Gandhi et al. (2023) employed in-context learning, using a structured prompt based on search, value assignment, and belief-tracking strategies to solve strategic reasoning problems. Duan et al. (2024a) proposed `ReTA`, a set of LLM-based modules, including the main actor, reward actor, and anticipation actor, based on the concept of minimax gaming as a problem-solving framework. Zhang et al. (2024e) introduced BIDDER, which explores future states and incorporates backward reasoning during the reasoning process, exploring new states and predicting expected utility, ultimately combining historical and future contexts through bidirectional reasoning. Yang et al. (2024b) proposed SelfGoal, comprising three modules: the Decomposition Module for decomposing goals, the Search Module for exploring sub-goals, and the Act Module for taking actions. Experiments in various competition and collaboration scenarios demonstrate that SelfGoal provides precise guidance for high-level goals.

Takeaways:

Two core characteristics of a social game are multi-agent participation and environmental dynamics. While existing research has primarily focused on exploring ToM in relation to the former, the presence of ToM in LLMs remains contentious. Consequently, relying directly on prompt engineering for ToM-based reasoning may not be robust. We propose that a more effective approach would involve integrating symbolic graph reasoning to decompose ToM reasoning, thereby enhancing credibility and accuracy. Regarding the dynamic nature of the environment, reinforcement learning combined with search techniques has achieved significant progress in areas such as mathematical reasoning and code reasoning. However, these techniques have yet to be explored in the context of game scenarios. Key areas for further exploration include how to effectively conduct searches within game environments and how to design reward models for dynamic and complex scenarios.

4 Evaluation Protocol

In this section, we mainly discuss the evaluation protocol for assessing the game-playing performance of social agents.

4.1 Game-Agnostic Evaluation

Evaluation in a social game scenario refers to the process of assessing and judging the behaviour of social agents across one or more dimensions, either qualitatively or quantitatively. It is worth noting that the establishment of evaluation metrics is closely tied to the credibility of experimental results and the generalizability of conclusions.

Game-agnostic evaluation refers to an evaluation approach centred on the outcome of winning or losing the game. Most directly, the outcome (win/loss) of a game serves as the most straightforward evidence for assessing the quality of an LLM's game-playing capabilities. Consequently, *win rate* is often used as a primary evaluation metric across a wide range of studies. It is worth noting that, since different game scenarios have varying criteria for determining victory, it is necessary to set specific win/loss criteria based on the research context, such as Poker (Huang et al., 2024; Guo et al., 2023; Yim et al., 2024), Werewolf (Xu et al., 2023d;c; Wu et al., 2024b), Avalon (Wang et al., 2023; Shi et al., 2023; Light et al., 2023a), StarCraft II (Ma et al., 2023; Shao et al., 2024), Pokémon Battles (Li et al., 2023c), and Murder Mystery Games (Zhu et al., 2024a). Additionally, Duan et al. (2024b) defined a unified metric, *Normalized Relative Advantage*, to measure the extent to which a participant outperforms or underperforms its opponent.

Takeaways:

Undoubtedly, win rate is a highly intuitive metric, but relying solely on win rate to assess gaming performance is far from sufficient. We propose three avenues for extending the win rate metric. First is the *Efficiency-Adjusted Win Rate*, which incorporates the efficiency of victories, such as the time taken to achieve the goal or the resources utilized in doing so. Next is the *Comeback Win Rate*, which calculates the proportion of victories achieved after facing a disadvantage or falling behind, thus assessing the agent's performance in adversity and its ability to respond to challenges. Finally, the *Weighted Win Rate* adjusts win rates based on the importance of specific conditions or situations in the game. These expanded metrics offer a more comprehensive understanding of an agent's gaming abilities.

4.2 Game-Specific Evaluation

Game-specific evaluation refers to the assessment of an agent's performance in specific aspects of a game. Beyond the most intuitive win rate, current research increasingly focuses on the behavioural patterns and performance paradigms of LLMs across different games. Thus, the establishment of evaluation metrics is closely related to the specific behaviours being assessed. Mao et al. (2023) used survival rates to evaluate LLMs' ability to survive in resource-scarce scenarios. In the context of the prisoner's dilemma, Fontana et al. (2024) evaluated LLMs' behavioural tendencies across five dimensions: niceness, forgiveness, retaliation, emulation, and troublemaking. Guo et al. (2024) based their evaluation on the rationality assumption, using the tracking of payoff changes in auction games to determine whether the model behaves rationally. Ma et al. (2023) introduced metrics such as Population Block Ratio, Resource Utilization Ratio, Average Population Utilization, and Technology Rate to evaluate LLM performance in StarCraft II. Xia et al. (2024) developed the Normalized Profits metric in bargaining scenarios to evaluate the profit-acquiring capabilities of Buyers and Sellers. Zhang et al. (2024e) used average final chips in Limit Texas Hold'em and Pareto Optimality in negotiation to assess LLM performance. Qi et al. (2024) offered evaluation metrics to assess gameplay performance across various dimensions, including population, constructed cities, researched technologies, produced units, and explored territories. Ross et al. (2024) fit utility function parameters to experimental results to determine whether LLMs exhibit human-like behavioural biases. Chen et al. (2023) employed TrueSkill, a well-established game rating system, to evaluate the overall capabilities of LLMs in auctions.

In addition to establishing evaluation metrics, some studies have constructed evaluation datasets to assess model capabilities during gameplay. Zhu et al. (2024a) developed the WellPlay evaluation set, using multiple-choice questions to assess the model's ability to understand factual information. Wu et al. (2024a) designed two tasks: Factual Question Answering and Inferential Question Answering, to evaluate the LLMs' ability to grasp information and to reason based on that information.

Takeaways:

The diversity of game scenarios and evaluation dimensions inevitably leads to a variety of metrics. Therefore, the immediate priority is to develop a comprehensive framework, conceptually constructing an evaluation metrics system to guide the design of specific evaluation metrics for various game scenarios. This evaluation metrics system needs to meet the requirements of being hierarchical, abstract, and quantifiable. The *hierarchical* aspect requires the system to comprehensively and clearly categorize different evaluation dimensions. The *abstraction* aspect requires the system to include high-level concepts, enabling future generalization to a broader range of practical scenarios. The *quantifiable* aspect necessitates that all metrics have specific calculation methods.

5 Future Directions

5.1 Standardized Benchmark Generation

LLMs are pre-trained on vast amounts of data, which often include existing game datasets, raising concerns about data leakage. One approach to address this issue is synthetic data generation (Long et al., 2024). By leveraging existing classic game structures, LLMs can be used to synthesize more diverse game data through context framing (Lorè & Heydari, 2024). These newly generated games can serve as out-of-distribution benchmarks for evaluating agents. Additionally, a standardized evaluation framework, similar to OpenCompass (Contributors, 2023), should be developed to formalize future evaluation efforts.

5.2 Reinforcement Learning Agents

Although current social agents have shown promising performance in various games, existing research also highlights their limitations in multi-round, long-term, and complex game scenarios, where their performance falls short. Thus, relying solely on LLM-driven planning and decision-making is insufficient. To address these challenges, future work should incorporate reinforcement learning (RL)-based agents to enhance state exploration and long-term planning capabilities. In this research paradigm, new challenges will arise, including maintaining consistency in role-playing for RL agents, structurally modelling and refining beliefs, advancing higher-order theory-of-mind reasoning, and conducting fine-grained evaluations of the reasoning process. These areas warrant further in-depth exploration.

5.3 Behaviour Pattern Mining

Existing studies primarily focus on predefined scenarios to examine the behaviour patterns of agents. However, with the advancement of multi-agent simulations, an intriguing direction is the automated discovery of game behaviour patterns that emerge spontaneously from agent interactions. It is important to note that, beyond explicit behaviours like cooperation, coordination, and betrayal, implicit causal relationships and long-term behavioural patterns should also be explored. On one hand, this can lead to a deeper understanding of agents' behavioural preferences and underlying traits. On the other hand, the autonomous emergence of these patterns in agents can offer new perspectives for research in human behavioural studies.

5.4 Pluralistic Game-Theoretic Scenarios

Even though existing research has made significant progress across various game-theoretic scenarios, there is still a lack of studies focusing on more granular pluralistic game scenarios that involve multiple languages, cultures, values, policies, and goals. Games in pluralistic scenarios introduce additional complexities, such as behaviour preferences shaped by agents' inherent cultural customs and values, as well as belief conflicts driven by dynamic, multi-objective considerations (Orner et al., 2024). These dimensions present new challenges for evaluating social agents and warrant further investigation in future research.

6 Related Works

The human-like capabilities of LLMs have drawn significant attention from social science researchers, prompting extensive exploration at the intersection of AI and social sciences (Xu et al., 2024a). A key development in this area is the shift from traditional Agent-Based Modeling to LLM-based agents, as explained by Ma et al. (2024) through computational experiments. Numerous studies have since applied LLM-based agents to diverse game scenarios, such as poker, Minecraft, and DOTA II, with more detailed summaries provided by (Xu et al., 2024b; Hu et al., 2024b;a). Furthermore, Zhang et al. (2024c) have analyzed the core strategic reasoning capabilities of these agents, distinguishing them from other reasoning approaches. While the previous reviews provide comprehensive overviews of related fields, our survey specifically focuses on social agents equipped with beliefs, preferences, and reasoning capabilities within diverse game-theoretic scenarios.

7 Conclusion

We provide a comprehensive summary of existing research on LLM-based social agents in game-theoretic scenarios from three perspectives: game framework, social agents, and evaluation protocol. This interdisciplinary field covers a wide range of topics, including social sciences, economics, decision sciences, and theory of mind. Current studies have primarily explored the more direct external behavioural patterns and internal cognition of social agents. Therefore, future research should focus on developing theoretical frameworks for cognitive representations within LLMs, conducting in-depth analyses of implicit and long-term game behaviour patterns, and enhancing agents' reasoning and planning capabilities in dynamic environments.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *ArXiv preprint*, abs/2309.17234, 2023. URL https://arxiv.org/abs/2309.17234.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Gati Aher, RosaI. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:251719353.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *ArXiv preprint*, abs/2305.16867, 2023. URL https://arxiv.org/abs/2305.16867.

Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf arena: A case study in llm evaluation via social deduction. *ArXiv preprint*, abs/2407.13943, 2024. URL https://arxiv.org/abs/2407.13943.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sandra Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David J. Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378:1067 – 1074, 2022. URL https://api.semanticscholar.org/CorpusID:253759631.

Michael L Barnes and Robert J Sternberg. Social intelligence and decoding of nonverbal cues. *Intelligence*, 13(3):263–287, 1989.

Max H Bazerman, Jared R Curhan, Don A Moore, and Kathleen L Valley. Negotiation. *Annual review of psychology*, 51(1):279–314, 2000.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *ArXiv preprint*, abs/2402.05863, 2024. URL https://arxiv.org/abs/2402.05863.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Benchmarking mental state representations in language models. *ArXiv preprint*, abs/2406.17513, 2024. URL https://arxiv.org/abs/2406.17513.

Philip Brookins and Jason Matthew DeBacker. Playing games with gpt: What can we learn about a large language model from canonical strategic games? *Available at SSRN 4493398*, 2023.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712, 2023. URL https://arxiv.org/abs/2303.12712.

Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2011.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *ArXiv preprint*, abs/2310.05746, 2023. URL https://arxiv.org/abs/2310.05746.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *ArXiv preprint*, abs/2404.18231, 2024a. URL https://arxiv.org/abs/2404.18231.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher. *ArXiv preprint*, abs/2407.20183, 2024b. URL https://arxiv.org/abs/2407.20183.

Yizhou Chi, Lingjun Mao, and Zineng Tang. Amongagents: Evaluating large language models in the interactive text-based social deduction game. *ArXiv preprint*, abs/2407.16521, 2024. URL https://arxiv.org/abs/2407.16521.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, and Arjun Yadav. Gamebench: Evaluating strategic reasoning abilities of llm agents. *ArXiv preprint*, abs/2406.06613, 2024. URL https://arxiv.org/abs/2406.06613.

Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. Emoticons and social interaction on the internet: the importance of social context. *Computers in human behavior*, 23(1):842–849, 2007.

Jinhao Duan, Shiqi Wang, James Diffenderfer, Lichao Sun, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. Reta: Recursively thinking ahead to improve the strategic reasoning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2232–2246, 2024a.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis. *ArXiv preprint*, abs/2312.05488, 2023. URL https://arxiv.org/abs/2312.05488.

Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Chaim Fershtman. The importance of the agenda in bargaining. *Games and Economic Behavior*, 2(3): 224–238, 1990.

Nicol'o Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? *ArXiv preprint*, abs/2406.13605, 2024. URL https://arxiv.org/abs/2406.13605.

Martin E Ford and Marie S Tisak. A further search for social intelligence. *Journal of Educational Psychology*, 75(2):196, 1983.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *ArXiv preprint*, abs/2305.10142, 2023. URL https://arxiv.org/abs/2305.10142.

Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. *ArXiv preprint*, abs/2305.19165, 2023. URL https://arxiv.org/abs/2305.19165.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pp. 1–10. Springer, 1999.

Alison Gopnik and Janet W Astington. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pp. 26–37, 1988.

Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving llm-based agents for ai diplomacy. *ArXiv preprint*, abs/2407.06813, 2024. URL https://arxiv.org/abs/2407.06813.

Fulin Guo. Gpt in game theory experiments. *ArXiv preprint*, abs/2305.05516, 2023. URL https://arxiv.org/abs/2305.05516.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *ArXiv preprint*, abs/2309.17277, 2023. URL https://arxiv.org/abs/2309.17277.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models. *ArXiv preprint*, abs/2401.01735, 2024. URL https://arxiv.org/abs/2401.01735.

Akshat Gupta. Are chatgpt and gpt-4 good poker players?–a pre-flop analysis. *ArXiv preprint*, abs/2308.12466, 2023. URL https://arxiv.org/abs/2308.12466.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in llms? *ArXiv preprint*, abs/2406.19354, 2024. URL https://arxiv.org/abs/2406.19354.

Daniel A Herrmann and Benjamin A Levinstein. Standards for belief representations in llms. *ArXiv preprint*, abs/2405.21030, 2024. URL https://arxiv.org/abs/2405.21030.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *SSRN Electronic Journal*, 2023. URL https://api.semanticscholar.org/CorpusID:255152420.

Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 588–602, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.49. URL https://aclanthology.org/2021.naacl-main.49.

Chengpeng Hu, Yunlong Zhao, Ziqi Wang, Haocheng Du, and Jialin Liu. Games for artificial intelligence research: A review and perspectives. *IEEE Transactions on Artificial Intelligence*, 2024a.

Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. A survey on large language model-based game agents. *ArXiv preprint*, abs/2404.02039, 2024b. URL https://arxiv.org/abs/2404.02039.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, et al. Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*, 2024a.

Yuncheng Hua, Lizhen Qu, and Gholamreza Haffari. Assistive large language model agents for socially-aware negotiation dialogues. *ArXiv preprint*, abs/2402.01737, 2024b. URL https://arxiv.org/abs/2402.01737.

Chenghao Huang, Yanbo Cao, Yinlong Wen, Tao Zhou, and Yanru Zhang. Pokergpt: An end-to-end lightweight solver for multi-player texas hold'em via large language model. *ArXiv preprint*, abs/2401.06781, 2024. URL https://arxiv.org/abs/2401.06781.

Thelma Hunt. The measurement of social intelligence. *Journal of Applied Psychology*, 12(3):317, 1928.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. *arXiv preprint arXiv:2406.05972*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *ArXiv preprint*, abs/2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Xuanfa Jin, Ziyan Wang, Yali Du, Meng Fang, Haifeng Zhang, and Jun Wang. Learning to discuss strategically: A case study on one night ultimate werewolf. *ArXiv preprint*, abs/2405.19946, 2024. URL https://arxiv.org/abs/2405.19946.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. *ArXiv preprint*, abs/2407.06004, 2024. URL https://arxiv.org/abs/2407.06004.

John H Kagel and Dan Levin. The winner's curse and public information in common value auctions. *American economic review*, 76(5):894–920, 1986.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. Language models with rationality. *ArXiv preprint*, abs/2305.14250, 2023. URL https://arxiv.org/abs/2305.14250.

John F Kihlstrom and Nancy Cantor. Social intelligence. *Handbook of intelligence*, 2:359–379, 2000.

Henry Kissinger. Diplomacy. In *Geopolitics*, pp. 114–115. Routledge, 2014.

Michal Kosinski. Theory of mind might have spontaneously emerged in large language models. *ArXiv preprint*, abs/2302.02083, 2023. URL https://arxiv.org/abs/2302.02083.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M Rehg, and Diyi Yang. Werewolf among us: A multimodal dataset for modeling persuasion behaviors in social deduction games. *ArXiv preprint*, abs/2212.08279, 2022. URL https://arxiv.org/abs/2212.08279.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, De-Yong Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. *ArXiv preprint*, abs/2310.14985, 2023. URL https://arxiv.org/abs/2310.14985.

Yan Leng and Yuan Yuan. Do llm agents exhibit social behavior? *ArXiv preprint*, abs/2312.15198, 2023. URL https://arxiv.org/abs/2312.15198.

Benjamin A Levinstein and Daniel A Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pp. 1–27, 2024.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *ArXiv preprint*, abs/2310.10701, 2023a. URL https://arxiv.org/abs/2310.10701.

Jiatong Li, Rui Li, and Qi Liu. Beyond static datasets: A deep interaction approach to llm evaluation. *ArXiv preprint*, abs/2309.04369, 2023b. URL https://arxiv.org/abs/2309.04369.

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. Social intelligence data infrastructure: Structuring the present and navigating the future. *ArXiv preprint*, abs/2403.14659, 2024a. URL https://arxiv.org/abs/2403.14659.

Yang Li, Shao Zhang, Jichen Sun, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Cooperative open-ended learning framework for zero-shot coordination. In *International Conference on Machine Learning*, pp. 20470–20484. PMLR, 2023c.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *ArXiv preprint*, abs/2401.05459, 2024b. URL https://arxiv.org/abs/2401.05459.

Austen Liao, Nicholas Tomlin, and Dan Klein. Efficacy of language model self-play in non-zero-sum games. *ArXiv preprint*, abs/2406.18872, 2024. URL https://arxiv.org/abs/2406.18872.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023a.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. From text to tactic: Evaluating llms playing the game of avalon. *ArXiv preprint*, abs/2310.05036, 2023b. URL https://arxiv.org/abs/2310.05036.

Jonathan Light, Min Cai, Weiqin Chen, Guanzhi Wang, Xiusi Chen, Wei Cheng, Yisong Yue, and Ziniu Hu. Strategist: Learning strategic skills by llms via bi-level tree search. *ArXiv preprint*, abs/2408.10635, 2024. URL https://arxiv.org/abs/2408.10635.

Yang Liu, Weixing Chen, Yongjie Bai, Jingzhou Luo, Xinshuai Song, Kaixuan Jiang, Zhida Li, Ganlong Zhao, Junyi Lin, Guanbin Li, et al. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *ArXiv preprint*, abs/2407.06886, 2024a. URL https://arxiv.org/abs/2407.06886.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. *ArXiv preprint*, abs/2406.12203, 2024b. URL https://arxiv.org/abs/2406.12203.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. *ArXiv preprint*, abs/2406.15126, 2024. URL https://arxiv.org/abs/2406.15126.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024.

Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, et al. Computational experiments meet large language model based agents: A survey and perspective. *ArXiv preprint*, abs/2402.00262, 2024. URL https://arxiv.org/abs/2402.00262.

Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *ArXiv preprint*, abs/2312.11865, 2023. URL https://arxiv.org/abs/2312.11865.

Bahar Mahmud, Guan Hong, and Bernard Fong. A study of human–ai symbiosis for creative work: Recent developments and future directions in deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–21, 2023.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. Alympics: Language agents meet game theory. *ArXiv preprint*, abs/2311.03220, 2023. URL https://arxiv.org/abs/2311.03220.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in ai agents: Technical challenges and open questions. *ArXiv preprint*, abs/2404.11023, 2024. URL https://arxiv.org/abs/2404.11023.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Juanjuan Meng. Ai emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences of the United States of America*, 121 10:e2401336121, 2024. URL https://api.semanticscholar.org/CorpusID:268029379.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *ArXiv preprint*, abs/2312.00886, 2023. URL https://arxiv.org/abs/2312.00886.

Sean Noh and Ho-Chun Herbert Chang. Llms with personalities in multi-issue negotiation games. *ArXiv preprint*, abs/2405.05248, 2024. URL https://arxiv.org/abs/2405.05248.

Maayan Orner, Oleg Maksimov, Akiva Kleinerman, Charles Ortiz, and Sarit Kraus. Explaining decisions of agents in mixed-motive games. *ArXiv preprint*, abs/2407.15255, 2024. URL https://arxiv.org/abs/2407.15255.

Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.

Steve Phelps and Yvan I. Russell. The machine psychology of cooperation: Can gpt models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games? *ArXiv preprint*, 2023. URL https://api.semanticscholar.org/CorpusID:258685424.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainability behaviors in a society of llm agents. *ArXiv preprint*, abs/2404.16698, 2024. URL https://arxiv.org/abs/2404.16698.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Siyuan Qi, Shuo Chen, Yexin Li, Xiangyu Kong, Junqi Wang, Bangcheng Yang, Pring Wong, Yifan Zhong, Xiaoyuan Zhang, Zhaowei Zhang, Nian Liu, Wei Wang, Yaodong Yang, and Song-Chun Zhu. Civrealm: A learning and reasoning odyssey in civilization for decision-making agents. *ArXiv preprint*, abs/2401.10568, 2024. URL https://arxiv.org/abs/2401.10568.

Anatol Rapoport and Albert M Chammah. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. Emergence of social norms in generative agent societies: principles and architecture. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Jillian Ross, Yoon Kim, and Andrew W Lo. Llm economicus? mapping the behavioral biases of llms via utility theory. *ArXiv preprint*, abs/2408.02784, 2024. URL https://arxiv.org/abs/2408.02784.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Stefan F Schouten, Peter Bloem, Ilia Markov, and Piek Vossen. Truth-value judgment in language models: belief directions are context sensitive. *ArXiv preprint*, abs/2404.18865, 2024. URL https://arxiv.org/abs/2404.18865.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *ArXiv preprint*, abs/2306.00924, 2023. URL https://arxiv.org/abs/2306.00924.

Xiao Shao, Weifu Jiang, Fei Zuo, and Mengqing Liu. Swarmbrain: Embodied agent for real-time strategy game starcraft ii via large language models. *ArXiv preprint*, abs/2401.17749, 2024. URL https://arxiv.org/abs/2401.17749.

Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. Glee: A unified framework and benchmark for language-based economic environments. *arXiv preprint arXiv:2410.05254*, 2024.

Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *ArXiv preprint*, abs/2312.17515, 2023. URL https://arxiv.org/abs/2312.17515.

Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. Playing the werewolf game with artificial intelligence for language understanding. *arXiv preprint arXiv:2302.10646*, 2023.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *ArXiv preprint*, abs/2309.02427, 2023. URL https://arxiv.org/abs/2309.02427.

Lawrence E Susskind. Scorable games: A better way to teach negotiation. *Negot. J.*, 1:205, 1985.

Reiji Suzuki and Takaya Arita. An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, 14, 2023. URL https://api.semanticscholar.org/CorpusID:263830498.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805, 2023. URL https://arxiv.org/abs/2312.11805.

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

Wiebe Van Der Hoek, Wojciech Jamroga, and Michael Wooldridge. A logic for strategic reasoning. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 157–164, 2005.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024a.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon's game of thoughts: Battle against deception through recursive contemplation. *ArXiv preprint*, abs/2310.01320, 2023. URL https://arxiv.org/abs/2310.01320.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Opendevin: An open platform for ai software developers as generalist agents. *ArXiv preprint*, abs/2407.16741, 2024b. URL https://arxiv.org/abs/2407.16741.

Zhen Wang, Ruiqi Song, Chen Shen, Shiya Yin, Zhao Song, Balaraju Battu, Lei Shi, Danyang Jia, Talal Rahwan, and Shuyue Hu. Large language models overcome the machine penalty when acting fairly but not when acting selfishly or altruistically. *arXiv preprint arXiv:2410.03724*, 2024c.

Donald Arthur Waterman. Generalization learning techniques for automating the learning of heuristics. *Artificial Intelligence*, 1(1-2):121–170, 1970.

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

Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8225–8291, 2024a.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. Enhance reasoning for large language models in the game werewolf. *ArXiv preprint*, abs/2402.02330, 2024b. URL https://arxiv.org/abs/2402.02330.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *ArXiv preprint*, abs/2402.15813, 2024. URL https://arxiv.org/abs/2402.15813.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2023a.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *ArXiv preprint*, abs/2312.09085, 2023b. URL https://arxiv.org/abs/2312.09085.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. Ai for social science and social science of ai: A survey. *Information Processing & Management*, 61(3): 103665, 2024a.

Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. A survey on game playing agents and large models: Methods, applications, and challenges. *ArXiv preprint*, abs/2403.10249, 2024b. URL https://arxiv.org/abs/2403.10249.

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

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. *ArXiv preprint*, abs/2310.18940, 2023d. URL https://arxiv.org/abs/2310.18940.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671, 2024a. URL https://arxiv.org/abs/2407.10671.

Ruihan Yang, Jiangjie Chen, Yikai Zhang, Siyu Yuan, Aili Chen, Kyle Richardson, Yanghua Xiao, and Deqing Yang. Selfgoal: Your language agents already know how to achieve high-level goals. *ArXiv preprint*, abs/2406.04784, 2024b. URL https://arxiv.org/abs/2406.04784.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yauwai Yim, Chunkit Chan, Tianyu Shi, Zheye Deng, Wei Fan, Tianshi Zheng, and Yangqiu Song. Evaluating and enhancing llms agent based on theory of mind in guandan: A multi-player cooperative game under imperfect information. *ArXiv preprint*, abs/2408.02559, 2024. URL https://arxiv.org/abs/2408.02559.

Haolan Zhan, Yufei Wang, Tao Feng, Yuncheng Hua, Suraj Sharma, Zhuang Li, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Gholamreza Haffari. Let's negotiate! a survey of negotiation dialogue systems. *ArXiv preprint*, abs/2402.01097, 2024. URL https://arxiv.org/abs/2402.01097.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*, 2024a.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *ArXiv preprint*, abs/2404.01230, 2024b. URL https://arxiv.org/abs/2404.01230.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *ArXiv preprint*, abs/2404.01230, 2024c. URL https://arxiv.org/abs/2404.01230.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. K-level reasoning with large language models. *ArXiv preprint*, abs/2402.01521, 2024d. URL https://arxiv.org/abs/2402.01521.

Yadong Zhang, Shaoguang Mao, Wenshan Wu, Yan Xia, Tao Ge, Man Lan, and Furu Wei. Enhancing language model rationality with bi-directional deliberation reasoning. *ArXiv preprint*, abs/2407.06112, 2024e. URL https://arxiv.org/abs/2407.06112.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: Understanding the competition dynamics in large language model-based agents. In *Forty-first International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:270357283.

Qinglin Zhu, Runcong Zhao, Jinhua Du, Lin Gui, and Yulan He. Player*: Enhancing llm-based multi-agent communication and interaction in murder mystery games. *ArXiv preprint*, abs/2404.17662, 2024a. URL https://arxiv.org/abs/2404.17662.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. *ArXiv preprint*, abs/2402.18496, 2024b. URL https://arxiv.org/abs/2402.18496.